Proceedings

The 14th International Conference on
Green and Human Information Technology

ICGHIT 2026

Feb.4 – 6, 2026 / Bali, Indonesia



http://icghit.org/

ISSN : 2466-121X

The ICGHIT 2026 Committee

General Co-Chairs:

- Hyunsik Ahn (Tongmyong Univ., Korea)
- Kyutae Lee (Kongju National Univ., Korea)
- Imam Wahyudi Farid (Institut Teknologi Sepuluh Nopember, Indonesia)

Organizing Co-Chairs:

- Eun Sung Jung (Hongik Univ., Korea)
- Gde Wikan Pradnya Dana (Universitas Warmadewa, Indonesia)
- Eko Setijadi (Institut Teknologi Sepuluh Nopember, Indonesia)
- Yurid Eka Nugraha (Institut Teknologi Sepuluh Nopember, Indonesia)

International Advisory Co-Chairs:

- Sang-ug Kang (Sangmyung Univ., Korea)
- Seungcheon Kim (Hansung Univ., Korea)
- Sungwon Kim (Yeungnam University, Korea)
- Gyu Sang Choi (Yeungnam University, Korea)
- Duckki Lee (Yonam Institute of Technology, Pres. Of IEIE CIS, Korea)

Local Management Co-Chairs:

- Lienggar Rahadiantino (Institut Teknologi Sepuluh Nopember, Indonesia)
- Siska Arifiani (Institut Teknologi Sepuluh Nopember, Indonesia)
- Akhmad Ibnu Hija (Institut Teknologi Sepuluh Nopember, Indonesia)
- Faiza Alif Fakhrina (Institut Teknologi Sepuluh Nopember, Indonesia)
- Ridho Hendra Yoga Perdana (Chungbuk National University, Korea)

International Journal Chair:

- Byung Seo Kim (Hongik Univ., Korea)

Publication Co-Chairs:

- Ki Young Lee (Incheon National Univ., Korea)
- Soyoung Rho (Wolsong Books Co., Korea)

Finance and Administration Chair:

- Alexis Jinhong Kim (Paichai Univ., Korea)

Promotion Co-Chairs:

- DoHyun Kim (JeJu Nat'l Univ., Korea)
- Yong-Gyu Jung (Eulji Univ., Korea)

Industrial Cooperation Co-Chairs:

- SungChul Yu (ITCEN TECT., Korea)
- Harisatul Agustin (Institut Teknologi Sepuluh Nopember, Indonesia)

Workshop Co-Chairs:

- Soohyun Park (Kookmin Univ., Korea)
- Sungwook Park (GangneungWonju Nat. Univ., Korea)

Technical Program Committee Co-Chairs:

- Young-Hoon Park (Sookmyung Women's Univ., Korea)
- Kyusung Shim (Hankyong Nat. Univ., Korea)
- Changheun Oh (Hongik Univ., Korea)

Steering Committee Members:

- Beongku An (Hongik Univ., Korea)
- Moonsik Kang (GangneungWonju Nat. Univ., Korea)
- Seong Oun Hwang (Gacheon Univ., Korea)
- Yongsoo Choi (Shinhan Univ., Korea)

Technical Program Committee Members:

- Afizan Azman (Multimedia Univ., Malaysia)
- Anil Dubey (Poornima Institute of Eng. & Technology, India)
- Ali Reza Masoum (Urmia Univ. of Technology, Iran)
- Balázs Sonkoly (Budapest University of Technology and Economics, Hungary)
- Bilal Ahmed Lodhi (Teesside University, UK)
- Byeungwoo Jeon (SungKyunKwan Univ., Korea)
- Bok-Min Goi (Universiti Tunku Abdul Rahman, Malaysia)
- Bui Huu Phu (HCMUTE, Vietnam)

- Bongsang Suh (All4land, Korea)
- Claudia Linnhoff-Popien (Institut für Informatik, Germany)
- Chin Ji Jian (Multimedia Univ., Malaysia)
- Chibong Song (Wavers Co., Korea)
- Chom Kimpan (Panyapiwat Institute of Management, Thailand)
- Cheol-Min Park (NIMS, Korea)
- Chiung Ching Ho (Multimedia Univ., Malaysia)
- Duckki Lee (Yonam Institute of Technology, Korea)
- Derya Yiltas (Istanbul Univ., Turkey)
- Dao Thanh Toan (University of Transport and Communications)
- Dayu Kao (Central Police Univ., Taiwan)
- Derya Yiltas-Kaplan (Istanbul Univ., Turkey)
- Dongseop Lee (EzInno Lab Co. Ltd., Korea)
- Eun Sung Jung (Hongik Univ., Korea)
- Franklin Bien (UNIST, Korea)
- Gilbert M. Tumibay (AUF, Philippines)
- Gnana Vathanam (VIT Univ. Chennai Campus, India)
- Goh Hock Guan (Universiti Tunku Abdul Rahman, Malaysia)
- Geong Sen Poh (MIMOS Berhad, Malaysia)
- Goh Kah Ong Michael (Multimedia Univ., Malaysia)
- Gnana Swathika (Vellore Institute of Technology, India)
- Gokhan Secinti (Istanbul Technology University, Türkiye)
- Hoo-Jin Lee (Hansung Univ., Korea)
- Hyung-Woo Lee (Hanshin Univ., Korea)
- Hyeonsang Eom (Seoul National Univ., Korea)
- Hoang Van Xiem (Vietnam National Univ., Vietnam)
- Hong Min (Hoseo Univ., Korea)
- Ivan Ku (Multimedia Univ., Malaysia)
- Imran Khan (Cardiff Metropolitan University, UK)
- Jang-Geun Ki (Kongju National Univ., Korea)
- Ji-Jian Chin (Multimedia Univ., Malaysia)
- Jiyoung Kang (Dankook Univ., korea)
- Jungwoo Ryoo (Penn State Univ., USA)
- Jianping He (Penn State Univ., USA)
- Junbeom Hur (Korea Univ., Korea)
- Jayabalan Sudharsan (MRIET, India.)
- Junseok Lee (ETRI, Korea)
- Justinian Anatory (Univ. of Dodoma, Tanzania)
- Jiwa Abdullah (Univ. Tun Hussein Onn Malaysia, Malaysia)
- Jaeseok Kim (Yonsei Univ., Korea)
- Jun Yan (Nanjing Univ. of Posts and Tele., China)

- Jong-Yun Kim (Kyungdong Univ., Korea)
- Jawad Ahmed (Edinburgh Napier University, UK)
- Kyusung Shim (Hankyong National Univ., Korea)
- Kyung-Rak Sohn (Korea Maritime and Ocean Univ., Korea)
- Le Minh Thuy (HaNoi University of Science and Technology)
- Le Sun (Nanjing Univ. of Info. Science & Technology, China)
- Lim Kian Ming (Multimedia Univ., Malaysia)
- Lee Chin Poo (Multimedia Univ., Malaysia)
- Laszlo Toka (Budapest University of Technology and Economics, Hungary)
- Moon S. Lee (Univ. of Luxembourg, Luxembourg)
- Michele Fiorini (The IET Council Chairman, Italy)
- Md Shohel Sayeed (Multimedia Univ., Malaysia)
- Muazzam Khattak (NUST, Pakistan)
- Myung-Sup Kim (Korea Univ., Korea)
- Muhammad Usman Akram (NUST, Pakistan)
- Mohammad Haseeb Zafar (Cardiff Metropolitan University, UK)
- Nguyen Minh Dat (Institut National de recherche Scientifique, Canada)
- Navneet Agrawal (MPUAT, India)
- Nguyen Van Trung (Le Quy Don Tech. Univ., Vietnam)
- Nguyen Quoc Dinh (Le Quy Don Tech. Univ., Vietnam)
- Nguyen Minh Giang (Le Quy Don Tech. Univ., Vietnam)
- Nguyen Tuan Hung (Le Quy Don Tech. Univ., Vietnam)
- Nguyen Van Huynh (Edinburgh Napier University, UK)
- Noel Tay Nuo Wi (Tokyo Metropolitan Univ., Japan)
- Nalini Iyer (KLE Technological Univ., India)
- Ong Thian Song (Multimedia Univ., Malaysia)
- Pang Ying Han (Multimedia Univ., Malaysia)
- Pham Viet Hung (Vietnam Maritime Univ., Vietnam)
- Piotrowski Zbigniew (Military Univ. of Technology, Poland)
- Parinya Sanguansat (Panyapiwat Institute of Management, Thailand)
- Quoc Huy Nguyen (Saigon Univ., Vietnam)
- Omar Dakkak (Karabük University, Türkiye)
- Rashid Ali (Universitat Pompeu Fabra, Spain)
- Rana Asif Rehman (National University of Computer and Emerging Sciences, Pakistan)
- Rajkumar Kettimuthu (Argonne National Lab., USA)
- Renny Badra (Universidad Simón Bolívar, Caracas, Venezuela)
- Roy Chang (Multimedia Univ., Malaysia)
- Renny Badra (Universidad Simon Bolivar, Venezuela)
- Rashid Mehmood (King Abdul Aziz Univ., Saudi Arabia)
- Ronnel Atole (Partido State Univ., Philippines)
- Seung-Hun Jin (ETRI, Korea)

- Sarhan Musa (Prairie View A&M Univ., USA)
- Shivani Sud (Intel, USA)
- Shing Chiang Tan (Multimedia Univ., Malaysia)
- Sweety Chauhan (Microsoft, USA)
- Sansanee Auephanwiriyakul (Chiang Mai Univ., Thailand)
- Seon Wook Kim (Korea Univ., Korea)
- Stenio Fernandes (Federal Univ. of Pernambuco, Brazil)
- Sejong Kim (SJ Telecom, Korea)
- Shahid Butt (NUST, Pakistan)
- Shing-Chiang Tan (Multimedia Univ., Malaysia)
- SangYep Nam (Kookje Univ., Korea)
- Sana Ullah Jan (Edinburgh Napier University, UK)
- Taehwa Han (Yonsei Univ. Health System, Korea)
- Tekin Bicer (Argonne National Lab., USA)
- Tom Oh (Rochester Institute of Technology, USA)
- Truong Xuan Tung (Le Quy Don Tech. Univ., Vietnam)
- Tran Thi Thu Huong (Le Quy Don Tech. Univ., Vietnam)
- Tri Nhu Do (Polytechnique Montreal, Canada)
- Taekyoung Kwon (Yonsei Univ., Korea)
- Tan Tran Duc (Vietnam National Univ., Vietnam)
- Tan Syh Yuan (Multimedia Univ., Malaysia)
- Tingjun Xie (Univ. of Virginia, USA)
- Tudor Palade (Technical Univ. of Cluj-Napoca, Romania)
- Tee Connie (Multimedia Univ., Malaysia)
- Thumrongrat Amornraksa (KMUTT, Thailand)
- Thai Son Nguyen (Tra Vinh Univ., Vietnam)
- Tammam Tillo (Libera Univ. of Bozen-Bolzano, Italy)
- Tho Quan (HCMUT, Vietnam)
- Tariq Rahim (Kingston University London, UK)
- Vo Nguyen Quoc Bao (Posts and Telecommunications Institute of Technology)
- Vineet Kumar (CG Power and Industrial Solutions, India)
- Woontack Woo (KAIST., Korea)
- Wonsoo Kim (Texas Instruments, USA)
- Woomin Hwang (NSRI, Korea)
- Wathiq Mansoor (Univ. of Dubai, UAE)
- Wai-Kong Lee (Universiti Tunku Abdul Rahman, Malaysia)
- Wali Ullah Khan (University of Luxembourg, Luxembourg)
- Young B. Choi (Regent Univ., USA)
- Yap Wun She (Universiti Tunku Abdul Rahman, Malaysia)
- Yong-Gyu Jung (Eulji Univ., Korea)
- Yan-Ming Cheng (Beihua Univ., China)

Message from General Co-Chairs of ICGHIT 2026



Dr. Imam Wahyudi Farid
General Co-Chairs
Institut Teknologi Sepuluh Nopember, Indonesia



Prof. Kyutae Lee
General Co-Chairs
Kongju National University, Korea



Prof. Hyunsik Ahn
General Co-Chairs
Tongmyong University, Korea

We are very happy to welcome you to the International Conference on Green and Human Information Technology (ICGHIT) 2026. This year, we are excited to hold the conference from February 4 to 6 at Warmadewa University in the beautiful Bali, Indonesia.

ICGHIT began in 2013 to help researchers from all over the world share their technical ideas and work together. Our mission is to use IT to improve human welfare and happiness. Today, AI technology is growing faster than ever, changing our economy, society, and culture. By combining AI with robotics, the metaverse, and IoT, we are entering a new era. This conference is a place to share these exciting changes.

The theme for 2026 is “Vertical AI.” While General AI is already helping us in our daily lives, “Vertical AI” focuses on solving specific problems in fields like healthcare, finance, manufacturing, and law. We look forward to deep academic discussions on how AI can provide professional support in these specific areas. We also hope you enjoy meeting other researchers and companies for future collaboration.

In ICGHIT 2026, more than 150 scientists from 12 countries (Korea, Indonesia,

Egypt, Philippines, Vietnam, India, Malaysia, USA, Pakistan, Bangladesh, Germany, and Jordan) will present their research.

We have prepared four plenary talks, multiple technical sessions for oral and poster presentations, and a special workshop.

We have received 131 paper submissions, which were peer-reviewed by experts in the area of the paper's subject. The selected papers, 97 high-quality research papers, consist of 67 oral presentation papers (Korea: 24, other countries: 43) and 30 poster presentation papers.

Preparing this conference required a lot of support. First, we sincerely thank our sponsors: IEEE, The Institute of Electronics and Information Engineers (IEIE), the Green Human IT Association (GHIT), University of Warmadewa(UW), and Institut Teknologi Sepuluh Nopember (ITS).

We extend our deepest gratitude to Prof. Dr. Ir. I Gde Suranaya Pandit, M.P., Rector of Warmadewa University (UW), for his strong leadership and support in hosting this conference.

We also deeply appreciate the dedication of Prof. Eun-Sung Jung in Hongik University, Korea and Prof. Dr. Ir. I Nengah Sinarta, S.T., M.T., IPM., ASEAN.Eng., Acting Dean of the Faculty of Engineering at Warmadewa University, for his tireless efforts in preparing all aspects of the event.

Our heartfelt thanks go to Gde Wikan Pradnya Dana and the dedicated faculty members of Warmadewa University for their outstanding local organizing efforts, which significantly contributed to the success of the conference.

Furthermore, we express our appreciation to Prof. Dr. Wahyu Wibowo, S.Si., M.Si., Dean of the Faculty of Vocational Studies at Institut Teknologi Sepuluh Nopember (ITS), Lucky Putri Rahayu, S.Si., M.Si., Head of the Department of Electrical Automation Engineering, and Dr. Eng. Imam Wahyudi Farid, S.T., M.T., Director of the Center of Excellence in Mechatronics and Industrial Automation at ITS, for their exemplary collaboration and successful local management.

In particular, we would like to thank the companies who provided financial support for this conference. We also appreciate the committee members who have been responsible and fulfilled successfully in their fields, and reviewers who carried out

the careful evaluation of the submitted papers.

Above all, we would like to express our deep appreciation to all authors, participants, and speakers for making ICGHIT 2026 prosper. We wish you to have a great time during your stay in this beautiful place, Bali, Indonesia.

Thank you!

Welcome Message from the Rector of Warmadewa University



Prof. Dr. Ir. I Gde Suranaya Pandit, MP.
Rector of Warmadewa University

Dear Distinguished Guests, Respected Professors, Esteemed Colleagues, and Participants, Ladies and Gentlemen,

It is my great honor and pleasure to warmly welcome you to the ICGHIT 2026, held at Warmadewa University in the beautiful island of Bali, Indonesia.

We are truly delighted to have the opportunity to gather scholars, researchers, and professionals from around the world in this inspiring and culturally rich setting. Bali is internationally renowned not only for its natural beauty, but also for its strong cultural heritage, creativity, and openness—values that closely align with the spirit of academic collaboration and innovation fostered by this conference.

This conference represents an important milestone for Warmadewa University, reflecting our continuous commitment to advancing education, research, and international academic cooperation. Through this forum, we aim to provide a meaningful platform for the exchange of ideas, the presentation of cutting-edge research, and the strengthening of global research networks across disciplines.

We are pleased to note the enthusiastic participation in this year's conference, with contributions from researchers representing diverse institutions and countries. This strong engagement highlights the growing global interest in the conference themes and underscores the relevance of the discussions that will take place over the coming days.

Warmadewa University is a dynamic and forward-looking institution dedicated to excellence in education, research, and community engagement. We continuously strive to nurture innovation, critical thinking, and human-centered development to contribute positively to society at both national and international levels.

On behalf of Warmadewa University, I would like to express my sincere appreciation to the organizing committee, reviewers, sponsors, keynote speakers, and all contributors whose dedication and hard work have made this conference possible.

I sincerely hope that ICGHIT 2026 will serve as a fruitful forum for insightful discussions, inspiring presentations, meaningful collaborations, and the generation of impactful ideas for the future. Once again, welcome to Warmadewa University and to Bali. I wish you a successful, rewarding, and memorable conference experience.

Thank you.

Sincerely yours,

Prof. Dr. Ir. I Gde Suranaya Pandit, MP.
Rector of Warmadewa University

Welcome to ICGHIT 2026



Prof. Duckki Lee
International Advisory Chair and President of IEIE Computer Information Society
Yonam Institute of Technology, Korea

Welcome to Bali, Indonesia!

On behalf of IEIE CIS (The Institute of Electronics and Information Engineering Computer Information Society), it is a great pleasure to welcome you all to ICGHIT 2026 (International Conference on Green and Human Information Technology), in Bali, Indonesia. I am very happy to hold this conference in Bali, one of the most famous islands in the world, well known for its beautiful nature and rich cultural heritage.

First of all, I express my appreciation to all speakers, participants, professors, researchers, and students coming from multiple countries. And I thank the General Co-Chairs of this conference, Prof. Hyunsik Ahn, Prof. Kyutae Lee, and Prof. Imam Wahyudi Farid, the Organizing Co-Chairs, Prof. Eun Sung Jung, Prof. Gde Wikan Pradnya Dana, Prof. Eko Setijadi, and Prof. Yurid Eka Nugraha, and all the committee members who have eagerly devoted themselves to the preparation of ICGHIT 2026.

ICGHIT first began as a small workshop to encourage friendship and partnership between Korean IEIE CIS members and researchers and professors from other countries. At the beginning, it was modest and limited in scale, but through continuous efforts and sincere cooperation, it has steadily grown. Today, ICGHIT has evolved into one of the active international conferences, serving as an important platform for the advancement of green technology and human-centered information

technology in various engineering and industrial fields.

The IEIE is the largest institute of electronics and information engineering in Korea, consisting of professors, researchers, and students in IT-related areas. Moreover, it has developed its vision to encourage international relations. As a founder and sponsor of ICGHIT, IEIE CIS will continue to support this conference to pursue its philosophy, which is to realize the dream of green and human information technology that improves human welfare and happiness, together with friends and colleagues from many countries.

I would like to express my sincere appreciation to Warmadewa University, its leadership, faculty members, and staff for their generous hospitality and strong support in hosting ICGHIT 2026. Their dedication and excellent organization have created a welcoming academic environment that greatly contributes to the success of this conference and provides valuable opportunities for academic exchange and cooperation.

Once again, I would like to express my heartfelt welcome to all participants and speakers. I sincerely wish you to have a great academic experience and an enjoyable stay during your time in this beautiful island, Bali, Indonesia.

Thank you.

Table of Contents

CI1: Communication & IoT1

IoT-Based Wearable Device for Real-Time Sitting Posture Monitoring and Early Kyphosis Prevention ····· **001p**

Eka Septia Amini (Universitas Telkom, Indonesia); Sevia Indah Purnama (Universitas Telkom, Indonesia & Telkom University, Indonesia); Fikra Titan Syifa (Institut Teknologi Telkom Purwokerto, Indonesia)

Bandwidth Improvement in a Compact 3D-Printed Conical Patch Antenna by Substrate Profiling ····· **006p**

Manasa Chinnam and Vakula D (NIT Warangal, India)

BER Analysis of Index Modulation in STAR RIS NOMA System ····· **010p**

Mangali Shanti Priya (National Institute of Technology Warangal, India); Anuradha Sundru (NIT Warangal, India)

A Prototype of Non-Invasive Blood Sugar Measurement Device Using MAX30102 and IoT Connectivity ····· **016p**

Evan Fajar Firmansyah (Universitas Telkom, Indonesia); Dodi Zulherman and Slamet Indriyanto (Telkom University, Indonesia)

Design and Development of an IoT-Based Blood Oxygen (SpO_2) Monitoring Device for Asthma Patients ····· **022p**

Irvine Kornelius (Telkom University Purwokerto, Indonesia); Sevia Indah Purnama (Universitas Telkom, Indonesia & Telkom University, Indonesia); Slamet Indriyanto (Telkom University, Indonesia); Alvera Carolina (Jendral Soedirman University, Indonesia)

W1: Workshop – AI-RSFMS 2026 (1)

Secure OTA Firmware Update Using BlockChain and Role Based Access Control ····· **031p**

Pak Hyeonggeon and Seung Yeob Nam (Yeungnam University, Korea (South))

An IRS-UAV-Assisted EMF-Aware PD-NOMA Framework for Uplink Communication in Dense Urban Networks ····· **036p**

Maliha Binte Mukhles and Sung Won Kim (Yeungnam University, Korea (South)); M Ishtiaque Ahmed (Jahangirnagar University, Bangladesh)

Let's Fight the Dirt: SIR-GAN for Robust Camera Perception in Autonomous Vehicles ····· **044p**

Rahman Shafique (Yeungnam University South Korea, Korea (South)); Gyu Sang Choi (Yeungnam University, Korea (South))

SemiSwin-DehazeNet: Wavelet-Transformer Restoration for Low-Visibility Navigation ···· **052p**

Hira Khan (Yeungnam University, Korea (South)); Muhammad Khalil Afzal (COMSATS University Islamabad Wah Campus, Pakistan); Sung Won Kim (Yeungnam University, Korea (South))

TrajFormer: A Unified One-Stage Transformer for Pedestrian Detection, and Trajectory Prediction in Complex Urban V2X Environments ····· **058p**

Memoona, Muhammad Haseeb Hashir and Sung Won Kim (Yeungnam University, Korea (South))

V2XNet: Vehicle-To-Everything Network for LLM-Guided Multi-Agent Path Planning Under Complex Urban Mobility Conditions with Convergent Nash Equilibrium ····· **064p**

Muhammad Haseeb Hashir, Memoona and Sung Won Kim (Yeungnam University, Korea (South))

AI1: AI & Big Neural Network 1

Modeling User Comfort and Usability in Human-AI Interaction on Social Media Using One-Way ANOVA for Vertical AI Applications ····· **070p**

Stiven Hizkia Hardana (Bina Nusantara University, Indonesia); Nuril Kusumawardani (Binus University, Indonesia)

Authentic Dataset Development and Exploration of Cross-Language Transfer Learning for Aksara Jawa Recognition ····· **076p**

Bryan Myer Setiawan, Aura Kristian Sumowidjojo and Andry Chowanda (Bina Nusantara University, Indonesia)

Smartphone-Based Melanoma Detection Using Edge Impulse and Convolutional Neural Networks ····· **083p**

Nurul Vita Dwi Faraya (Universitas Telkom, Indonesia); Sevia Indah Purnama (Universitas Telkom, Indonesia & Telkom University, Indonesia); Irmayatul Hikmah (Telkom University, Indonesia & Institut Teknologi Telkom Purwokerto, Indonesia)

Efficient Sentiment Classification via RoBERTa-BiLSTM and Nature-Inspired Feature Optimization ····· **088p**

Uzma Shakoor and Muhammad Khalil Afzal (COMSATS University Islamabad Wah Campus, Pakistan); Byung-Seo Kim (Hongik University, Korea (South))

Analyzing User Acceptance of News on Social Media Using Machine Learning ····· **094p**

Mochammad Risyad Aukasyah Fatahillah and Jeremia Johanes Mikhael Paulus (Binus University, Indonesia); Eka Miranda (Bina Nusantara University, Indonesia)

Digital Marketing Website for Gender Equality: Empowering Women Entrepreneurs in Labuan Bajo's Culinary Tourism ····· **100p**

Wahyu Sardjono (University of Bina Nusantara, Indonesia); Trias Septyoari Putranto, Maryani Maryani and Nora Fitriawati (Bina Nusantara University, Indonesia); Hasyiya Karimah Adli (Universiti Malaysia Kelantan, Malaysia)

CI2: Communication & IoT2

Design of Reconfigurable Modulation System for Selective Cellular Interference with SDR ····· **106p**

Diky Arif Hidayat (Universitas Telkom, Indonesia); Eka Nadiyah Purwanianti, Daniel Alden Susetyo, Amalia Muhammad Jasmin and Budi Prasetya (Telkom University, Indonesia)

A Portable Intelligent System for Lung Disease Prediction Using Machine Learning Models 112p

Dipta Paul (American International University-Bangladesh (AIUB), Bangladesh); Ekramul Hasan (Westcliff University, USA); Md Al Amin and Sk Rakib Ul Islam Rahat (International American University, USA); Mustafizur Rahaman (Westcliff University, USA); Saiful Islam Dewan (American International University Bangladesh (AIUB), Bangladesh)

Performance Evaluation of Hybrid Machine Learning Models for Heart Disease Prediction in U.S. Clinical Decision Support Systems 118p

Dipta Paul (American International University-Bangladesh (AIUB), Bangladesh); Ekramul Hasan (Westcliff University, USA)

Psoriasis Classification Based on Convolutional Neural Network (CNN) with Explainable AI Interpretation Using Grad-CAM 124p

Putri Anggrawati Cahyani (Universitas Telkom, Indonesia); Sevia Indah Purnama (Universitas Telkom, Indonesia & Telkom University, Indonesia); Irmayatul Hikmah (Telkom University, Indonesia & Institut Teknologi Telkom Purwokerto, Indonesia)

Automated Greenwashing Risk Detection: An LLM-Driven Linguistic Index for Sustainability Auditing 133p

Silvi Agustina, Irni Yunita and Andry Alamsyah (Telkom University, Indonesia)

Trends in Online Monitoring of OT- and IIOT-Networks 139p

Axel Sikora (University of Applied Sciences Offenburg, Germany)

W2: Workshop – AI-RSFMS 2026 (2)

Adaptive Weighted Fairness Power Allocation for Vehicular NOMA 143p

Kagudde Abbas and Kwonhue Choi (Yeungnam University, Korea (South))

Shaft Dynamic Behavior Analysis and Optimal Sensor Placement Strategy for Crack Diagnosis in Drive Shaft Splines 149p

Jaewon Song, Jaehyun Lee and Byeongil Kim (Yeungnam University, Korea (South))

Adaptive Fault Detection and Monitoring Frequency Reduction in NoC Systems Using Reinforcement Learning for Energy Efficiency 153p

Shaheerullah and Sung Won Kim (Yeungnam University, Korea (South))

Road Traffic Collision Severity Modeling Using Bayesian Multi-Level Multinomial Logistic Regression 159p

Emmanuel Byamugisha, Sheriff Murtala, Ingyu Lee and Gyu Sang Choi (Yeungnam University, Korea (South))

P1: Poster Session

Integrated Digital Medicine Record System as a Service Innovation Pharmacy Based on User Centered Design 165p

Randy Tejonegoro (Institut Teknologi Sepuluh November, Indonesia); Feby Artwodini Muqtadiroh (Institut Teknologi Sepuluh Nopember (ITS), Indonesia); Mauridhi Hery Purnomo (Institut of Technology Sepuluh Nopember, Indonesia)

Optimization of Radix-3 NTT for NTRU+ on GPU ····· **171p**

Ryeo Dam Kim, Kyung Hyun Han and Seong Oun Hwang (Gachon University, Korea (South))

The Impact of TAS on RIS-Aided ISAC ····· **174p**

Kyusung Shim (Hankyong National University, Korea (South)); Soo-Hyun Park (Kookmin University, Korea (South))

Enhancing Multi-Agent Reinforcement Learning Robustness via Intrinsic Motivation and Emergent Emotional Dynamics ····· **176p**

Sospeter Birir, Hyoungju Kim and Pankoo Kim (Chosun University, Korea (South))

AI-Based Adaptive Audio Generation with Frequency-Specific Processing ····· **183p**

Taesoo Park, MuJung Kim, Dohyun Park, Hyunjun Park and Jisang Yoo (Kwangwoon University, Korea (South)); Hoyun Lee (Ewha Womans University, Korea (South)); Sanghoon Kim (KyungHee University, Korea (South)); Soon Chul Kwon (Kwangwoon University, Korea (South))

DeepSets-s: An Encoder-Decoder Model for Real-Time Resource Allocation in in-Network Edge Metaverse Systems ····· **186p**

Rashid Muhammad Sulaiman, Ibrahim Aliyu, Seungmin Oh, Yonggwan Won, Jaehyung Park and Jinsul Kim (Chonnam National University, Korea (South))

Prediction of Atmospheric Refractivity from Clutter Power Using Deep Learning Based on the Tri-Linear Atmospheric Model and Ocean Spectrum ····· **192p**

Taekyeong Jin and Hosung Choo (Hongik University, Korea (South))

Analysis of Solar Interference Impact on Future Networks in Extreme Cold Regions ·· **194p**

ByoungWoo Jeon and Ari Hwang (Kookmin University, Korea (South)); Dongjin Yoon (Office of Technology Development and Service of Korea Polar Research Institute, Korea (South)); Soo-Hyun Park (Kookmin University, Korea (South))

Deep Learning-Based Clustering Protocol for Digital Twin-Enabled IoT-MANETs ····· **201p**

Junaedi Adi Prasetyo (Politeknik Negeri Banyuwangi, Indonesia); Ridho Hendra Yoga Perdana (Chungbuk National University, Korea (South)); Yushintia Pramitarini (Korea University, Korea (South)); Axel Sikora (University of Applied Sciences Offenburg, Germany)

A Study on a Privacy-Preserving Platform for Protecting and Using Medical Data ····· **206p**

Muneeb Muhammad and Oudone Phengkhamlar (Sang Ji University, Korea (South)); Kwangman Ko (Sangji University, Korea (South))

Design of a Security Event-Driven Monitoring System Architecture for Software-Defined Vehicles ····· **213p**

Tai-Gil Kwon (Korea Electronics Technology Institute, Korea (South))

Evaluation of Acceptance and Fidelity Scaling for Topology-Aware Zero-Level Magic-State Distillation in Lattice and Heavy-Hex Architecture ····· **215p**

Leanghok Hour (Pukyong Nation University, Korea (South)); Dongmin Kim, Sanghyeon Lee and Youngsun Han (Pukyong National University, Korea (South))

Wireless Power Transfer System Primary Current Detector Based for Wide Range Resonant Frequency Tracking ····· **218p**

Youngseok Ko (Dankook University, Korea (South)); Eunho Choi (Newhealthvision Co., Ltd., Suwon-si, South Korea); Hyunkyeong Jo (BWRIGHT Inc., Gyeongsan-si, South Korea); Heeyoung Chae (Dong-A University, Korea (South)); Shihong Park (Dankook University, Korea (South))

Design Enhancement of a Lifecycle-Based Verification Simulator for Digital Medical Products ····· **221p**

Sung-Min Jeon and JungHyun Woo (Sang Ji University, Korea (South)); Kwangman Ko (Sangji University, Korea (South)); LeeJinHyong (Sang Ji University, Korea (South)); Byung-Suk Seo (Sangji University, Korea (South)); Hanyong Choi (Shinhan University, Korea (South))

SWKD: Scalable Wavelet Knowledge Distillation for Vision Transformers ····· **227p**

JungMin Kim, Changhee Kang and Sang-ug Kang (Sangmyung University, Korea (South))

CNS: Computer & Network Security

An Initial Study Toward a Protocol-Level Cybersecurity Maturity Model: The Evolution of MQTT Authentication ····· **229p**

Arfan Ridwan Hartawan and Muhammad Salman (Universitas Indonesia, Indonesia)

UVASS-FDD: A Multimodal Deepfake Detection Framework Integrating Audio-Visual Consistency to Enhance Content Integrity ····· **235p**

Laode Alif Ma'sum Sidrajat Raja Ika and Prima Dewi Purnamasari (Universitas Indonesia, Indonesia)

Evaluation of ML Models Using Feature Scaling, ADASYN and K-Means Clustering for Intrusion Detection System ····· **241p**

Faizal Riza (Universitas Putra Indonesia YPTK Padang, Indonesia); Iskandar Fitri (Universitas Putra Indonesia YPTK, Indonesia); Sumijan Sumijan (Universitas Putra Indonesia YPTK Padang, Indonesia)

Who Prompted What? Privacy-Preserving Incident Detection for LLM Systems Using Minimal Audit Logs ····· **247p**

Hariprasad Holla (Ieee.org, USA); Bhaskar Bharat Sawant (Cornerstone Building Brands, USA); Chakra Pavan Kumar Kota (Walmart, USA); Ankur Vora (Ieee.org, USA)

Enhancing SnortML Intrusion Prevention Using LightGBM to Neural Network Distillation ····· **253p**

Bintang Pandu Anugrah, Dwaynimay Farrel Rajendra and Alika Widuri Kartika (Universitas Telkom, Indonesia); Salna Siraz Maulidia and Nyoman Karna (Telkom University, Indonesia); Arif Indra Irawan (Okayama University, Japan & Universitas Telkom, Indonesia)

Design and Deployment of an Autoencoder-Binary Classifier Intrusion Prevention Model for SnortML ····· **259p**

Alika Widuri Kartika, Bintang Pandu Anugrah and Dwaynimay Farrel Rajendra (Universitas Telkom, Indonesia); Salna Siraz Maulidia and Nyoman Karna (Telkom University, Indonesia); Arif Indra Irawan (Okayama University, Japan & Universitas Telkom, Indonesia)

W3: Workshop – SACT 2026

IOTA Tangle-Based Digital Evidence Anchoring System for International Investigative Assistance 265p

Chang-Jun Choi and Hyunsoo Kim (LSware Inc., Korea (South)); Kyung-Yeob Park (Lsware Inc., Korea (South)); Dong-Myung Shin (LSWare. Inc., Korea (South))

YOLO-PT: One-Shot Framework for 3D-Based Progress Tracking 268p

Lan Anh Nguyen (Hongik University, Korea (South)); Jaemo Sung (DesignedAI, Korea (South)); Eun-Sung Jung (Hongik University, Korea (South))

Design of a ContractNLI-Based Intermediate-Task Trained Natural Language Inference Model for Game and Software License Violation Detection 274p

Dong-Wan Kim and Chang-Jun Choi (LSware Inc., Korea (South)); YongJoon Joe (LSware Inc. & PPUZZL Group Inc., Korea (South)); Dong-Myung Shin (LSWare. Inc., Korea (South))

Multi-Layered Framework for Evaluating Requirements Completeness in User-Story-Driven Agile Development 278p

Yukyong Kim (Sookmyung Women's University, Korea (South))

Digital Forensics of the Embedded On-Devices AI 284p

Kyutae Lee, Hyun Mook Cho and Jang-Geun Ki (Kongju National University, Korea (South)); DoHyeun Kim (Jeju National University, Korea (South))

DAD: SW/HW Design, Architecture & Development

Exploring User Experience Components for the HSE Metaverse Evaluation Framework 286p

Al-amin Abdulhamid (Universiti Teknologi Petronas, Malaysia); Khairul Shafee Kalid and KS Savita (Universiti Teknologi PETRONAS, Malaysia); Shamsu Abdullahi (University of Technology PETRONAS, Malaysia & Hassan Usman Katsina Polytechnic, Nigeria); Muhammad Sabo Yahaya (Universiti Teknologi PETRONASA (UTP), Malaysia & Federal Polytechnic Daura, Katsina, Nigeria); Sarah Aisyah Isnani (Universiti Teknologi Petronas, Malaysia)

Prototype of Chronic Obstructive Pulmonary Disease Therapeutic Devices Based on Hall Effect Sensor 291p

Anggita Dewi Puspita (Telkom University Purwokerto, Indonesia); Nur Afifah Zen and Dodi Zulherman (Telkom University, Indonesia)

Managerial Pragmatism and Tactical Responses in Requirements Traceability: An Interview-Based Study 298p

Raditya Pratama, Auzan Taqi Syafega, Fausta Irsyad Ramadhan and Anisah Mutmainah (Institut Teknologi Sepuluh Nopember, Indonesia); Feby Artwodini Muqtadiroh (Institut Teknologi Sepuluh Nopember (ITS), Indonesia)

Optimized Differential Ring VCO Architecture Featuring Wide Tuning Range, Low Phase Noise, and Enhanced Linearity in CMOS Technology 304p

O Rakesh (National Institute of Technology Warangal, India); Patri Sreeharirao (National Institute of Technology, Warangal, India)

Designing an Integrated Digital Governance Architecture for Disaster-Resilient Villages 309p

Astuti Purwaningsih and Lukito Edi Nugroho (Universitas Gadjah Mada, Indonesia); Rudy Hartanto (Gadjah Mada University & Electrical Engineering and Information Technology Departmen, Faculty of Engineering Gadjah Mada University, Indonesia)

MSP: Multimedia and Signal Processing

Design of a High-Rate Modulation Code for 4-Ary Holographic Data Storage Systems 315p

Chi Dinh Nguyen and Khanh Long Tran (FPT University, Vietnam)

Augmentation Effect on Indonesian Sign Language Image Classification 321p

Stefani Siahaan (Telkom University Purwokerto, Indonesia); Sevia Indah Purnama (Universitas Telkom, Indonesia & Telkom University, Indonesia); Eko Fajar Cahyadi and Dodi Zulherman (Telkom University, Indonesia)

Development and Evaluation of a Low-Cost EMG-Based System for Muscle Weakness Detection 327p

Kezia Hening Ata (Universitas Telkom, Indonesia); Sevia Indah Purnama (Universitas Telkom, Indonesia & Telkom University, Indonesia); Fikra Titan Syifa (Institut Teknologi Telkom Purwokerto, Indonesia)

Realtime 3D Animation Capture Using Pose Estimation and Video Segmentation with YOLOv11n-Seg Model 332p

Steven Steven (Binus University, Indonesia); Andry Chowanda (Bina Nusantara University, Indonesia)

Hybrid Scale Factor Adjustment for Spatial Resampling of VCM 338p

Min-suk Kim and Eun-bin An (Yonsei University, Korea (South)); Sangwoon Kwak (Electronics and Telecommunications Research Institute (ETRI), Korea (South)); Ayoung Kim (ETRI, Korea (South)); Soon-heung Jung (Electronics and Telecommunications Research Institute, Korea (South)); Kwang-deok Seo (Yonsei University, Korea (South))

Daily Municipal Solid Waste Forecasting: A Comparison of SARIMA and LSTM 341p

Syalom Veninda Runtuwene (Universitas Gadjah Mada, Indonesia & Technology Information, Indonesia); Rudy Hartanto (Gadjah Mada University & Electrical Engineering and Information Technology Departmen, Faculty of Engineering Gadjah Mada University, Indonesia); Adhistya Erna Permanasari (Universitas Gadjah Mada, Indonesia)

W4: Workshop – SACS 2026

Joint Resource Management and Trajectory Optimization Through Quantum-Inspired Reinforcement Learning in UAVs 347p

Muhammad Nadeem Ali (Hongik University, Korea (South)); Ihsan Ullah (Hongik Unviersity, Korea (South)); Ghulam Musa Raza (Hongik University, SEJONG CAMPUS, Korea (South)); Muhammad Asim, Bareera Anam and Byung-Seo Kim (Hongik University, Korea (South))

A Cross Layer Design for Joint Caching and Routing in Information-Centric Internet of Vehicles 350p

Muhammad Nadeem Ali (Hongik University, Korea (South)); Ihsan Ullah (Hongik Unviersity, Korea (South)); Ghulam Musa Raza (Hongik University, SEJONG CAMPUS, Korea (South)); Muhammad Asim, Bareera Anam and Byung-Seo Kim (Hongik University, Korea (South))

A Cross-Layer ResDNN-Based Multicast Routing Protocol in Flying Ad-Hoc Networks ··· **352p**

Amalia Amalia (Hongik University, Korea (South)); Yushintia Pramitarini (Korea University, Korea (South)); Ridho Hendra Yoga Perdana (Chungbuk National University, Korea (South)); Kyusung Shim (Hankyong National University, Korea (South)); Beongku An (Hongik University, Korea (South))

Transfer Learning for Joint Precoding and Phase Shift Optimization in Mobile RIS-Aided Networks ····· **358p**

Yosefine Triwidyastuti (Hongik University, Korea (South)); Tri Nhu Do (Polytechnique Montréal, Canada); Kyusung Shim (Hankyong National University, Korea (South)); Beongku An (Hongik University, Korea (South))

Large-Scale Benchmarking of AI Models for Predictive Resource Allocation in Open RAN ····· **364p**

Md Shakirul Islam and Eun-Sung Jung (Hongik University, Korea (South))

Dynamic Optimization of Input Context Length for Reducing Computational Cost and Power Consumption in Decision Transformers ····· **370p**

TaeHyun Han (Hongik University, Korea (South)); Janghwan Kim (Hongik University, Korea (South) & Software Engineering Laboratory, Korea (South)); Chanyeol Yang (Da-on, Korea (South)); Sanho Lee (RasTech, Korea (South)); Chaeyun Seo (SE Laboratory, Hongik University, Korea (South)); Robert Youngchul Kim (Hongik University, Korea (South))

GT: Green Information Technology

Sustainable Loops: Mapping the Scientific Landscape of Digital Circular Economy Solutions in Textile Waste ····· **372p**

Edison M Esberto and Mideth Abisado (National University, Philippines)

An Automatic Transfer Switch Scheme on a 7.7 kWp Solar Panel System with Metaheuristic Approximation ····· **379p**

Ryan Yudha Adhitya (Politeknik Perkapalan Negeri Surabaya, Indonesia & Institut Teknologi Sepuluh Nopember, Indonesia)

Optimization of the YOLOv12s Architecture with the Addition of 160×160 and C3k2 Heads for Floating Plastic Bottle Detection ····· **384p**

Novy N.R.A. Mokobombang (Hasanuddin University, Indonesia & Auckland University of Technology, New Zealand); Alief Akbar Agung El-Mahirah (Universitas Hasanuddin, Indonesia)

Optimal Control of a Nonlinear Tourism Environment Dynamic System for Sustainable Development ····· **390p**

Ida Ayu Putu Ari Utari (Institut Teknologi Sepuluh November (ITS), Indonesia)

GAT-Assisted Risk-Aware Isolation in a Bio-Inspired IoT Botnet Defense Simulation ····· **395p**

Vaishali Ravi, Mohd Anuaruddin Ahmadon, Shakirah Mohd Taib, Iswardi Yuzrin, Che Azizi Che Rus and Ahmad Hafiz Ghadzali (Universiti Teknologi Petronas, Malaysia)

Integrating Motivation and Organizational Context: A Conceptual Model for Gamification-Based Cybersecurity Awareness in SMEs ····· **401p**

Qinthara Putrie Zuvianne (Universitas Telkom, Indonesia); Puspita Kencana Sari (Telkom University, Indonesia)

AI2: AI & Big Neural Network2

Orbital Intelligence: A Vision for Sustainable Large Language Model Training and Inference in Space ····· **407p**

Mohamed Khalil Brik (The American University in Cairo, Egypt)

Prescriptive and Contextual Technical Debt Management with LLM and SonarQube ···· **413p**

Daffa Raygama (Universitas Telkom, Indonesia); Jati H. Husen and Villy Satria Praditha (Telkom University, Indonesia); Yijun Lu (Waseda University, Japan); Adam Rafif Faqih (Telkom University, Indonesia)

Comparative Analysis of LSTM and ARIMA Models for Stock Price Prediction in Sharpe Ratio-Based Portfolio Construction ····· **420p**

Dwi Fitrizal Salim (Telkom University, Indonesia)

An Efficient Hybrid Transformer-XGBoost Model for Improved Accuracy in Thermal-Based Rice Seed Infection Detection ····· **426p**

Ani Dijah Rahajoe, I Gede Susrama Mas Diyasa, Basuki Rahmat and Agussalim Agussalim (Universitas Pembangunan Nasional Veteran Jawa Timur, Indonesia)

Multi-Task IndoBERT for Depression Detection: Intent, Emotion, and PHQ-9 Classification in Indonesian Chatbot ····· **431p**

Rabiatun Nisa, Bimo Sunarfri Hantono and Indriana Hidayah (Universitas Gadjah Mada, Indonesia); Fadhilla Najmi Qinthara (Universitas Padjajaran, Indonesia)

Real-Time Protection Scheme for Embedded Systems Using a Slack Manager ····· **437p**

Kun Woo Park (Sangmyung University, Korea (South)); Ingeol Chun (Electronics and Telecommunications Research Institute, Korea (South)); Daehee Seo (Sangmyung University, Korea (South))

P2: Poster Session

Hippocampal-Inspired Memory-Enhanced YOLO for Context-Aware Urban Crowd DensityAnalysis ····· **443p**

Hye-Youn Lim (Dong-A University, Korea (South)); Ming-Shou An and Dae-Seong Kang (Dong-A University)

Design and Implementation of Multi-Level Voice Phishing Detection Model Based on LLM ····· **449p**

DoHyeun Kim, Sihoon Kim, Syed Shehryar Ali Naqvi, Kwangil Jeon and Sohyun Beak (Jeju National University, Korea (South))

Dental Condition Diagnosis Using U-Mamba ····· **455p**

Hyunsik Ahn (Tongmyoung University, Korea (South))

Enhanced Quantum-Based Mining Strategy for Bitcoin ····· **458p**

Yeonsong Suh and Younghoon Park (Sookmyung Women's University, Korea (South)); Kwangman Ko (Sangji University, Korea (South))

Real-Time Object Distance Estimation and LLM-Based Surrounding Scene Description · **462p**

DoHyeun Kim, Misbah Bibi, Sihoon Kim and Munho Kang (Jeju National University, Korea (South))

Design Challenges of on-Device AI for Maritime Communication Systems ····· **467p**

Shrutika Sinha and Soo-Hyun Park (Kookmin University, Korea (South))

A Proposal of Adaptive Threshold Determination Model for Coastal Fishing Vessel Safety Index Using Reinforcement Learning from Human Feedback ····· **471p**

Sun-Ho Yum, Ari Hwang and Soo-Hyun Park (Kookmin University, Korea (South))

A Self-Supervised Energy Flow Signature-Based Fault Detection Framework for DER Systems ····· **476p**

DoHyeun Kim, Murad Ali Khan, Syed Shehryar Ali Naqvi, Misbah Bibi and Faiza Qayyum (Jeju National University, Korea (South))

An Analysis of Distributed Energy Resources Ecosystem for Efficient Surplus Energy Trading and Distribution ····· **482p**

DoHyeun Kim, Syed Shehryar Ali Naqvi, Muhammad Adnan Ejaz, Syed Ali Yazdan and Misbah Bibi (Jeju National University, Korea (South))

Exploiting User-Intent Priors for Robust Road Damage Categorization in Citizen-Reported Imagery ····· **488p**

Gyeyeong Kim, Wonjun Jeong, Seunghun Song, Junseo Ok and Duckki Lee (Yonam Institute of Technology, Korea (South))

The Industrial Internet of Things (IIoT) and Edge Computing: Key Drivers of Industrial Digital Transformation ····· **494p**

Ala Khalifeh (German University of Jordan, Jordan); Muath Alhwetat (German Jordanian University, Jordan); Khalid A. Darabkh (The University of Jordan, Jordan); Iacovos Ioannou (European University of Cyprus, Cyprus & University of Cyprus and CYENS Centre of Excellence, Cyprus)

Comparative Evaluation of Machine Learning Algorithms for Smartphone-Based Human Activity Recognition ····· **500p**

Ala Khalifeh (German University of Jordan, Jordan); Hamza Damra (German Jordanian University, Jordan); Iacovos Ioannou (European University of Cyprus, Cyprus & University of Cyprus and CYENS Centre of Excellence, Cyprus); Khalid A. Darabkh (The University of Jordan, Jordan)

IoT Mesh Networking: Performance Comparison of Z-Wave, Zigbee, BLE, and Thread with a Generic Mathematical Model ····· **506p**

Ala Khalifeh (German University of Jordan, Jordan); Hamza Damra (German Jordanian University, Jordan); Iacovos Ioannou (European University of Cyprus, Cyprus & University of Cyprus and CYENS Centre of Excellence, Cyprus); Khalid A. Darabkh (The University of Jordan, Jordan)

AI-Driven Sleep Apnea Screening Based on Wireless Abdominal Wearable Sensor ····· **511p**

Thi Hang Dang (Ulsan National Institute of Science and Technology, Korea (South)); Sukmin Lee (SB Solutions Inc., Korea (South)); Seungcheon Kim (Hansung University, Korea (South)); Franklin Bien (Ulsan National Institute of Science and Technology, Korea (South))

Lightweight AI-Generated Text Detection Under Domain Ambiguity via Latent Lexical Domains ····· **515p**

Jisoo Lee (Korea (South)); Skyler Chun (Ridgewood Highschool, USA); Hyemyung Lee (Kookje University, Korea (South)); Nam Sang-Yep (KookJe, Korea (South))

IoT-Based Wearable Device for Real-Time Sitting Posture Monitoring and Early Kyphosis Prevention

1st Eka Septia Amini *, 2nd Sevia Indah Purnama †, 3rd Fikra Titan Syifa ‡
* *Department of Biomedical Engineering, Telkom University, Purwokerto 53147, Jawa Tengah, Indonesia*
† *Department of Biomedical Engineering, Telkom University, Purwokerto 53147, Jawa Tengah, Indonesia*
‡ *Department of Electrical Engineering, Telkom University, Purwokerto 53147, Jawa Tengah, Indonesia*
Emails: *1st ekaseptiaamini@student.telkomuniversity.ac.id, 2nd seviaindah@telkomuniversity.ac.id, 3rd fikra@ittelkom-pwt.ac.id*

***Abstract*—** Prolonged Poor sitting posture over long periods can lead to musculoskeletal disorders and increase the risk of postural kyphosis, particularly among students and office workers who engage in extensive digital activities. Continuous posture monitoring and immediate corrective feedback are required to prevent spinal deformities at an early stage. This research proposes an Internet of Things–based wearable system for real-time monitoring of sitting posture as an early prevention effort for kyphosis. The system utilizes an MPU6050 sensor and an ESP32 microcontroller to measure trunk inclination and transmit angle data to the Blynk platform. When the inclination exceeds the 40-degree threshold, the buzzer provides an auditory alert, while real-time visualization is displayed through the Blynk dashboard. Testing results show that the system achieves an accuracy level of approximately 98 percent based on comparison with a clinometer, and the average alert response time is 0.2 seconds. The system has proven effective in increasing posture awareness and demonstrates strong potential as an ergonomic solution to reduce the long-term risk of spinal disorders.

***Keywords*—** *IoT, wearable, posture monitoring, kyphosis prevention, real-time feedback*

I. INTRODUCTION

Prolonged sitting has become an increasingly prevalent issue due to the rapid growth of digital-based activities in workplaces, schools, and daily routines. Long-duration sitting with poor posture is strongly associated with musculoskeletal problems, particularly back pain, reduced physical performance, and long-term spinal deformities such as postural kyphosis [1]. Postural kyphosis is characterized by an excessive forward curvature of the upper thoracic spine, leading to a hunched back posture. The global impact is severe: Musculoskeletal Disorders (MSDs) are the second most common cause of disability globally, with the disability burden increasing by 45% over the past decade [2]. Furthermore, the World Health Organization (WHO) reports that sedentary lifestyles significantly contribute to various spinal issues, estimating that 250,000 to 500,000 new spinal disorders occur each year globally [3]. This significant health threat underscores the urgent need for effective intervention strategies.

The need for continuous monitoring has driven the exploration of wearable technologies in posture assessment due to their portability and ability to deliver real-time feedback. Previous studies have successfully demonstrated the feasibility of low-cost, IoT-based wearable systems. For example, research by Supriyanto et al. developed a smart vest prototype with an MPU6050 sensor to detect non-ergonomic sitting posture [4], while Setyawan and Wicaksono designed a compact posture alarm using NodeMCU and a flexible sensor to provide immediate feedback on deviations [5]. These efforts establish the potential for developing integrated systems that effectively alert users to poor posture, forming the foundation for the current study.

Building upon the promise of wearable solutions, the advancement of the Internet of Things (IoT) provides the necessary infrastructure for efficient, cloud-connected health monitoring. IoT systems enable continuous posture evaluation and instant feedback delivery via mobile applications. The Blynk platform is particularly suitable for this application, as it facilitates seamless and straightforward integration between the sensor hardware and smartphone interface [6]. This capability allows the system to not only collect data but also to offer real-time data visualization, which is essential for user engagement and enabling immediate, self-driven posture correction.

This paper proposes an IoT-enabled wearable system for real-time monitoring of sitting posture, addressing the critical need for early kyphosis prevention. The system utilizes an MPU6050 sensor and an ESP32 microcontroller to accurately measure and transmit posture angle data. The solution is designed to provide immediate warnings through an audible buzzer alert for incorrect posture and simultaneously display the quantitative posture data on the Blynk dashboard. By providing dual feedback mechanisms, the proposed system aims to significantly enhance user awareness and facilitate instant posture correction, thus supporting the early prevention of postural kyphosis.

II. RELATED WORK

A. Research Gap and Contribution

Previous posture-monitoring studies have used Load Cell and ultrasonic sensors [7], IMU-based classification models [4][8], flexible-sensor alarms [10], Bluetooth smart vests, and camera-based detection such as MediaPipe and Haar Cascade [9][11], yet these systems still show key limitations. Many are not wearable, rely on controlled visual environments, lack IoT connectivity, or do not provide real-time corrective alerts. Even recent IoT-based prototypes, such as the ESP32–MPU6050 wearable by Ziadeh [12], do not include early-kyphosis detection thresholds (≈40°) or integrated mobile dashboards. To address these gaps, this work introduces an IoT-enabled wearable system combining the MPU6050, ESP32, Blynk visualization, and buzzer alerts to deliver real-time posture angle monitoring and immediate feedback for early kyphosis prevention.

B. IoT and Sensor Technologies for Posture Monitoring

Advances in IoT have significantly improved remote and continuous health monitoring. The ESP32 microcontroller is widely utilized due to its integrated Wi-Fi, low power usage, and suitability for wearable systems. Several studies report its efficiency in transmitting sensor data to cloud platforms for real-time observation [5]. The MPU6050 IMU sensor is frequently applied in posture and motion-tracking applications because it combines a 3-axis accelerometer and 3-axis gyroscope. Previous studies show that IMU-based sensing provides reliable accuracy for estimating trunk inclination and detecting early postural deviations [14][15].

C. Blynk Platform for Remote Health Monitoring

Blynk has been widely used in IoT health applications due to its cloud-based architecture, real-time dashboard, and seamless integration with ESP32 and related microcontrollers [13]. Previous IoT healthcare systems have successfully used Blynk to visualize vital signs such as temperature, heart rate, and fall detection, demonstrating its reliability for low-latency data transmission and user-friendly mobile interfaces. However, posture-monitoring implementations using Blynk remain limited. Most existing prototypes only stream raw sensor data without contextual processing such as angle computation, posture classification, or early kyphosis warning thresholds. Studies such as Ziadeh's implementation confirmed the feasibility of Blynk for wearable posture monitoring, but did not provide detailed real-time analysis or corrective feedback mechanisms [12].

III. METHODOLOGY

A. System Overview

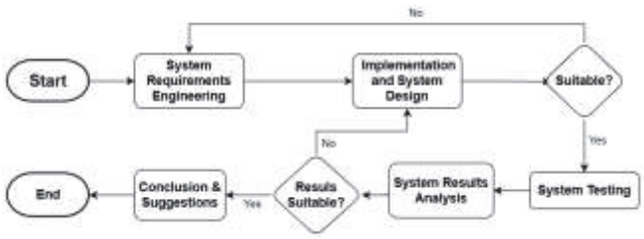


Figure 1. Block Diagram for Proposed Methodology

Figure 1 illustrates the flowchart that outlines the overall development process of the system. The workflow describes the steps involved in designing and implementing a wearable device capable of monitoring sitting posture using the MPU6050 sensor module integrated with the Blynk IoT platform.

B. Software System

Software development for this system encompasses the firmware programming process, sensor data processing, and integration with the IoT platform. The ESP32 firmware was written using the Arduino IDE, utilizing the Wire, MPU6050, and Blynk libraries to support sensor communication and internet connectivity. Raw accelerometer and gyroscope data from the MPU6050 are processed into a pitch angle value, which represents the degree of trunk inclination (forward slouching). The angle calculation utilizes a combination of X-Z axis acceleration data and Y-axis rotation to obtain stable readings. The resulting calculated angle is then transmitted to the Blynk Cloud via a Wi-Fi connection, where the data is displayed in real-time on the Blynk mobile application using widgets such as the gauge and value display. The Blynk platform also functions as the monitoring interface, enabling users to view instantaneous angle changes directly on their mobile devices.

C. Hardware System

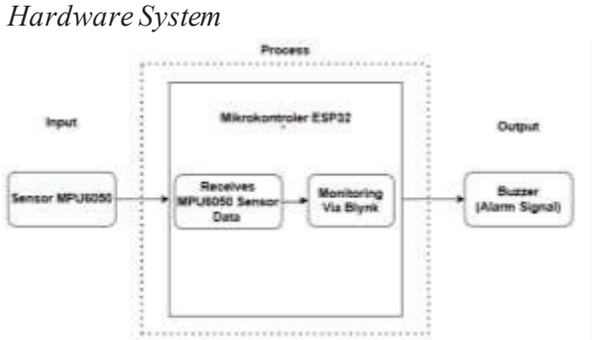
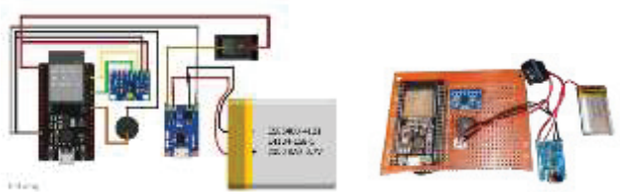


Figure 2. Block Diagram Hardware System



(a) (b)

Figure 3. (a) Hardware Schematic, (b) Hardware Appearance

The hardware system comprises the MPU6050 sensor module, the ESP32 microcontroller, a buzzer for the audible alert, and supporting components such as the power source and wiring connections. The MPU6050 serves as the primary sensor, detecting changes in the body's orientation based on acceleration and angular velocity. The sensor communicates with the ESP32 via the I2C communication protocol, utilizing the SDA and SCL pins. The ESP32 acts as the central data processing unit, responsible for calculating the inclination angle and transmitting the processed data to the IoT platform. The buzzer is connected to GPIO Pin 18 to generate an audible alert when the calculated inclination angle exceeds the defined threshold of 40°. All components are integrated into a wearable prototype (vest), which facilitates direct and continuous posture monitoring while the user is seated.

IV. Results and Discussion

A. *System Implementation*

The developed wearable prototype consists of a vest equipped with an MPU6050 sensor module, an ESP32 microcontroller, a buzzer, and supporting components. The system was tested on users seated on a chair to monitor posture in real-time. The ESP32 processes accelerometer and gyroscope data from the MPU6050, calculates the pitch angle, and transmits the data to the Blynk platform via Wi-Fi. Blynk displays the pitch value in real-time along with the posture status (Normal or Alert), while the buzzer provides an audible warning when the pitch angle exceeds the threshold of 40°.

B. *Testing Results*

1. MPU6050 Sensor Testing

To validate the sensor's accuracy, pitch angles measured by the MPU6050 were compared with reference values from a clinometer. The comparison focuses solely on the pitch angles and their differences. The accuracy percentage was calculated using the formula:

$$Accuracy\ (\%) = \left(1 - \frac{|Pitch\ MPU6050 - Pitch\ Reference|}{Pitch\ Reference}\right) \times 100\%$$

Where Pitch MPU6050 is the sensor reading and Pitch Reference is the value measured by the clinometer. The results are presented in Table 1 for clarity, showing the clinometer reference, MPU6050 readings, difference, and calculated accuracy percentage.

TABLE 1. MPU6050 SENSOR ACCURACY TESTING AGAINST CLINOMETER

No	Clino-meter (°)	MPU6050 (°)	Difference (°)	Accuracy (%)
1	0	0.5	0.5	99.0
2	10	10.2	0.2	98.0
3	30	30.7	0.7	97.7
4	40	40.3	0.3	99.2
5	60	60.5	0.5	99.1

2. Buzzer Function Testing

The buzzer was tested to verify its responsiveness when the pitch angle exceeded the threshold of 40°. The system successfully activated the buzzer every time the threshold was crossed, with an average response delay of 0.2 seconds. This confirms that the buzzer can reliably alert users of slouched postures that may increase the risk of kyphosis.

TABLE 2. BUZZER FUNCTION TESTING

Pitch Condition (°)	Buzzer Active	Response Delay (s)	Notes
≤40	No	–	Normal posture
>40	Yes	0.2	Consistently active

In the Blynk Web Dashboard display, the sitting posture monitoring system successfully shows the pitch angle in real time. The gauge widget indicates a value of 63°, which in the system's classification represents a severely slouched and unhealthy posture. This value triggers an alert status, informing the user that immediate posture correction is required to prevent excessive strain on the spine and back muscles.



Figure 4. Blynk Web Dashboard displaying a pitch angle of 63°

In the Blynk Web Dashboard display, the sitting posture monitoring system successfully shows the pitch angle in

real time. The gauge widget indicates a value of 63°, which in the system's classification represents a severely slouched and unhealthy posture. This value triggers an alert status, informing the user that immediate posture correction is required to prevent excessive strain on the spine and back muscles.

3. Blynk Application Testing

The Blynk platform was tested for real-time monitoring and visualization of pitch angles and posture status. Data was transmitted via Wi-Fi with an update interval of 1 second. The application successfully displayed pitch values and posture classifications without noticeable lag, demonstrating its reliability as a continuous monitoring interface.

TABLE 3. BLYNK APPLICATION TESTING

Parameter	Test Result	Notes
Pitch Update Interval	1 second	According to system setting
Pitch & Status Visualization	Real-time, accurate	Facilitates posture monitoring
Wi-Fi Connection	Stable	Data transferred without lag

C. Discussion

The testing results indicate that the developed wearable system is effective in monitoring sitting posture in real time. The MPU6050 sensor accurately detects changes in trunk inclination, providing detailed pitch measurements at one-degree resolution, and validation using a clinometer demonstrates high measurement accuracy. The buzzer responds promptly when the pitch angle exceeds 40°, with an average response delay of only 0.2 seconds. Integration with the Blynk application enables users to monitor pitch values and posture status in real time, supported by stable Wi-Fi connectivity and uninterrupted data transmission. Minor fluctuations in pitch readings were observed, likely caused by small involuntary body movements or sensor noise, highlighting the importance of proper sensor placement and calibration to ensure reliable measurements. Overall, the system offers a practical solution for early detection of kyphosis risk and promotes corrective actions to maintain healthy sitting posture and reduce the likelihood of spinal disorders.

Table 4. Visual Sitting Posture Analysis

Image	Posture Description	Pitch Category
	Neutral / Good posture	<20° (Low)
	Mild forward bending	20–40° (Moderate)
	Severe forward bending	>40° (High, warning)

ACKNOWLEDGMENT

The authors sincerely thank the supervising lecturers for their guidance and invaluable advice throughout the development of the IoT-enabled wearable for real-time monitoring of sitting posture and early kyphosis prevention. Special thanks are also extended to colleagues and friends for their support and collaboration during testing and data collection, as well as to family members for their continuous encouragement and motivation.

REFERENCES

[1] Hartanto, "Pengaruh Postur Duduk Membungkuk dalam Menimbulkan Kifosis Postural," *Jurnal Kedokteran Meditek*, vol. 28, no. 2, pp. 207–214, May 2022, doi: 10.36452/jkdoktmeditek.v28i2.2331.

[2] Annisa Hapsari, "Gangguan Muskuloskeletal (Tulang, Sendi, Otot)," hellosehat. https://hellosehat.com/muskuloskeletal/gangguan-muskuloskeletal/

[3] WHO, "Musculoskeletal health," WHO. https://www.who.int/news-room/fact-sheets/detail/musculoskeletal-conditions

[4] U. Brawijaya, Y. Dwi Kurnia, D. Syauqy, and E. R. Widasari, "Fakultas Ilmu Komputer Pengembangan Wearable Device Untuk Deteksi Postur Duduk Manusia Berbasis Data Sensor Mpu6050 Menggunakan Metode Support Vector Machine," 2017. [Online]. Available: http://j-ptiik.ub.ac.id

[5] O. : Candra, D. Setyawan, and A. Wicaksono, "Universitas Muhammadiyah Sidoarjo Prototype Alarm Posisi Postur Tubuh Berbasis NodeMCU ESP8266."

[6] BLYNK, "Blynk: a low-code IoT software platform for businesses and developers," https://blynk.io/.

[7] L. Alfarizy, D. Syauqy, and R. S. Perdana, "Sistem Monitoring Postur Duduk pada Pemain Game Online menggunakan Sensor Load Cell dan Ultrasonik dengan Metode K-Nearest Neighbor

(KNN) berbasis Arduino," 2023. [Online]. Available: http://j-ptiik.ub.ac.id.OP
[8] C. Qolby, R. Rizaputri, D. Syauqy, and R. Maulana, "Sistem Monitoring Postur Tubuh Lansia berbasis Wearable Embedded System dan Metode Klasifikasi Naïve Bayes," 2022. [Online]. Available: http://j-ptiik.ub.ac.id
[9] J. Perintis *et al.*, "Prosiding Seminar Ilmiah Sistem Informasi Dan Teknologi Informasi Pusat Penelitian dan Pengabdian Pada Masyarakat (P3M) Universitas Dipa Makassar Sistem Deteksi Postur Duduk Berbasis MediaPipe untuk Meningkatkan Ergonomi dan Kesehatan Pekerja."
[10] C. D. Setyawan and A. Wisaksono, "Body Posture Position Alarm Prototype Based on NodeMCU ESP8266," *Buletin Ilmiah Sarjana Teknik Elektro*, vol. 5, no. 4, pp. 614–622, Jan. 2024, doi: 10.12928/biste.v5i4.9543.
[11] N. Azizah, M. M. Parenreng, A. Gunawan, J. T. Elektro, P. Negeri, and U. Pandang, "Prosiding Seminar Nasional Teknik Elektro dan Informatika (SNTEI) 2022-Teknik Informatika."
[12] M. Ziadeh, "IoT-Based Posture Detection and Feedback System: A Wearable Solution for Maintaining Proper Posture." [Online]. Available: https://www.espressif.com.
[13] P. Utami Rakhmawati and K. Kunci, "Analisis Komunikasi Platform Internet of Things Aplikasi Blynk," vol. 9, p. 2024, 2024.
[14] U. Brawijaya, Y. Dwi Kurnia, D. Syauqy, and E. R. Widasari, "Fakultas Ilmu Komputer Pengembangan Wearable Device Untuk Deteksi Postur Duduk Manusia Berbasis Data Sensor Mpu6050 Menggunakan Metode Support Vector Machine," 2017. [Online]. Available: http://j-ptiik.ub.ac.id.
[15] C. D. Setyawan and A. Wisaksono, "Body Posture Position Alarm Prototype Based on NodeMCU ESP8266," *Buletin Ilmiah Sarjana Teknik Elektro*, vol. 5, no. 4, pp. 614–622, Jan. 2024, doi: 10.12928/biste.v5i4.9543.

Bandwidth Improvement in a Compact 3D-Printed Conical Patch Antenna by Substrate Profiling

Chinnam Manasa*, Damera Vakula† Department of ECE,
National Institute of Technology Warangal, India
Email: *cm22ecr2r01@student.nitw.ac.in, †vakula@nitw.ac.in

***Abstract*—Microstrip patches of antennas are widely used in many different applications, antenna engineers still face a significant challenge in meeting the growing demand for small, high performance designs. For 5G applications, a new small, high gain, conical-shaped 3D printed antenna with a wider bandwidth is introduced. The proposed antenna consists of two substrate layers layered on top of one another. The square substrate geometry is covered with a conical-shaped PLA substrate. Copper tape is applied to the raised portion of the conical substrate, which operates at 3.5GHz. The non-uniform rise in substrate height results in stable radiation patterns and greatly increases bandwidth.**

***Index Terms*—enhanced bandwidth, high gain, conical shaped, 3D printing, MIMO, 5G.**

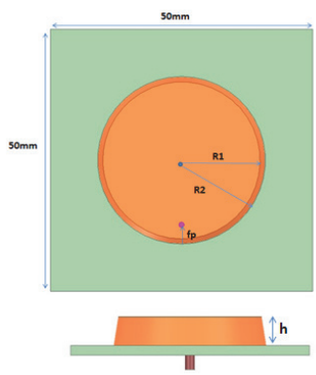


Fig. 1: Dimensional sketch of the proposed antenna

I. Introduction

Wireless connectivity has continuously improved as mobile networks have advanced from 1G to 4G, and 5G is currently the next big advancement. Due to this new generation's heavy reliance on sub-6 GHz frequency bands, system architecture and antenna design requirements have significantly changed [4]. Since microstrip patch antennas can be used in rectangular, circular, elliptical, and triangular geometries, they are generally regarded as appropriate for these systems. Their tiny structure, cheap production cost, and aptitude to function efficiently at millimeter-wave bands make them suitable for 5G. However, choosing the right substrate materials and meticulously tuning the antenna design are necessary to achieve excellent performance.

A number of techniques have been put out to increase patch antenna bandwidth. For instance, [1] presented a triple-band design based on an inverted-L monopole that was modified by adding two complimentary split-ring resonators to the ground plane. These technologies provide multi-band operation, but their usage in bigger platforms is limited since they often create end-fire radiation and are placed within small device frames. Dual-band microstrip antennas for terminal devices were introduced in [2], [3], where the bandwidth in the N78 and N79 bands was increased by using parasitic patches. Bandwidth of a microstrip antenna can also be achieved by etching slots on the patch surface to separate the radiator into a primary and a parasitic component [4]. To modify the resonance frequencies and increase the operating range, shorting pins can be positioned correctly. In a similar vein, [5] presented a dual-band setup for 5G applications utilizing a multilayer substrate and a U-shaped slot.

Gain enhancement is another important component that influences the dependability and coverage of wireless networks. Researchers have investigated many approaches to solve this, including antenna arrays [6]–[8], metasurface-based enhancements, cavity-backed techniques [9]–[12], and parasitic element loading [13], [14]. While these tactics can result in better gains, they frequently involve additional complexity, increase physical dimensions, or necessitate fine adjustment. The design described in [15] increased gain by inserting four shorting pins symmetrically along the diagonals of a square patch. In [16], a dual-band, dual-polarized 5G antenna was built using substrate-integrated suspended line (SISL) technology. The antenna features a driven patch, π-shaped slots, stacked layers, and parasitic components to facilitate high gain and dual-band operation. Despite these advancements, developing a compact, structurally simple patch antenna that both provides wide bandwidth and good gain remains a difficult research topic.

Additive manufacturing has lately attracted attention as an efficient way for boosting patch antenna performance, particularly by enabling substrate height variations and complicated under-patch designs [17], [18]. Fused Deposition Modeling (FDM) is one of the most cost-effective and extensively utilized 3D-printing techniques, making it ideal for developing antenna prototypes for microwave and lower millimeter-wave applications. Previous demonstrations using conductive thermoplastic filaments such as Electrifi include printed patch antennas and compact pyramidal horn structures, showing gains of approximately 3 dBi at 2.5 GHz and 14.3 dBi at 5.8 GHz; however, a loss of nearly 0.7 dB was observed due

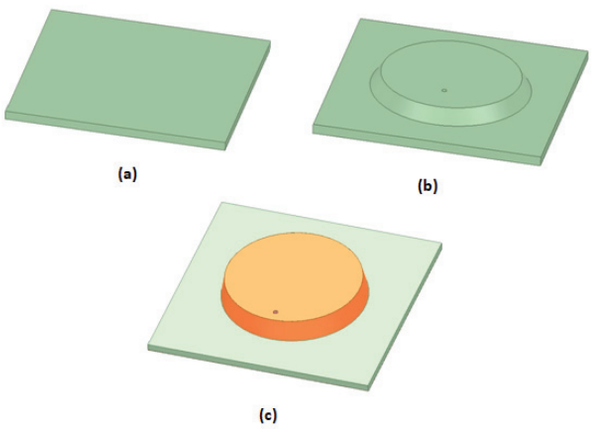


Fig. 2: Evolution of the suggested conical-shaped 3D printed patch antenna.

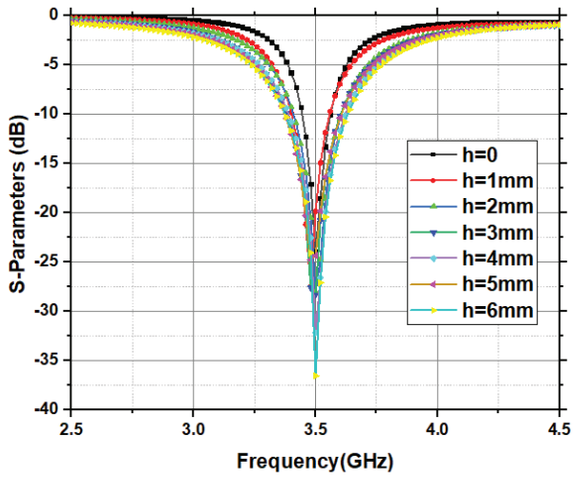


Fig. 3: Simulated reflection coefficient of the proposed antenna.

to surface-roughness effects [19]. In another instance, a multi-ring dielectric resonator antenna (DRA) produced using FDM obtained an impedance bandwidth of 4.3-8.0 GHz, realized gains between 0.65 and 2.45 dBi, and an average efficiency near 89% [20]. FDM has also enabled the production of gradient-index (GRIN) dielectric lenses, which generated significant aperture increases ranging from 15.1 to 23.6 dBi over the 12-40 GHz range with efficiencies around 82% [21].

The versatility of FDM has been further demonstrated in reflector antennas and space communication hardware. For example, a compact parabolic reflector designed for 24GHz radar systems reported a gain of 18.3 dBi across 23.5–25.5 GHz [22]. Likewise, a curved CubeSat patch antenna operating at 2.16 GHz yielded a gain of 6.7 dBi along with approximately 96% radiation efficiency [23]. Leaky-wave spiral antennas printed using PREPERM ABS materials reached a peak gain of 25 dBi at 18 GHz with a narrow 6° beamwidth [24]. Additionally, conformal rectenna structures intended for biomedical and IoT applications demonstrated dual-band operation at 2.4 GHz and 5.2 GHz, with gains ranging from 5.7 to 8.8 dBi and corresponding RF-to-DC conversion efficiencies of 67.3% and 42.2% [25]. Overall, FDM proves highly effective for dielectric-dominated components such as DRAs and GRIN lenses [26], [27], while its use for metallic radiators remains limited due to the low conductivity of available filaments and challenges associated with surface roughness.

This paper describes a conical-shaped D printed antenna for 5G base station applications. The design uses a conical-shaped substrate placed on a flat, square-shaped foundation. This entire substrate may be produced utilizing FDM-3D printing technology. Copper tape is applied to the top surface of the conical geometry to make the radiating element, and copper tape is applied beneath the square substrate to form the ground plane. This structural solution not only increases bandwidth but also maintains a small footprint. Adjusting critical structural elements allows you to fine-tune the antenna's operating frequency.

II. Design Configuration and Functional Overview

The detailed structure of the antenna is shown in Fig. 1. The antenna uses a PLA substrate with $\varepsilon_r = 2.25$. Initially, a square-shaped flat substrate layer with dimensions $50 \times 50 \times 2\ \mathrm{mm}^3$ is used.

A cone-shaped substrate layer with inner radius R1, outer radius R2, and thickness h is put on top of the flat substrate layer. Finally, an octagonal-shaped substrate layer with radius R2 and thickness h is anchored to the second substrate layer. The whole substrate geometry can be produced by the FDM (Fused Deposition Modeling) 3D printing process. The radiating patch is now created by applying a conductive copper tape with a thickness of 0.035mm to the top and sides of the two octagonal substrate layers. A coaxial probe feed is utilized to feed the planned antenna at a distance of fp from the center.

TABLE I: Effect of Substrate Height on Antenna Parameters

h (mm)	R2 (mm)	fp (mm)	Bandwidth (MHz)	Bandwidth (%)
0 (Flat)	19	12.5	120	3.42
1	18.3	11.8	150	4.28
2	17.8	10.3	210	6
3	17.3	9.2	230	6.6
4	16.7	8.3	250	7.14
5	16.1	7.5	310	8.8
6	15.8	6.6	340	9.7

A parametric study is performed for h and the results are presented in this section.

Initially, in the first step, a square-shaped substrate with height 2mm, as shown in Fig. 2. The second stage of the antenna is created by adding another conical-shaped substrate layer to the square-shaped substrate with height h. To complete the radiating patch, copper tape is glued to the entire elevated octagonal substrate.

In Fig. 3, the suggested antenna's reflection coefficient is reported for different values of height h. It can be seen that the elevated substrate with the conical patch outer radius $R2$ shifts the resonance frequency. To obtain 3.5GHz resonant

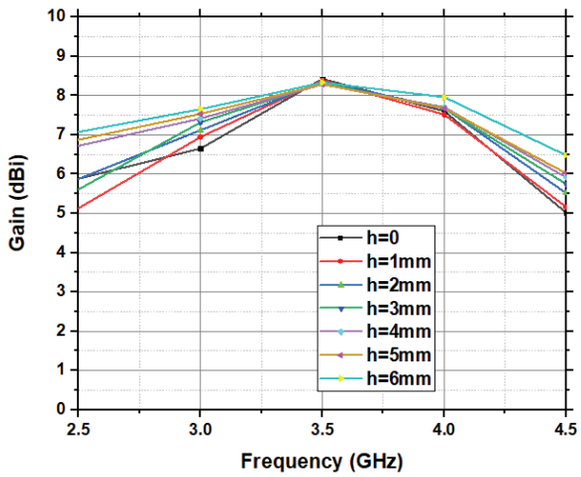


Fig. 4: Simulated gain of proposed antenna at various values of h

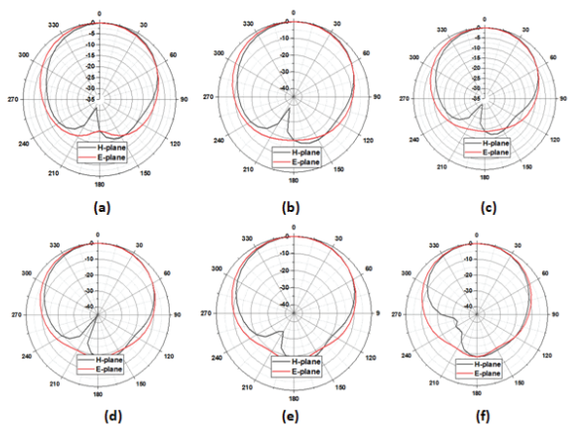


Fig. 5: Far-field radiation patterns of the proposed antenna at 3.5GHz (a) h=0, (b) h=1mm, (c) h=2mm, (d) h=3mm, (e) h=4mm, (f) h=5mm, (g) h=6mm

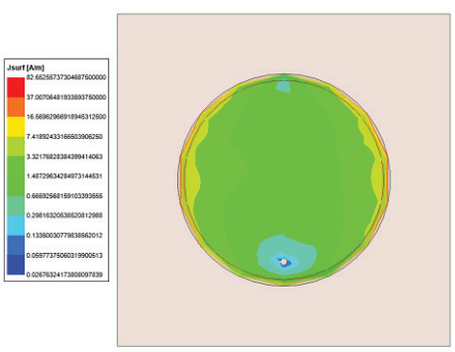


Fig. 6: Surface current distributions of the proposed antenna when at 3.5GHz

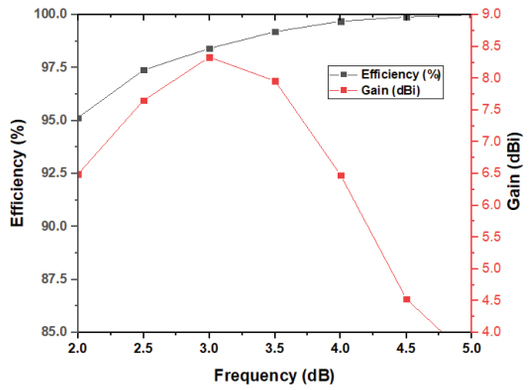


Fig. 7: Gain and Efficiency of the proposed antenna.

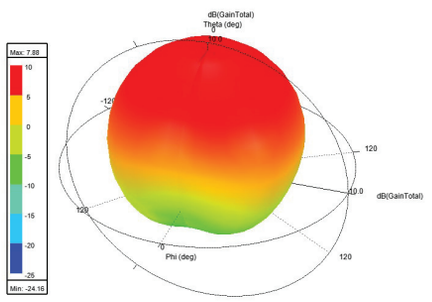


Fig. 8: 3D polar plot of the proposed antenna

frequency, the radius of the patch $R2$ must be reduced. As a result, the antenna becomes more compact. To achieve the appropriate frequency response at 3.5 GHz, the elevated conical substrate's radius decreases as its height increases. Figure 4 depicts the simulated gain of the proposed antenna for various h values. Gain and efficiency of the proposed antenna are demonstrated in Fig. 7. The bandwidth increases with the height of the substrate.However, if the substrate is too thick, particularly if it has a high dielectric constant, it may excite undesired surface waves, reducing radiation efficiency and distorting the pattern. So, rather of immediately increasing the overall substrate height, only a portion of the substrate beneath the radiating patch is raised. Table I summarizes the primary performance advantages of the proposed antenna over conventional flat circular patch antennas.

Fig. 6 shows the surface current distribution and Fig. 8 shows the far field radiation 3D polar plot of the proposed antenna at 3.5GHz.

III. Conclusion

This paper presents a low-cost, compact, high-gain conical shaped 3D patch antenna for 5G applications. The concept uses conical shaped substrate for bandwidth enhancement, operates at 3.5 with significant gain of 8.2 dBi. The proposed antenna is a good candidate for mid-band 5G applications. Further modifications in the substrate geometry and fabrication using FDM 3D printing makes it compatible with 5G/6G base-station arrays/MIMO.

REFERENCES

[1] Yousfi A. E., Lamkaddem, A. Abdalmalak, K. A. Segovia-Vargas, D. A miniaturized triple-band and dual-polarized monopole antenna based on a CSRR perturbed ground plane. IEEE Access. 9, 164292–164299 ,2021.

[2] Chang L., Liu, H. Low-profile and miniaturized dual-band microstrip patch antenna for 5G mobile terminals. IEEE Trans. Antennas Propag. 70, 2328–2333 ,2023.

[3] Chen X., Wang, J. Chang, L. Extremely low-profile dual-band microstrip patch antenna using electric coupling for 5G mobile terminal applications. IEEE Trans. Antennas Propag. 71, 1895–1900 ,2023.

[4] Lu Y. et al. Extremely low-profile micro dual-band patch antenna for 5G mobile terminals. IEEE Antennas Wirel. Propag. Lett. 23, 2326–2330 ,2024.

[5] Wei M. et al. Low-profile dual-band microstrip patch antenna with shorting pins. IEEE Antennas Wirel. Propag. Lett. 23, 1834– 1838 ,2024.

[6] J. Hao, N. Yan, Y. Luo, H. Fu and K. Ma, "A Low-Cost Dual-Band Multimode High-Gain Stacked-Patch Antenna Based on SISL for 5G Applications," in IEEE Antennas and Wireless Propagation Letters, vol. 21, no. 1, pp. 4-8, Jan. 2022.

[7] H. Wang, K. E. Kedze and I. Park, "A High-Gain and Wideband Series-Fed Angled Printed Dipole Array Antenna," IEEE Trans. Antennas Propag.,vol. 68, no. 7, pp. 5708-5713, July 2020.

[8] G. -H. Sun and H. Wong, "C-Shaped Open Slot Antenna Array for Millimeter-Wave Applications," IEEE Trans. Antennas Propag.,vol. 69, no. 12, pp. 8426-8435, Dec. 2021.

[9] A. Verma, M. Arrawatia and G. Kumar, "High Gain Wideband Circularly Polarized Microstrip Antenna Array," IEEE Trans. Antennas Propag.,vol. 70, no. 11, pp. 11183-11187, Nov. 2022.

[10] D. Chen, W. Yang, W. Che and Q. Xue, "Miniaturized Wideband Metasurface Antennas Using Cross-Layer Capacitive Loading," IEEE Antennas Wireless Propag. Lett.,vol. 21, no. 1, pp. 19-23, Jan. 2022.

[11] K. E. Kedze, H. Wang and I. Park, "A Metasurface-Based Wide-Bandwidth and High-Gain Circularly Polarized Patch Antenna," IEEE Trans. Antennas Propag.,vol. 70, no. 1, pp. 732-737, Jan. 2022.

[12] C. Zhao, X. Li, C. Sun, H. Huang and Z. Zhang, "Wideband Planar Sleeve Dipole Antenna With Back Cavity," IEEE Antennas Wireless Propag. Lett.,[14]vol. 19, no. 6, pp. 959-963, June 2020.

[13] K. Raha and K. P. Ray, "Broadband High Gain and Low Cross-Polarization Double Cavity-Backed Stacked Microstrip Antenna" IEEE Trans. Antennas Propag.,[15]vol. 70, no. 7, pp. 5902-5906, July 2022.

[14] Y. Cao et al, "Broadband and High-Gain Microstrip Patch Antenna Loaded With Parasitic Mushroom-Type Structure," IEEE Antennas Wire- less Propag. Lett.,vol. 18, no. 7, pp. 1405-1409, July 2019.

[15] T. Liu and Q. Feng, "Broadband and High-Gain Low-Profile Array Antenna Loaded With Parasitic Patches," IEEE Antennas Wireless Propag. Lett.,vol. 22, no. 10, pp. 2595-2599, Oct. 2023.

[16] X. Zhang and L. Zhu, "Gain-Enhanced Patch Antennas With Loading of Shorting Pins," IEEE Trans. Antennas Propag.,vol. 64, no. 8, pp. 3310-3318, Aug. 2016.

[17] N. Yan, H. Sun, Y. Li, Y. Luo and K. Ma, "A Low-Cost High-Gain Dual-Band Dual-Polarized Antenna Using Parasitic Patches Based on SISL for 5G Base Station Applications," in IEEE Antennas and Wireless Propagation Letters, 2025.

[18] Chinnam, Manasa, and Vakula Damera. "A wide band and high gain 3-D printed patch antenna for X-band applications." Frequenz 0 (2025).

[19] Chinnam, Manasa, and Vakula Damera. "An Open Book Shaped 3-D Patch Antenna Design for C Band Applications." 2024 IEEE Wireless Antenna and Microwave Symposium (WAMS). IEEE, 2024.

[20] E. Pizarro, R. Salazar, E. Rajo-Iglesias, M. Rodríguez, S. Fingerhuth, and G. Hermosilla, "Parametric study of 3D additive printing parameters using conductive filaments on microwave topologies," IEEE Access, vol. 7, pp. 106814–106823, 2019.

[21] Z.-X. Xia, K. W. Leung, and K. Lu, "3-D-printed wideband multi-ring dielectric resonator antenna," IEEE Antennas Wireless Propag. Lett., vol. 18, no. 10, pp. 2110–2114, Oct. 2019.

[22] S. Zhang, R. K. Arya, W. G. Whittow, D. Cadman, R. Mittra, and J. C. Vardaxoglou, "Ultra-wideband flat metamaterial GRIN lenses assisted with additive manufacturing technique," IEEE Trans. Antennas Propag., vol. 69, no. 7, pp. 3788–3799, Jul. 2021.

[23] J. R. Reis, C. Ribeiro, and R. F. S. Caldeirinha, "Compact 3D-printed reflector antenna for radar applications at K-band," IET Microw., Antennas Propag., vol. 15, no. 8, pp. 843–854, Jul. 2021.

[24] G. Muntoni, G. Montisci, A. Melis, M. B. Lodi, N. Curreli, M. Simone, G. Tedeschi, A. Fanti, T. Pisanu, I. Kriegel, A. Athanassiou, and G. Mazzarella, "A curved 3D-printed S-band patch antenna for plastic CubeSat," IEEE Open J. Antennas Propag., vol. 3, pp. 1351–1363, 2022.

[25] T. Lira-Valdés, E. Rajo-Iglesias, and F. Pizarro, "3-D-printed spiral leaky waveantenna with circular polarization," IEEE Open J. Antennas Propag., vol. 4, pp. 427–433, 2023.

[26] C. Yang and K. W. Leung, "3-D-Printed wideband circularly polarized MIMO dielectric resonator antenna," IEEE Trans. Antennas Propag., vol. 71, no. 7, pp. 5675–5683, Jul. 2023.

[27] D.D.Patil,K.S.Subramanian,andN.C.Pradhan,"3-D-printeddual-band rectenna system for green IoT application," IEEE Trans. Circuits Syst. II, Exp. Briefs, vol. 70, no. 8, pp. 2864–2868, 2023.

BER Analysis of Index Modulation in STAR RIS NOMA System

1st Mangali Shanti Priya*
2nd Anuradha Sundru†
* Dept. of ECE, NIT Warangal, Telangana, India
† Dept. of ECE, NIT Warangal, Telangana, India
ms24ecm3r22@student.nitw.ac.in, anuradha@nitw.ac.in

***Abstract*—Non-orthogonal multiple access (NOMA) and reconfigurable intelligent surfaces (RIS) are promising techniques to enhance spectral efficiency and connectivity in future wireless networks. Simultaneous transmitting and reflecting RIS (STAR-RIS) extends this concept by enabling full-space coverage, while index modulation (IM) introduces an additional dimension for data transmission by exploiting index domains. This paper investigates the performance of a STAR-RIS-assisted NOMA system with IM. Analytical and simulation results are presented in terms of bit error rate (BER) and transmission rate. The results show that integrating IM with STAR-RIS NOMA achieves superior spectral efficiency and reliability compared to conventional RIS-aided or NOMA only schemes, particularly under different fading environments.**

***Index Terms*—NOMA, IM, STAR RIS, CSI, BER, SNR.**

I. Introduction

Future wireless systems beyond 5G and 6G require very high spectral efficiency, massive connectivity, low latency, and reliable communication. Non-orthogonal multiple access (NOMA) supports these requirements by enabling power-domain multiplexing for multiple users over the same resources, improving spectral efficiency and user fairness [1]. Index Modulation (IM) further enhances efficiency by transmitting information through both modulation symbols and activation indices without additional bandwidth or power [2]. Reconfigurable Intelligent Surfaces (RIS) have emerged as an effective solution for controlling the wireless propagation environment through programmable reflections, enabling energy-efficient communication [3]. IM-NOMA schemes have also been explored to improve user separation and error performance [4]. Accurate channel estimation is essential for RIS-assisted systems, and several works have addressed channel estimation and performance analysis under different fading and coverage conditions [5], [6]. To enable full-space coverage, Simultaneously Transmitting and Reflecting RIS (STAR-RIS) has been proposed, offering enhanced flexibility in wave manipulation [7]. Practical limitations such as hardware impairments and phase errors have also been investigated [8], along with the effects of imperfect CSI on NOMA performance [9]. STAR-RIS has further shown capability for receiver-side spatial modulation [10]. Security, robust beamforming, and RIS partitioning have similarly been analyzed to improve reliability under realistic constraints [11]–[13], [15]. Advanced modulation approaches such as generalized quadrature spatial modulation have also been introduced for STAR-RIS–assisted NOMA networks [14]. Overall, combining IM with STAR-RIS-assisted NOMA offers a promising approach for future 6G systems by jointly enhancing spectral efficiency, coverage, and robustness under practical impairments.

II. System and Channel Models

A. STAR-RIS Assisted NOMA System

Because they allow for programmable propagation environments, reconfigurable intelligent surfaces (RIS) have become a crucial 6G technology. STAR-RIS elements can simultaneously transmit and reflect waves, offering full-space coverage and improved beamforming flexibility, in contrast to traditional RIS that only reflect signals.

We consider a downlink STAR-RIS–assisted NOMA system where a single-antenna access point serves two single-antenna users in the transmission and reflection

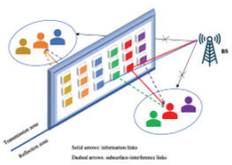


Fig. 1. STAR RIS NOMA

zones Figure 1 via an N-element STAR-RIS with independently controllable transmission and reflection coefficients. The NOMA transmit signal generated using superposition coding is:

$$x = \sqrt{\alpha_1 P}\, s_1 + \sqrt{\alpha_2 P}\, s_2, \tag{1}$$

where s_1 and s_2 are BPSK symbols of Users 1 and 2, respectively, P is the total transmit power, and power allocation factors satisfy $\alpha_1 + \alpha_2 = 1$ with $\alpha_1 > \alpha_2$, ensuring fairness for the weak user.

The received signals at the users are expressed as:

$$y_1 = \mathbf{h}_1^H \, \mathbf{\Theta}_T \, \mathbf{G} x + n_1, \tag{2}$$

$$y_2 = \mathbf{h}_2^H \, \mathbf{\Theta}_R \, \mathbf{G} x + n_2, \tag{3}$$

where:

- $\mathbf{G} \in \mathbb{C}^{N \times 1}$: AP-to-STAR-RIS channel,
- $\mathbf{h}_1, \mathbf{h}_2 \in \mathbb{C}^{N \times 1}$: STAR-RIS to user channels,
- $n_i \sim \mathcal{CN}(0, N_0)$: additive white Gaussian noise.

The STAR-RIS transmission and reflection coefficient matrices are:

$$\mathbf{\Theta}_T = \mathrm{diag}\left(\sqrt{\beta_{T,1}} e^{j\theta_{T,1}}, \ldots, \sqrt{\beta_{T,N}} e^{j\theta_{T,N}}\right), \tag{4}$$

$$\mathbf{\Theta}_R = \mathrm{diag}\left(\sqrt{\beta_{R,1}} e^{j\theta_{R,1}}, \ldots, \sqrt{\beta_{R,N}} e^{j\theta_{R,N}}\right), \tag{5}$$

where $\beta_{T,n} + \beta_{R,n} \leq 1$ ensures the conservation of energy.

B. IM Enhanced STAR-RIS-NOMA System

IM enhances spectral efficiency by transmitting additional information bits through the selection of indices, such as subcarriers, antennas, or RIS elements. In the proposed system, IM is implemented through subsurface activation of the STAR-RIS.

The STAR-RIS is partitioned into K subsurfaces, each with $\bar{N} = \lfloor N/K \rfloor$ elements. Let b_1 denote the index bits used to select one activation pattern $\Omega(i)$ among K possible subsurfaces. During each transmission interval, only one subsurface is activated, carrying IM bits.

The IM-assisted transmit signal becomes:

$$x_{\mathrm{IM}} = \sqrt{P}\left(\sqrt{\alpha_1}\, \mathbf{w}_1 s_1 + \sqrt{\alpha_2}\, \mathbf{w}_2 s_2\right), \tag{6}$$

where $\mathbf{w}_i$ is the activation vector corresponding to the selected subsurface, carrying index bits implicitly.

C. Transmitter Design

In the proposed IM-assisted STAR-RIS-NOMA system, the AP is responsible for jointly performing power-domain multiplexing and index-domain mapping. The goal is to exploit both modulation and index dimensions to further improve the spectral efficiency under low complexity constraints.

The AP processes the input bitstream through the following stages.

1) Bitstream Partitioning: The incoming data is divided into two parallel branches:

- **Index bits (b_1):** The index bits decide which STAR-RIS subsurface will be activated at every interval of transmission. Given that the STAR-RIS has K subsurfaces, $\log_2(K)$ index bits can be transmitted per channel use via subsurface selection alone.
- **Symbol bits (b_2):** The symbol bits are modulated based on conventional digital modulation schemes like BPSK, QPSK, or QAM. These bits carry user data over the traditional modulation domain.

This parallel mapping allows the proposed scheme to convey information simultaneously through two orthogonal domains: a signal constellation and a subsurface index domain. Parallel mappings like this increase throughput without the need for more bandwidth or increased power.

2) Subsurface Selection IM Mapping: The star-RIS is divided into K sub-surfaces and one subsurface is activated at every transmission. The activation pattern is determined by a predefined IM mapping function:

$$\Omega(i) = f_{\mathrm{IM}}(b_1), \quad i = 1, \ldots, K. \tag{7}$$

Only the subsurface corresponding to $\Omega(i)$ contributes to transmission, while the remaining subsurfaces remain inactive. This selection effectively embeds the index bits into the spatial domain of the STAR-RIS. Since the STAR-RIS influences both the transmission and reflection paths, IM offers further possibilities for cascaded channel shaping.

3) Superposition Coding (NOMA): NOMA relies on power-domain multiplexing to enable multiple users to share the same time–frequency resources. The AP obtains the final transmit signal as:

$$x = \sum_{k=1}^{K_u} \sqrt{\alpha_k P}\, s_k, \tag{8}$$

where, in this system, $K_u = 2$; $\alpha_1 > \alpha_2$ is to guarantee that the user with worse channel conditions receives higher power. Combining IM with NOMA thus allows for simultaneous multiplexing in the power domain and modulation domain, as well as in the STAR-RIS index domain for the first time, thereby significantly boosting both spectral efficiency and reliability.

D. Receiver Design

In particular, at the user side, detection consists of two major steps: IM index detection followed by NOMA symbol detection. Because the activation pattern of STAR-RIS affects the cascaded channel, the index bits must be recovered correctly before symbol demodulation.

1) Index Detection : First, the active subsurface pattern needs to be detected at the STAR-RIS. Maximum likelihood detection is performed by the receiver:

$$\widehat{\Omega} = \arg\min_{\Omega(i)} \left\| y_k - h_{\mathrm{eq},k}(\Omega(i))\, x \right\|^2 . \tag{9}$$

Here, $h_{\mathrm{eq},k}$ depends on the active subsurface $(\Omega(i))$ The ML detector compares every possible subsurface pattern and picks the one which gives the minimum reconstruction error. Correct detection of $\widehat{\Omega}$ ensures correct recovery of index bits b_1.

2) Symbol Detection : Once the active subsurface is determined, the receiver then applies linear detection followed by SIC to recover the modulated bits.

a) MMSE Filtering To suppress multi-user interference, the receiver calculates:

$$\mathbf{W}_{\mathrm{MMSE}} = (\mathbf{H}^H\mathbf{H} + N_0\mathbf{I})^{-1}\mathbf{H}^H. \tag{10}$$

The MMSE detection balances the trade-off between noise enhancement and interference suppression, yielding better performance in noisy conditions. The detected symbols from:

$$\hat{s} = \mathbf{W}_{\mathrm{MMSE}} y. \tag{11}$$

b) Successive Interference Cancellation (SIC) SIC then exploits the power allocation of NOMA by first decoding the signal of the high-power user and subtracting its interference from the received signal:

$$y'_k = y_k - \sum_{\ell=1}^{k-1} \sqrt{\alpha_\ell P}\, h_{\mathrm{eq},k}\, \hat{s}_\ell. \tag{12}$$

After interference subtraction, the remaining lower-power users can be decoded with improved reliability.

Together, IM-based index detection and SIC-based symbol detection enable the receiver to recover both index bits (b_1) and symbol bits (b_2), ensuring full utilization of the additional degrees of freedom introduced by STAR-RIS and IM.

E. Channel Models

In STAR-RIS-assisted systems, where the AP–STAR-RIS and STAR-RIS–user channels make up the end-to-end link, channel modeling is crucial. The cascaded channel depends only on these links because STAR-RIS is passive. It is assumed that every channel experiences i.i.d. Rayleigh fading, which is represented by circularly symmetric complex Gaussian random variables that are common in NLOS settings.

$$\mathbf{G}, \mathbf{h}_1, \mathbf{h}_2 \sim \mathcal{CN}(0, \mathbf{I}_N). \tag{13}$$

This assumption is realistic for dense urban, indoor, and non-line-of-sight (NLOS) propagation scenarios.

1) Effective Cascaded Channels: Since STAR-RIS modifies the incident signal through programmable transmission and reflection coefficients, the effective channels observed by the users are cascaded channels. Specifically,

$$H_1 = \mathbf{h}_1^H\, \mathbf{\Theta}_T \mathbf{G}, \tag{14}$$

$$H_2 = \mathbf{h}_2^H\, \mathbf{\Theta}_R \mathbf{G}, \tag{15}$$

where $\mathbf{\Theta}_T$ and $\mathbf{\Theta}_R$ represent the transmission and reflection matrices.

These cascaded channels exhibit a multiplicative structure. As a result, they show more severe fading fluctuations than direct links, making intelligent optimization of STAR-RIS coefficients essential for achieving high link reliability and improved NOMA performance.

2) Rayleigh Fading PDF: Rayleigh fading characterizes environments with rich scattering and no LOS component. The magnitude of a Rayleigh-distributed channel gain follows:

$$f_{|h|}(x) = \frac{x}{\sigma^2} \exp\left(-\frac{x^2}{2\sigma^2}\right), \quad x \geq 0. \tag{16}$$

This model is suitable for highly obstructed AP-RIS or RIS-user links such as indoor, vehicular, and urban microcell environments.

3) Rician Fading PDF: Rician fading is used when a dominant LOS component exists in addition to scattered components. This is practical when STAR-RIS is placed on building facades or indoors with partial visibility:

$$f_{|h|}(x) = \frac{x}{\sigma^2} \exp\left(-\frac{x^2 + A^2}{2\sigma^2}\right) I_0\left(\frac{Ax}{\sigma^2}\right), \quad x \geq 0, \quad (17)$$

where A denotes the LOS strength and $I_0(\cdot)$ is the modified Bessel function of the first kind. The Rician K-factor ($K = A^2/2\sigma^2$) determines the dominance of LOS.

4) Nakagami-m Fading PDF: Nakagami-m fading is a generalized model capable of capturing both severe and mild fading conditions. It is suitable for outdoor-to-indoor scenarios and clustered scattering environments:

$$f_{|h|}(x) = \frac{2m^m}{\Gamma(m)\Omega^m} x^{2m-1} \exp\left(-\frac{mx^2}{\Omega}\right), \quad (18)$$

where m is the fading severity parameter and Ω is the average fading power.

- $m = 1$ corresponds to Rayleigh fading,
- $m > 1$ represents less severe fading,
- $0.5 < m < 1$ represents highly severe fading.

Thus, Nakagami-m provides a more flexible modeling option for STAR-RIS systems deployed in diverse locations.

F. Imperfect CSI Model

Since RIS configuration, user detection, and IM recovery rely on precise cascaded channel knowledge, channel state information (CSI) has a significant impact on STAR-RIS-assisted NOMA performance. In reality, noise, inadequate training, pilot contamination, and hardware flaws make CSI ineffective.

To model realistic channel estimation, the estimated cascaded channel is expressed as:

$$\widehat{H}_i = \sqrt{1 - \xi^2}\, H_i + \xi E_i, \quad (19)$$

where $E_i \sim \mathcal{CN}(0, \mathbf{I}_N)$ is the estimation error, and $\xi \in [0, 1]$ controls the accuracy level.

- $\xi = 0$: perfect CSI,
- $\xi \to 1$: extremely poor CSI,
- intermediate values: practical moderate CSI accuracy.

Imperfect CSI affects:

- **NOMA power-domain separation** (SIC error propagation increases),
- **STAR-RIS phase optimization** (beamforming mismatch),
- **IM pattern detection** (misclassification of active subsurfaces),
- **Overall BER performance** (error floors under severe CSI uncertainty).

In STAR-RIS-assisted IM-NOMA systems, the degradation due to channel uncertainty is more pronounced because the cascaded channel consists of multiple multiplicative components, making accurate CSI estimation significantly more challenging.

III. Performance Analysis

A. Transmission Rate

In the IM-based STAR-RIS NOMA system, the achievable transmission rate for User i over a Rayleigh fading channel is expressed using the Shannon capacity formula, considering the superposition-coded NOMA signal as

$$R_i(\text{SNR}) = \log_2(1 + \gamma_i) \quad \text{[bps/Hz]},$$

where γ_i is the effective received SNR for User i. For a two-user NOMA system, the received SNRs are given by

$$\gamma_1 = \frac{\alpha_1 |H_1|^2 P_{\text{total}}}{\alpha_2 |H_1|^2 P_{\text{total}} + N_0}, \qquad \gamma_2 = \frac{\alpha_2 |H_2|^2 P_{\text{total}}}{N_0}.$$

Here, $|H_i|^2$ represents the cascaded channel power gain from AP $\to$ STAR-RIS $\to$ User i, P_{total} is the total transmit power, and $\alpha_1 > \alpha_2$ are the NOMA power allocation coefficients.

The total transmission rate for IM-STAR-RIS NOMA, including index modulation bits, becomes

$$R_i^{\text{IM}}(\text{SNR}) = b_1 + b_2 \log_2(1 + \gamma_i),$$

where b_1 denotes the number of index bits for subsurface selection and b_2 represents the number of modulated bits per symbol.

B. BER Analysis

The BER for IM-based STAR-RIS NOMA depends on both index detection errors and modulated symbol detection errors. It can be expressed as

$$P_{e,i}(\text{SNR}) \approx P_{e,i}^{\text{IM}} + P_{e,i}^{\text{mod}}.$$

For Rayleigh fading channels, the symbol error probability based on the effective SNR γ_i can be approximated as

$$P_{e,i}^{\text{mod}}(\text{SNR}) \approx \frac{1}{\pi} \int_0^{\frac{(M-1)\pi}{M}} \prod_{n=1}^{N} \left(1 + \frac{\gamma_i \sin^2\theta}{N_0}\right)^{-1} \mathrm{d}\theta,$$

where M denotes the modulation order. The index detection error $P_{e,i}^{\mathrm{IM}}$ depends on the minimum Euclidean distance between feasible subsurface activation patterns under the effective cascaded channel H_i.

BER with Imperfect CSI: In practical systems, the channel is not perfectly known. The estimated channel is modeled as

$$\hat{H}_i = \sqrt{1-\tau^2}\, H_i + \tau e_i,$$

where τ represents the CSI error factor (0 for perfect CSI), and $e_i \sim \mathcal{CN}(0,1)$ is the estimation error. Due to this imperfect channel estimate, the effective SNR used for detection becomes

$$\hat{\gamma}_i = \frac{\alpha_i |\hat{H}_i|^2 P_{\mathrm{total}}}{\sigma_n^2 + I_i},$$

where I_i represents the interference term (from NOMA or residual SIC).

Since $\hat{H}_i$ deviates from the true channel H_i, the BER under CSI error is obtained by averaging over the distribution of channel estimation error:

$$P_{e,i}^{\mathrm{CSI}}(\mathrm{SNR}) = \mathbb{E}_{\hat{H}_i}\left[P_{e,i}^{\mathrm{mod}}(\hat{\gamma}_i)\right].$$

Thus, the total BER with imperfect CSI is

$$P_{e,i}^{\mathrm{total}}(\mathrm{SNR}) \approx P_{e,i}^{\mathrm{IM}}(\hat{H}_i) + P_{e,i}^{\mathrm{mod}}(\hat{\gamma}_i).$$

A higher CSI error factor τ reduces the accuracy of $\hat{H}_i$, lowers the effective SNR $\hat{\gamma}_i$, and results in a higher BER, especially for User 2 due to residual SIC errors.

IV. Simulation Results

A. Comparison: STAR-RIS NOMA vs IM-STAR-RIS NOMA

The comparison of the BER performance of both STAR-RIS NOMA and IM-STAR-RIS NOMA under Rayleigh Fading conditions presented in Figure 2 indicates that for each user, IM-STAR-RIS's performance is better than that of STAR-RIS NOMA and also evidenced lower BER values when SNR levels were high due to increases in the index domain.

B. BER of IM-STAR-RIS NOMA under Different Modulation Schemes

The BER performance of IM-STAR-RIS NOMA with BPSK, QPSK, and 8-QAM is displayed in Figure 3. The trade-off between modulation order and error performance is highlighted by the fact that User 1 consistently performs better due to favorable power allocation, while BPSK achieves the lowest BER, followed by QPSK and 8-QAM.

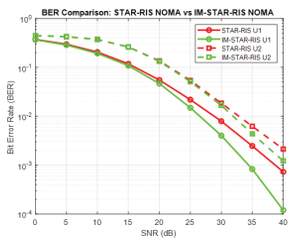


Fig. 2. BER vs SNR comparison between STAR-RIS NOMA and IM-STAR-RIS NOMA.

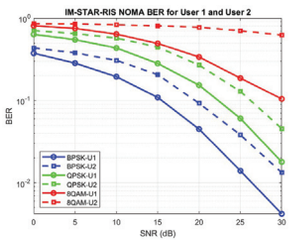


Fig. 3. BER performance of IM-STAR-RIS NOMA under various modulation schemes.

C. IM-STAR-RIS NOMA Performance under Different Fading Channels

BER performance under AWGN, Rayleigh, Rician, and Nakagami-m fading is compared in Figures 4 and 5. Rayleigh does the worst, AWGN does the best, and Nakagami-m does better than Rician. The robustness of IM-STAR-RIS NOMA across fading scenarios is confirmed by User 2, who exhibits higher BER due to lower power allocation.

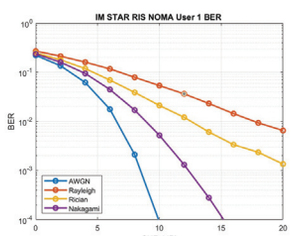


Fig. 4. BER of User 1 under different fading environments.

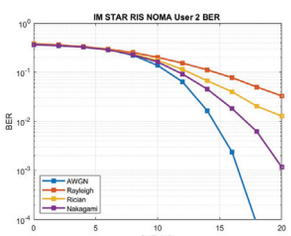


Fig. 5. BER of User 2 under different fading environments.

D. IM-STAR-RIS NOMA under Perfect and Imperfect CSI Conditions

System performance under perfect and imperfect CSI is contrasted in Figures 6 and 7. BER deteriorates as CSI accuracy declines, with User 2 being more impacted by residual SIC interference while User 1 stays largely stable. This emphasizes how crucial precise CSI estimation is to dependable IM-STAR-RIS NOMA operation.

E. Effect of Different No. of STAR-RIS Elements

The BER performance for STAR-RIS with $N = 64$, 128, and 256 elements is displayed in the Figure 8. Due to improved beamforming and array gain, increasing N improves both users' BER; $N = 256$ yields the best results.

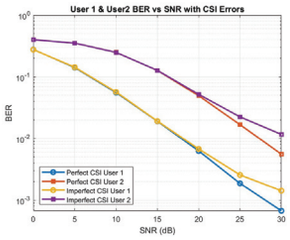


Fig. 6. BER under perfect and imperfect CSI.

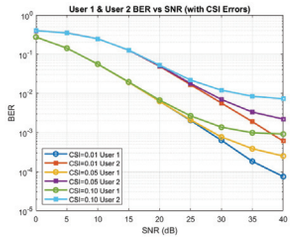


Fig. 7. Impact of CSI errors on IM-STAR-RIS NOMA BER.

F. Transmission Rate Analysis

The achievable rates of IM-STAR-RIS NOMA and conventional NOMA are contrasted in Figure 9. Because of index modulation and improved RIS-assisted channel gains, IM-STAR-RIS achieves higher throughput for both users across all SNRs, with the benefit becoming more noticeable at high SNR.

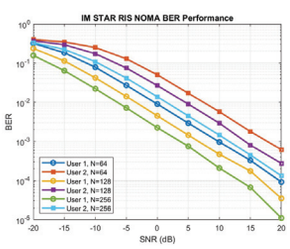


Fig. 8. BER performance for different STAR-RIS elements

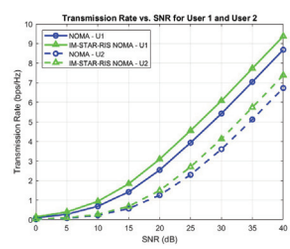


Fig. 9. Transmission rate

V. Conclusion

The combination of STAR-RIS-assisted NOMA and index modulation provides increased spectral efficiency through both improved spectral efficiency and increased reliability and greater coverage capabilities without having to increase bandwidth or power. The ability to transmit and reflect signals at the same time increases user equity, fading resistance, and increased resource use. As such, this technology has great potential for utilization in the next generation of wireless networks.

References

[1] L. Dai, B. Wang, Y. Yuan, S. Han, I. Chih-lin and Z. Wang, "Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends," *IEEE Communications Magazine*, vol. 53, no. 9, pp. 74–81, Sep. 2015, doi: 10.1109/MCOM.2015.7263349.

[2] Y.-T. Lai, Y.-R. Ciou and J.-M. Wu, "Index Modulation Multiple Access," in *Proc. IEEE PIMRC*, Bologna, Italy, 2018, pp. 1–5, doi: 10.1109/PIMRC.2018.8580958.

[3] Y. Liu *et al.*, "Reconfigurable Intelligent Surfaces: Principles and Opportunities," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 3, pp. 1546–1577, 3rd Quarter 2021, doi: 10.1109/COMST.2021.3077737.

[4] A. Almohamad, M. O. Hasna, S. Althunibat and K. Qaraqe, "A Novel Downlink IM-NOMA Scheme," *IEEE Open Journal of the Communications Society*, vol. 2, pp. 235–244, 2021, doi: 10.1109/OJCOMS.2021.3051082.

[5] L. Wei, C. Huang, G. C. Alexandropoulos, C. Yuen, Z. Zhang and M. Debbah, "Channel Estimation for RIS-Empowered Multi-User MISO Wireless Communications," *IEEE Transactions on Communications*, vol. 69, no. 6, pp. 4144–4157, Jun. 2021, doi: 10.1109/TCOMM.2021.3063236.

[6] C. Wu, Y. Liu, X. Mu, X. Gu and O. A. Dobre, "Coverage Characterization of STAR-RIS Networks: NOMA and OMA," *IEEE Communications Letters*, vol. 25, no. 9, pp. 3036–3040, Sep. 2021, doi: 10.1109/LCOMM.2021.3091807.

[7] X. Mu, Y. Liu, L. Guo, J. Lin and R. Schober, "Simultaneously Transmitting and Reflecting (STAR) RIS Aided Wireless Communications," *IEEE Transactions on Wireless Communications*, vol. 21, no. 5, pp. 3083–3098, May 2022, doi: 10.1109/TWC.2021.3118225.

[8] W. Khalid, M. A. Ur Rehman, T. Van Chien and H. Yu, "Simultaneously Transmitting and Reflecting-Reconfigurable Intelligent Surfaces with Hardware Impairment and Phase Error," in *Proc. IEEE ICAIIC*, Bali, Indonesia, 2023, pp. 654–656, doi: 10.1109/ICAIIC57133.2023.10067009.

[9] G. Cao, M. Li, H. Yuan, W. Chen, L. Li and A. Raouf, "Error Performance of RIS-Assisted NOMA Networks with Imperfect Channel State Information," in *Proc. IEEE VTC-Spring*, Florence, Italy, 2023, pp. 1–5, doi: 10.1109/VTC2023-Spring57618.2023.10199367.

[10] S. Kumari, S. Srirangarajan and V. Bhatia, "STAR-RIS Assisted NOMA with Receive Spatial Modulation," in *Proc. IEEE CICT*, Jabalpur, India, 2023, pp. 1–5, doi: 10.1109/CICT59886.2023.10455667.

[11] X. Liu, T. Wei, H. Liang, C. Guo and B. Liao, "Robust Beamforming for RIS-Assisted NOMA Systems With CSI Imperfection and Low-Resolution Phase Shifters," in *Proc. IEEE ICICSP*, Xi'an, China, 2023, pp. 654–658, doi: 10.1109/ICICSP59554.2023.10390794.

[12] K. M. Hamza, S. Basharat, H. Jung, M. Gidlund and S. A. Hassan, "Secrecy Analysis of RIS-Assisted Uplink NOMA Systems Under Nakagami-m Fading," in *Proc. IEEE ICC Workshops*, Denver, CO, USA, 2024, pp. 1511–1516, doi: 10.1109/ICCWorkshops59551.2024.10615519.

[13] J. Zhou, J. Hou, Y. Zhang, X. Hu, L. Yang and S. Liu, "Secure Transmission in Segmented RIS-Assisted NOMA Networks with Imperfect SIC," in *Proc. ISAPE*, Hefei, China, 2024, pp. 1–5, doi: 10.1109/ISAPE62431.2024.10840408.

[14] G. Yonghao, D. Shuping, L. Jun, S. Wenli, H. Jia and H. Yu, "Generalized Quadrature Spatial Modulation for STAR-RIS Aided NOMA Networks," *China Communications*, vol. 22, no. 4, pp. 1–12, Apr. 2025, doi: 10.23919/JCC.fa.2024-0368.202504.

[15] A. Chauhan, S. Ghosh and A. Jaiswal, "RIS Partition-Assisted Non-Orthogonal Multiple Access (NOMA) and Quadrature-NOMA With Imperfect SIC," *IEEE Transactions on Wireless Communications*, vol. 22, no. 7, pp. 4371–4386, Jul. 2023 , doi: 10.1109/TWC.2022.3224645.

A Prototype of Non-invasive Blood Sugar Measurement Device using MAX30102 and IoT Connectivity

Evan Fajar Firmansyah[§], Dodi Zulherman[§‡ *], Slamet Indriyanto[†‡]
[§]*Departement of Biomedical Engineering*, *Telkom University*, Purwokerto, Indonesia
[†]*Department of Electrical Engineering*, *Telkom University*, Purwokerto, Indonesia
[‡] *Center of Excellence for Healthcare and Well-Being Technologies*, *Telkom University*, Purwokerto, Indonesia
[*]Corresponding author: zulherman@telkomuniversity.ac.id

***Abstract*—Non-invasive blood glucose monitoring is increasingly being developed to reduce discomfort associated with invasive methods, such as finger pricking. This study proposes a system that uses the MAX30102 optical sensor based on Near-Infrared Spectroscopy (NIRS) to estimate blood sugar levels using RED and Infrared signals. Two machine learning models were compared: Linear Regression and Random Forest. The Linear Regression model showed low accuracy (R^2 = 0.0233), while the Random Forest model achieved a much higher R^2 (0.8453) with lower error. The main contribution of this study is the development of a non-invasive glucose-monitoring approach that integrates optical sensing with an ensemble machine learning model, providing a more comfortable alternative to invasive methods. Furthermore, the proposed model can be deployed on ESP32 using lightweight frameworks such as TensorFlow Lite, Edge Impulse, or TinyML, enabling the future development of wearable medical devices within the IoMT paradigm.**

***Index Terms*—Non-invasive, MAX30102, NIRS, IoMT, Linear Regression, Random Forest.**

I. Introduction

Diabetes mellitus is a chronic disease that has become a global health problem with a far-reaching impact on the quality of life of those affected. According to the International Diabetes Federation (IDF), in 2021, approximately 537 million adults worldwide were estimated to be living with diabetes, and this number continues to increase annually [1]. This is consistent with the findings of the 2018 Basic Health Research (Riskesdas), which reported that the national prevalence of diabetes mellitus, based on doctor diagnosis, was 2.0%, while blood tests reached 10.9% in the population aged 15 years and over [2]. Hirsch et al. (2024) also confirmed this increasing prevalence, stating that the number of diabetes cases has increased significantly in many countries. This raises the need for glucose monitoring methods that are more practical, accessible, and can be performed routinely by patients [3].

Commonly used methods for monitoring glucose levels are still dominated by invasive procedures, such as finger or arm pricks with a needle. This method, while accurate, often causes discomfort for patients, especially those who perform repeated measurements throughout the day [4]. This discomfort has prompted various studies to develop non-invasive blood glucose monitoring methods using light-based sensor technology or other approaches that do not require direct blood sampling. As an alternative to invasive methods, recent studies have begun developing non-invasive approaches to measuring blood glucose levels, one of which uses optical technology based on Near-Infrared Spectroscopy (NIRS). This technology works by detecting changes in glucose concentration through the absorption patterns of infrared light as it penetrates body tissue [5]. Various studies have shown that the use of specific wavelengths in the infrared spectrum has great potential for non-invasive glucose monitoring, enabling rapid, practical detection without blood sampling [6].

In this context, optical sensors such as the MAX30102 are widely used due to their ability to detect variations in light absorbance in human tissue [7]. This sensor operates using two primary wavelengths: red light at 660 nm and infrared light at 880 nm, both of which are effective for detecting changes in tissue characteristics related to blood glucose concentration. The combination of these two wavelengths plays a crucial role in improving the accuracy of non-invasive blood glucose estimation. However, the development of non-invasive methods is not without challenges, particularly regarding measurement accuracy. Various physiological factors,

such as skin surface temperature, tissue thickness, and variations in local blood circulation, can affect the stability and accuracy of the data obtained [8]. This suggests the need for advanced strategies to strengthen the validity and improve the correlation between the optical signals recorded by the sensor and actual blood glucose levels. In addition to physiological constraints, hardware limitations are also significant. One example is the variety of infrared wavelengths, which are generally available commercially in limited quantities. Several recent studies indicate that using a combination of several wavelengths simultaneously can help improve the accuracy of blood glucose level prediction [9]. Therefore, selecting the right sensor type and developing more adaptive optical technology are crucial to supporting the success of non-invasive methods.

Based on this description, this research focuses on designing a non-invasive blood sugar measurement system using the MAX30102 sensor. This research not only emphasizes the use of the sensor as a measuring tool but also evaluates the extent to which variations in the wavelength of light used can affect the accuracy of glucose level estimation. It is hoped that this research will encourage the development of health technologies that are more comfortable to use, more accurate, and affordable for the wider community [10]. However, several previous studies still have limitations. Most research relies on simulated datasets or focuses solely on linear models, which are insufficient for complex physiological signals [11]. Additionally, deploying machine learning models on embedded systems such as the ESP32 is still rare, even though real-time monitoring is essential for medical applications. Therefore, this study addresses these gaps by using real experimental data and comparing Linear Regression with Random Forest to find a more accurate non-invasive prediction model. This approach also opens the opportunity to be implemented in wearable IoMT medical devices using lightweight deployment frameworks such as TensorFlow Lite, Edge Impulse, or TinyML in future research.

II. METHOD

A. System Overview

This research used an experimental method by designing a tool to read optical signals from the MAX30102 sensor. This sensor uses the Near-Infrared Spectroscopy (NIRS) principle, emitting red and infrared light onto the skin and receiving the reflected signals. This reflected signal, called a PPG (Photoplethysmography) signal, is used to determine whether it correlates with a person's blood sugar levels. The ESP32 microcontroller was used because it supports I2C communication, has Wi-Fi, and is suitable for developing systems towards the Internet of Medical Things (IoMT) [12] [13].

B. Research Workflow

The research workflow is illustrated in Fig. 1. In the first stage, the RED and Infrared signals from the MAX30102 sensor are acquired to verify signal stability prior to data recording. Once stable readings are obtained, the data are transmitted to a computer via serial communication and stored as a dataset.

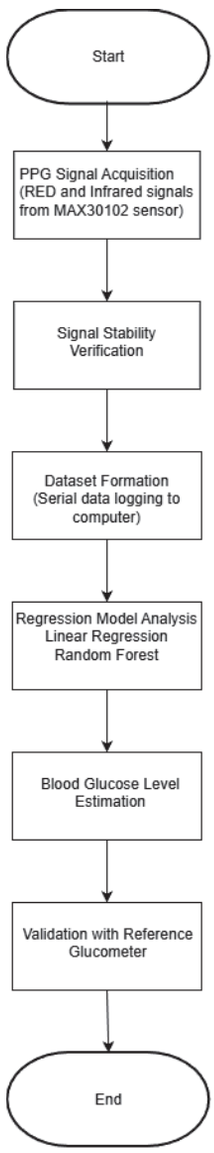


Fig. 1: Research Methodology Workflow

Unlike previous studies that apply extensive signal pre-processing techniques, this research directly analyzes raw PPG signals to determine whether they can be

used for preliminary estimation of blood glucose levels. The recorded data were processed using two machine learning models: Linear Regression and Random Forest. Linear Regression was used to examine potential linear relationships between the optical signals and blood glucose levels and to generate an interpretable mathematical model [11]. If the data did not exhibit linear behavior, Random Forest was utilized as it is more flexible and suitable for complex physiological data patterns [15].

C. Data Acquisition

Data collection was conducted on multiple subjects. Each subject produced hundreds of RED and Infrared samples, paired directly with blood glucose values measured using a glucometer as reference ground truth. These raw data points were then used for model training and evaluation. The use of real experimental data distinguishes this work from many previous studies that relied on simulated or secondary datasets.

D. Machine Learning Analysis

The collected data were used to train two regression models. The first model was Linear Regression, applied to determine whether a linear relationship exists between optical signals and blood glucose levels. This model was also chosen because it can generate an interpretable mathematical formula that may be suitable for implementation on a microcontroller in future developments [15]. However, if the signals did not follow a linear pattern, an alternative method was required. Therefore, Random Forest was selected as a second model because it is an ensemble-based method that can better adapt to nonlinear and complex physiological patterns [11]. The performance of both models was evaluated using three commonly used metrics in biomedical regression: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the coefficient of determination (R^2). These metrics were used to assess how well each model estimated blood glucose levels from optical signals recorded by the MAX30102 sensor.

E. Hardware Implementation

After the model analysis, the hardware prototype was assembled on a breadboard for easy testing and modification. The MAX30102 sensor was connected to the ESP32 microcontroller via I2C on the SDA and SCL pins. The sensor was powered with 3.3V according to its datasheet specifications, and a 16×2 LCD was also connected via I2C to minimize digital pin usage. The complete data-collection circuit is shown in Fig. 2.

The initial circuit of this non-invasive blood sugar measurement system uses a MAX30102 sensor connected to an ESP32 microcontroller. The ESP32 reads optical signals from the sensor and displays them directly on a 16x2 LCD screen. The MAX30102 sensor and LCD are connected to the ESP32 via I2C on the SDA and SCL pins. The sensor receives a 3.3-volt electrical voltage according to its operating requirements. The system uses a breadboard for easy testing and repair before being designed into a permanent circuit.

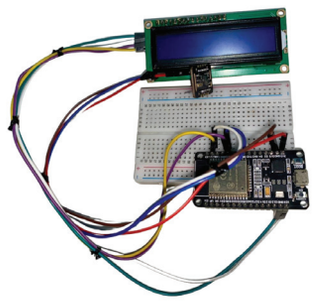


Fig. 2: Hardware implementation of the proposed non-invasive blood glucose measurement system using ESP32 and MAX30102 sensor

III. Result

The initial system testing was conducted to ensure that the MAX30102 sensor could acquire stable RED and Infrared optical signals. The results showed that the sensor successfully detected intensity changes when the fingertip was placed consistently on the sensor surface. During signal acquisition, it was found that signal stability was significantly affected by finger pressure, ambient light, and movement. Therefore, the preprocessing stage was crucial for improving signal quality and enhancing model performance.

Once the PPG signal stability was achieved, data collection began and was matched with invasive glucometer readings as reference values. The experiment was conducted on several subjects with a total of 900 data samples, which were used to build and evaluate two regression models. A sample of the collected dataset is shown in Table I.

From the initial 500 data samples, the Linear Regression model only achieved an R^2 of 0.019. After increasing the dataset to 900 samples, the accuracy slightly improved to $R^2 = 0.0233$ using Linear Regression. However, this value remains far from clinically acceptable. The performance of both models is compared in Table II, which shows that Linear Regression is inadequate for

TABLE I: Sampel of Collected Datasheet

Name	Age (years)	Gender	red	infrared	glucometer (mg/dL)
S1	22	man	23538	31784	123
S2	16	woman	29670	33332	110
S3	28	man	22184	27576	175
S4	29	woman	26552	31308	213
S5	16	man	20887	26092	126
S6	20	woman	25685	32478	134
S7	24	woman	26461	31586	110
S8	18	man	26716	33640	120
S9	30	woman	21216	29059	108
S10	20	man	24624	31834	154

complex physiological patterns, while Random Forest produces significantly better results.

TABLE II: Perbandingan Model

Model	$R\sqrt{2}Score$	Information
Regresi Linear	0.019	500 data
Regresi Linear	0.023	900 data
Random Forest Regressor	0.845	900 data

To visualize the prediction performance, Figure 3 presents a scatter plot comparing actual glucose values with model predictions. The Linear Regression model shows a concentration of values within a narrow range, indicating poor responsiveness to signal variation. In contrast, Random Forest predictions more closely follow the actual values.

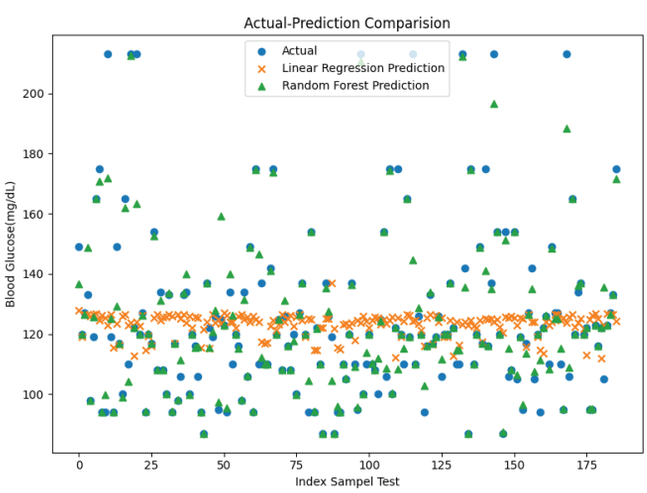


Fig. 3: Comparison of actual blood glucose values with predictions from Linear Regression and Random Forest models.

Figure 3 compares actual blood glucose values with those predicted by Linear Regression and Random Forest models. The Linear Regression model shows a narrow prediction range with limited variation, indicating its inability to capture physiological signal dynamics. In contrast, the Random Forest model shows a more responsive pattern that closely tracks the actual glucose values, indicating better adaptability to the nonlinear relationships in the optical signals.

These results confirm that the relationship between the PPG signals and glucose levels is nonlinear, which explains why Random Forest significantly outperforms Linear Regression. Therefore, ensemble-based regression models are more suitable for estimating non-invasive blood glucose levels using optical sensors.

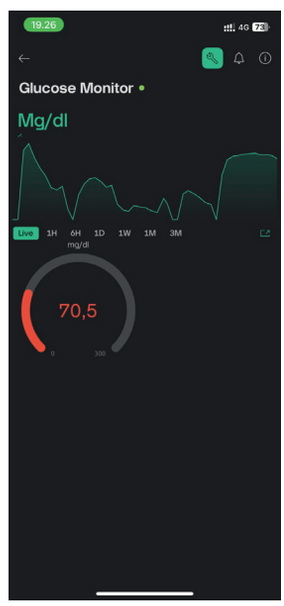


Fig. 4: Real-time Glucose Monitoring Display using Blynk Platform

After the Random Forest model was selected as the best, an initial test was conducted to display the prediction results on the Blynk mobile platform. The glucose value was sent from the microcontroller to the Blynk server, allowing real-time visualization on a smartphone. This test serves as a preliminary IoMT prototype and shows the potential of the system to be developed into a wearable monitoring device in future research. The display from the Blynk application is shown in Fig. 4.

IV. Discussions

The experimental results demonstrate that the MAX30102 sensor can be used as an optical signal

acquisition device for non-invasive blood glucose estimation. However, the quality of the collected signal is highly influenced by several physiological and environmental factors, including finger pressure, ambient lighting, and motion artifacts. These sources of noise often lead to unstable PPG signals, making the preprocessing stage an essential part of this study to ensure that the extracted features are clean and applicable for further machine learning analysis. The implementation of a Moving Average Filter successfully reduced high-frequency noise and improved the signal stability, making the data more suitable for regression model development.

Linear Regression was initially applied as a baseline model to examine whether the relationship between optical signals (RED and Infrared) and blood glucose levels is linear. However, the obtained R^2 value of only 0.0233 indicates that the model was unable to capture the underlying physiological patterns effectively. This finding confirms that simple linear statistical methods are insufficient for estimating glucose levels solely from optical data. As a result, a more advanced machine learning method was required.

The Random Forest model was then used to handle the nonlinear characteristics of the input signals. The model produced a notable increase in prediction accuracy, achieving an R^2 value of 0.8453 and an MAE of approximately 4 mg/dL. These results indicate that ensemble learning strategies—by combining multiple decision trees—are more adaptive to physiological variations among subjects and can better generalize across different signal patterns. This also aligns with previous research suggesting that ensemble-based methods outperform linear models for complex biomedical data.

Overall, this study confirms that integrating optical sensors with machine learning models is a promising approach for non-invasive blood glucose monitoring. Although direct implementation on the ESP32 microcontroller has not yet been carried out in this research, the model's performance has shown sufficient potential to serve as a foundation for future embedded system development. Deployment on lightweight platforms such as TensorFlow Lite, Edge Impulse, or TinyML remains a feasible direction, although further model optimization and quantization are required to ensure compatibility with limited hardware resources.

Future work should focus on expanding the number of test subjects, enhancing robustness to motion artifacts, and validating the system through clinical trials to ensure medical reliability. By achieving these goals, the proposed approach could be adapted into an IoMT-based wearable biomedical device capable of real-time blood glucose monitoring without the need for invasive finger-prick methods. This represents a substantial step toward more comfortable, efficient, and cost-effective healthcare solutions for individuals with diabetes who require continuous monitoring.

V. Conclusion

This study successfully designed and tested a non-invasive blood glucose measurement system using the MAX30102 sensor based on the Near-Infrared Spectroscopy (NIRS) principle. The MAX30102 captured RED and Infrared signals effectively, though preprocessing was required to reduce noise from finger pressure, ambient light, and motion artifacts. Signal stabilization was found to be essential for reliable machine learning analysis.

Linear Regression was used as the baseline model; however, it yielded an R^2 of 0.0233, indicating that the relationship between optical signals and blood glucose levels is not linear. The Random Forest model provided significantly better performance, with an MAE of 4.0516 and an R^2 of 0.8453, demonstrating its suitability for non-invasive estimation and its closer alignment with clinical standards.

These findings indicate that integrating optical sensors with machine learning methods has strong potential to develop a more comfortable and practical non-invasive glucose monitoring system. Although an ESP32 implementation was not conducted in this study, the model's performance is promising and may serve as a foundation for future embedded IoMT systems. Further work is required to optimize the model, increase the number of test subjects, and conduct clinical validation to ensure applicability in real-world medical settings.

Acknowledgment

The authors would like to express their sincere gratitude to the Research and Community Service Institute (LPPM) of Telkom University Purwokerto and the Biomedical Engineering Study Program for their guidance, facilitation, and institutional support throughout the implementation of this research.

References

[1] International Diabetes Federation, "IDF Diabetes Atlas 10th Edition," Brussels, 2021.

[2] Kementerian Kesehatan Republik Indonesia, "Laporan Nasional Riset Kesehatan Dasar (Riskesdas) 2018," Jakarta, 2018.

[3] “I. B. Hirsch, A. Tirosh, and A. Navon, “Noninvasive Real-Time Glucose Monitoring Is in the Near Future,” Diabetes Technol Ther, vol. 26, no. 9, pp. 661–666, Sep. 2024, doi: 10.1089/dia.2024.0009

[4] J. Wu, Y. Liu, H. Yin, and M. Guo, “Review Article A new generation of sensors for noninvasive blood glucose monitoring,” 2023. [Online]. Available: www.ajtr.org

[5] J. L. C. Perez et al., “Fiber optic sensors: A review for glucose measurement,” Mar. 01, 2021, MDPI. doi: 10.3390/bios11030061.

[6] S. Ali et al., “Development and Evaluation of a Sensor-Based Non-Invasive Blood Glucose Monitoring System Using Near-Infrared Spectroscopy †,” Engineering Proceedings, vol. 82, no. 1, 2024, doi: 10.3390/ecsa-11-20395.

[7] A. Hina and W. Saadeh, “Noninvasive Blood Glucose Monitoring Systems Using Near-Infrared Technology—A Review,” Jul. 01, 2022, MDPI. doi: 10.3390/s22134855.

[8] A. Kandwal et al., “A comprehensive review on electromagnetic wave based non-invasive glucose monitoring in microwave frequencies,” Sep. 30, 2024, Elsevier Ltd. doi: 10.1016/j.heliyon.2024.e37825.

[9] M. Shokrekhodaei, D. P. Cistola, R. C. Roberts, and S. Quinones, “Non-Invasive Glucose Monitoring Using Optical Sensor and Machine Learning Techniques for Diabetes Applications,” IEEE Access, vol. 9, pp. 73029–73045, 2021, doi: 10.1109/ACCESS.2021.3079182.

[10] Mr. K. G. Mr. C. S. Dr. R. Athilingam, “A Review on Non-Invasive Technology for Monitoring Blood Glucose in the Management of Diabetes,” International Research Journal of Modernization in Engineering Technology and Science, 2024, doi: https://www.doi.org/10.56726/IRJMETS60492.

[11] J. Allen, “Photoplethysmography and its application in clinical physiological measurement,” Physiological Measurement, vol. 28, no. 3, pp. R1–R39, 2007.

[12] Man B. Noninvasive Spectroscopic Detection of Blood Glucose and Analysis of Clinical Research Status. J Healthc Eng. 2022 Feb 7;2022:8325451. doi: 10.1155/2022/8325451. Retraction in: J Healthc Eng. 2023 Jul 12;2023:9856086. doi: 10.1155/2023/9856086. PMID: 35178236; PMCID: PMC8844100.

[13] Y. Zhang et al., “Two-Stage Motion Artifact Reduction Algorithm for rPPG Signals from MAX30102 Sensor Using Moving Average and Normalization,” Sensors (Basel), vol. 23, no. 8, p. 3874, Apr. 2023.

[14] A. Kermani and H. Esmaeili, ”Photoplethysmography based non-invasive blood glucose estimation using systolic-diastolic framing MFCC features and machine learning regression,” BioImpacts, vol. 15, p. 30589, 2025. doi: 10.34172/bi.2025.30589.

[15] G. Zhang et al., ”A Noninvasive Blood Glucose Monitoring System Based on Smartphone PPG Signal Processing and Machine Learning,” IEEE Trans. Ind. Informat., vol. 16, no. 11, pp. 7209-7218, Nov. 2020.

Design and Development of an IoT-Based Blood Oxygen (SpO_2) Monitoring Device for Asthma Patients

Irvine Kornelius , Sevia Indah Purnama †, Slamet Indriyanto ‡, Alvera Carolina §
* *Departement of Biomedical Engineering, Telkom University, Purwokerto, Indonesia*
† *Departement of Biomedical EngineeringTelkom University, Purwokerto, Indonesia*
‡ *Departement of Biomedical Engineering, Telkom University, Purwokerto, Indonesia*
§ *Department of International Business, Jenderal Soedirman University, Purwokerto, Indonesia*
Emails: *irvinekornelius@student.telkomuniversity.ac.id , seviaindah@telkomuniversity.ac.id , slamet@telkomuniversity.ac.id , alvera.carolina@mhs.unsoed.ac.id*

***Abstract*— Health complications related to impaired airway function often arise from persistent irritation within the respiratory system, which gradually reduces airflow capacity and can escalate into critical conditions when left unmanaged. In many regions, the measurement of physiological parameters associated with breathing—particularly the level of oxygen circulating in peripheral blood (SpO_2)—remains dependent on instruments available only in clinical settings. The limited accessibility of personal monitoring tools creates a need for a device capable of delivering continuous readings without requiring hospital visits. To address this gap, the present work introduces a portable SpO_2 observation system built around the ESP32 microcontroller. Data acquisition is performed using a MAX30102 photoplethysmography module, and the processed information is visualized locally through an OLED interface while simultaneously being forwarded to an Internet-connected platform utilizing an IoT architecture. Remote users can view their physiological status in real time through the online dashboard, enabling routine self-evaluation of respiratory performance. Such a system offers a practical and economical alternative to conventional monitoring methods and has the potential to support earlier identification of declining oxygenation levels, contributing to better management of chronic airway conditions and promoting increased public attention to respiratory well-being.**

***Index Terms*— Air Disorder, Asthma, ESP32, IoT, Health Monitoring**

I. INTRODUCTION

This Asthma is a disorder of the respiratory system characterized by intermittent airway obstruction, inflammation of the respiratory tract, and bronchial hypersensitivity. The number of asthma cases has increased rapidly during the 20th century, particularly since the 1960s [1]. Pulse rate is a wave felt in the arteries caused by the pumping of blood from the heart into the blood vessels. Several factors can cause changes in blood vessels, including shock, changes in oxygen saturation, physical activity, changes in body temperature, and intracranial pressure, which is the pressure inside the cranial cavity that contains the brain, cerebrospinal fluid, and blood vessels [2]. Currently, blood oxygen level monitoring is generally performed in healthcare facilities using conventional pulse oximeters. However, the availability of these devices is still limited for independent home monitoring. In addition, many existing devices operate standalone without remote monitoring features or long-term data logging, making them less effective for patients who require continuous observation. These limitations become an obstacle for patients at risk of hypoxemia who need more intensive supervision. Several previous studies have developed Internet of Things (IoT)-based pulse oximetry devices using various approaches and technologies. For example, one study designed a device using the MAX30100 sensor integrated with a WeMos D1 Mini to measure blood oxygen levels (SpO_2) and heart rate in real time, with the measurement results displayed through a web server to facilitate remote monitoring [3]. In addition, another study developed a similar device capable of displaying SpO_2 and heart rate readings via an internet connection to a server, achieving a relatively low error rate—0.36% for SpO_2 and 1.12% for heart rate—indicating high measurement accuracy. The use of sensors such as the MAX30100, WeMos D1 Mini or ESP8266 modules, as well as cloud-based and mobile applications, demonstrates a strong trend toward continuous and contactless patient health monitoring [4]. The solution proposed in this study is the design and development of a blood oxygen monitoring device based on the ESP32 for asthma patients. This system uses the MAX30102 sensor connected to the ESP32 microcontroller to perform real-time SpO2 measurements. The acquired data will be displayed on an OLED screen and can be transmitted to an IoT-based application, enabling remote monitoring via the internet. The MAX30102 sensor is a pulse sensor that functions

to detect heart rate and blood oxygen levels. It operates with a 3.3-volt power supply and consists of four pins: VCC, GND, SDA, and SCL. The sensor works by measuring changes in blood volume within the blood vessels beneath the skin using IR light, which is then reflected back to a photodiode. Blood volume changes with each heartbeat, causing variations in the amount of reflected light. These variations are processed to produce IR values, which are then further analyzed by the microcontroller [5].

A. Research by Kemalasari and Rochmad (2022)

This study discusses the design of an oxygen saturation and heart rate measuring device using the MAX30100 sensor, which operates based on the photoplethysmography method. The research emphasizes that measurements can be performed non-invasively without requiring additional medical procedures, making the device practical to use. However, the MAX30100 sensor used has relatively low sensitivity, especially when exposed to ambient light or user finger movement, resulting in less stable measurement accuracy[6].

B. Research by Onubogu and Ayuk (2022)

This study is a clinical investigation that examines the relationship between SpO_2 values and Peak Expiratory Flow Rate in children experiencing asthma attacks. The results show that a decrease in SpO_2 levels can reflect declining respiratory function in asthma patients, making this parameter useful as an indicator of severity. Although the study provides a strong clinical basis for the relevance of SpO_2 in asthma monitoring, it does not design a monitoring device nor provide a system that can be directly used by patients at home. Therefore, the study does not yet address the need for real-time, independent monitoring for asthma patients [7].

C. Research by Baiq Nurul Laili and her colleagues (2020)

They developed an SpO2 monitoring system based on the WeMos D1 Mini that can be accessed through a web server. This system demonstrates that remote monitoring can be performed at a low cost. However, the system does not provide a direct display on the device, requiring users to open a browser to view the measurement results. In addition, the stability of the internet connection greatly affects the continuity of data monitoring, which can become an obstacle in situations that require a fast response [3].

D. Research by Refni and her colleagues (2022)

developed an Internet of Things–based SpO_2 monitoring device aimed at early detection of Covid-19 symptoms. The system is capable of transmitting data to a cloud platform, allowing remote monitoring. However, the system has limitations in accuracy because the sensor used is prone to noise, and the device design is not yet portable, making it less comfortable for long-term use. This research also does not include heart rate monitoring features, and its focus is not directed toward chronic respiratory conditions such as asthma [4].

E. Research by Muthmainnah and her colleagues (2022)

used the MAX30102 sensor integrated with the ESP8266 and the Blynk application to display real-time heart rate data. The use of the MAX30102 sensor provides improved sensitivity compared to previous versions, and IoT integration enables remote monitoring. However, this study focuses solely on heart rate without considering SpO_2 values, even though SpO_2 is an important indicator for asthma patients. Therefore, this research cannot yet be used for monitoring respiratory conditions [8].

F. Research by Sehrash and her colleagues (2023)

developed an ESP32-based health monitoring system capable of recording several physiological parameters such as heart rate, body temperature, and the user's location. The system is open-source and can be further developed by other researchers. However, the study does not emphasize the accuracy of oxygen saturation measurements and has not been applied in the context of diseases that require strict monitoring of SpO_2 fluctuations, such as asthma [9].

G. Research by Naval Indra Waskita and his colleagues (2024)

utilized the ESP32 and the MAX30100 sensor with the addition of a Chebyshev II filter to improve signal quality, resulting in more stable SpO_2 readings. Nevertheless, the MAX30100 sensor used still has limitations in sensitivity when there are changes in light intensity or hand movement. In addition, this research focuses more on improving algorithm accuracy rather than implementing the device for direct patient monitoring, particularly for individuals with asthma [10].

H. Asthma

Asthma is a respiratory disease influenced by various factors. To better understand this condition, the following explanation highlights several major causes that contribute to the increasing number of asthma cases. Asthma is a chronic lung disease characterized by airway inflammation, variable airflow obstruction, and bronchial hyperresponsiveness to various stimuli.

This condition leads to symptoms such as shortness of breath, coughing, and wheezing, which may vary in frequency and severity. Pathophysiologically, asthma involves an excessive T helper 2 (Th2) immune response mediated by cytokines such as interleukin-4 (IL-4), IL-5, and IL-13. This response causes eosinophil infiltration, allergen-specific immunoglobulin E (IgE) production, and airway remodeling, including smooth muscle hypertrophy and increased mucus production [11].

Chronic inflammation of the airways can lead to narrowing of the bronchial lumen and increased airflow resistance, which contributes to the clinical symptoms of asthma. Factors such as allergens, viral infections, air pollution, and stress can trigger asthma exacerbations by worsening airway inflammation and obstruction. Monitoring blood oxygen saturation (SpO2) is an essential aspect of asthma management, especially during acute exacerbations. A decrease in SpO_2 may indicate hypoxemia, which requires immediate medical intervention. Pulse oximetry is a commonly used non-invasive method for monitoring SpO2 and has been proven useful in assessing the severity of asthma attacks [12].

I. Blood Oxygen Saturation (SpO2)

Blood oxygen saturation (SpO2) is an important indicator that shows the percentage of hemoglobin bound to oxygen in arterial blood. Normal SpO2 values range from 95% to 100%. A decrease in this value may indicate hypoxemia, a condition in which arterial blood lacks sufficient oxygen, potentially leading to impaired function of vital organs. In asthma patients — especially during acute attacks — airway narrowing can disrupt gas exchange in the lungs, resulting in a drop in SpO2 levels. Therefore, regular SpO2 monitoring is essential for early detection of hypoxemia and prevention of further complications [13].

Non-invasive SpO_2 measurement is generally performed using the photoplethysmography (PPG) method. This method utilizes an optical sensor consisting of a light source (red and infrared LEDs) and a photodetector to detect changes in blood volume in the capillaries. Oxygenated and deoxygenated hemoglobin have different light absorption characteristics, so the absorbance ratio at these two wavelengths can be used to calculate the SpO2 value [14].

J. Internet of Things (IoT)

The Internet of Things (IoT) is a technological concept that enables various physical devices, such as sensors and actuators, to connect and communicate through the internet. The main goal of IoT is to continuously extend the benefits of internet connectivity, allowing real-time data collection, exchange, and analysis to support better decision-making. In the healthcare sector, IoT plays a crucial role in developing real-time patient monitoring systems. By integrating wearable devices and medical sensors, physiological data such as heart rate, body temperature, and blood oxygen levels can be collected and transmitted to cloud platforms for analysis. This enables early detection of critical conditions, remote monitoring by medical professionals, and improved access to healthcare services, especially for hard-to-reach populations [15].

K. ESP32 Microcontroller

The ESP32 is a System on Chip (SoC) microcontroller developed by Espressif Systems. This microcontroller is designed for Internet of Things (IoT) applications, featuring wireless connectivity and high processing capability. The ESP32 integrates a dual-core 32-bit Xtensa LX6 processor operating at frequencies up to 240 MHz, and is equipped with 520 KB of SRAM and up to 4 MB of flash memory [16].

L. MAX30102 Sensor

The MAX30102 sensor is an optical module used for non-invasive measurement of blood oxygen saturation (SpO_2) and heart rate. This sensor integrates two LEDs (red and infrared), a photodetector, optics, and signal processing elements into a single chip. The MAX30102 operates based on the photoplethysmography (PPG) principle, where the signal is generated from variations in light absorption by capillary blood beneath the skin. The sensor works by emitting red light (660 nm) and infrared light (880–940 nm) into the skin tissue. Oxygenated hemoglobin (HbO_2) and deoxygenated hemoglobin (Hb) have different absorption levels for these wavelengths. Light that is not absorbed is reflected back and detected by the photodetector. This data is then processed into information about blood oxygen levels and heart rate [17].

II. METHODS

A. System Overview

The proposed system is an Internet of Things (IoT)-based blood oxygen saturation (SpO_2) monitoring device designed to support real-time respiratory assessment for individuals with asthma. The system integrates a MAX30102 optical sensor to measure SpO_2 and heart rate using the photoplethysmography method by emitting red and infrared light and detecting variations in blood volume through reflected signals. These raw signals are processed by an ESP32 microcontroller, which performs signal conditioning, computes physiological parameters, and manages system control while utilizing its built-in Wi-Fi module to transmit data to a cloud-based IoT platform. The

processed results are displayed locally on a 0.96-inch OLED screen, allowing users to obtain immediate readings without internet access, while the same data is simultaneously sent to the Blynk IoT platform for remote monitoring and real-time visualization via a mobile application. This integrated approach provides an efficient, portable, and low-cost solution that enables continuous observation of respiratory conditions and supports early detection of decreased oxygen saturation levels, which is essential for asthma management and preventive healthcare.

B. Research Workflow

The flowchart illustrates the operational process of the early asthma detection system, which utilizes sensors and an ESP32 module. The process begins with the hardware setup, including the oxygen saturation (SpO2) sensor and the ESP32 microcontroller module. Once the system is initialized, the sensor starts collecting SpO2 data from the user's body. The acquired data is then transmitted to the ESP32 for analysis.

The flowchart illustrates the operational process of the early asthma detection system, which utilizes sensors and an ESP32 module. The process begins with the hardware setup, including the oxygen saturation (SpO2) sensor and the ESP32 microcontroller module. Once the system is initialized, the sensor begins collecting SpO2 data from the user's body, and the acquired readings are transmitted to the ESP32 for analysis. The next stage is to evaluate whether the SpO2 value exceeds 95%. If the SpO2 level is above 95%, the system concludes that the user is in a normal condition and displays a "Normal condition" notification. Conversely, if the SpO2 value is below or equal to 95%, the system indicates that the user is at risk of asthma and displays a warning notification stating "Alert! Indicated asthma risk." After the notification is delivered, the process ends. This system functions as an efficient and automated health-monitoring tool capable of quickly detecting potential asthma symptoms.

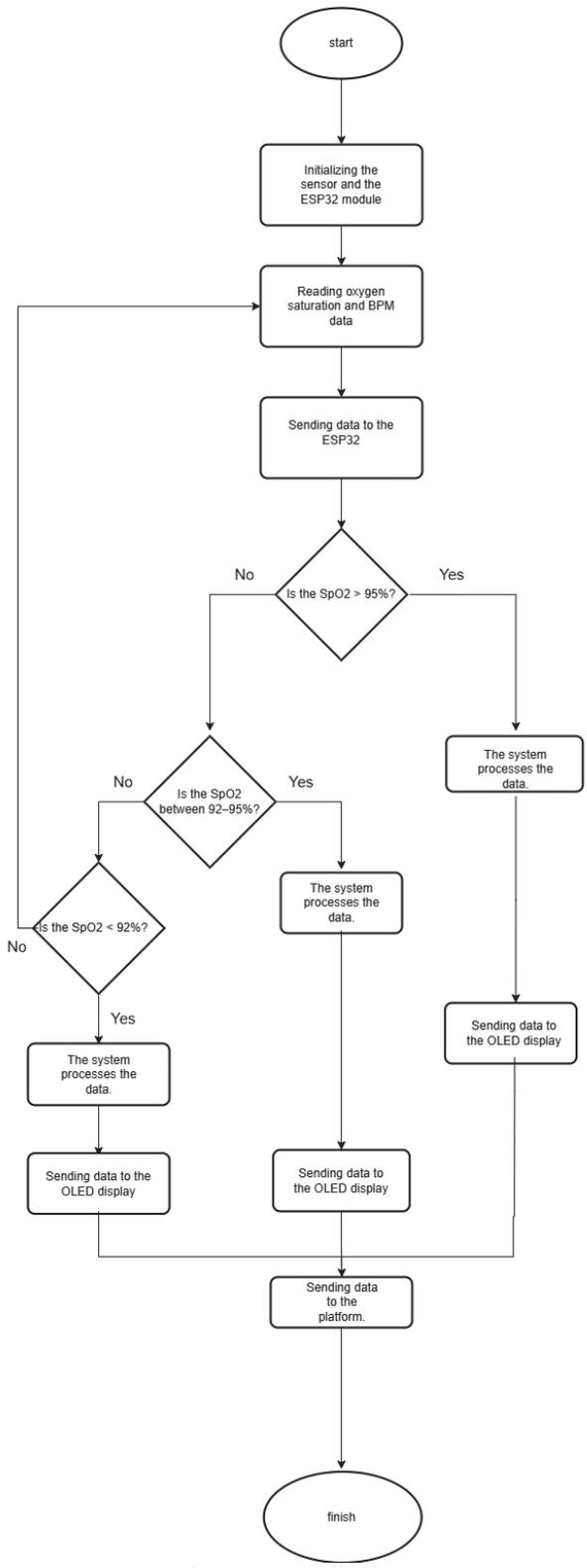


Fig. 1 : Research Flow

C. Data Acquisition

The data acquisition process in this study was conducted using a single subject under two different physiological conditions: resting and post-activity. For each condition, ten measurement trials were performed using both the developed SpO_2 monitoring device and a

commercial pulse oximeter as the reference tool. During each trial, the MAX30102 sensor captured photoplethysmography signals by emitting red and infrared light onto the subject's fingertip and detecting variations in the reflected intensity caused by blood volume changes. These raw optical signals were transferred to the ESP32 microcontroller via the I2C interface for initial processing, including noise reduction and digital conversion into SpO_2 and heart rate values. Each processed measurement was timestamped to ensure consistent comparison across trials and conditions. The calculated values were displayed locally through the OLED screen and simultaneously transmitted via Wi-Fi to the Blynk application, enabling real-time mobile monitoring in addition to local visualization. Reference readings from the commercial pulse oximeter were recorded for every trial to evaluate accuracy and validate system performance. This acquisition workflow ensured controlled, repeatable, and reliable data collection for assessing the effectiveness of the proposed monitoring device under both resting and post-activity states.

D. Signal Processing Analysis

The collected data were used to analyze the performance of the proposed SpO2 monitoring device under two physiological conditions: resting and post-activity. Measurements were obtained from a single subject, where each condition consisted of ten individual trials. For every trial, SpO2 and heart rate values were recorded simultaneously using the developed device and a commercial pulse oximeter, which served as the reference instrument. The optical signals captured by the MAX30102 sensor were processed using the built-in photoplethysmography (PPG) algorithm, which includes light-intensity acquisition, noise reduction, and peak detection for heart rate estimation SpO2values were computed using the standard red-to-infrared ratio method, which is widely used in pulse oximetry for extracting oxygen saturation from reflected light signals.

To assess the reliability and consistency of the system, the output of the developed device was compared with the reference measurements obtained from the commercial pulse oximeter. Three evaluation metrics were employed: percentage error, accuracy, and precision. Percentage error was used to quantify the deviation of the measured values from the reference values, while accuracy indicated the proportion of measurements that fell within an acceptable tolerance range. Precision was calculated to determine the consistency of repeated measurements under the same experimental condition. These metrics provided a comprehensive assessment of how effectively the proposed device estimated SpO2 values during both resting and post-activity states.

E. Hardware Implementation

In conducting a simulation of the IoT-based blood oxygen level (SpO_2) monitoring device using Wokwi, the component setup must follow the pin configuration that complies with the I2C communication standard. The MAX30102 sensor and the SSD1306 OLED display both utilize the I2C protocol, allowing them to be connected to the ESP32 microcontroller. For the MAX30102 sensor connection, the VIN pin is connected to the ESP32's 3.3V pin, GND to GND, SDA to GPIO 21, and SCL to GPIO 22. Meanwhile, for the SSD1306 OLED display, the VCC pin is connected to 3.3V, GND to GND, SDA to GPIO 4, and SCL to GPIO 15. Additionally, the buzzer's positive pin is connected to GPIO 18 and its negative pin is connected to ground. After all components are connected, the next step is to run the simulation and program through the Arduino IDE to virtually test the device's functionality beforehand. This simulation stage aims to ensure that the circuit operates as expected before physical assembly, thereby simplifying the implementation process and minimizing errors during hardware construction.

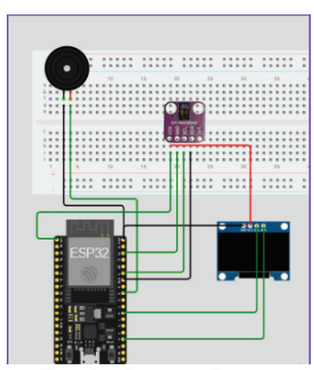


Fig. 2 : System Circuit

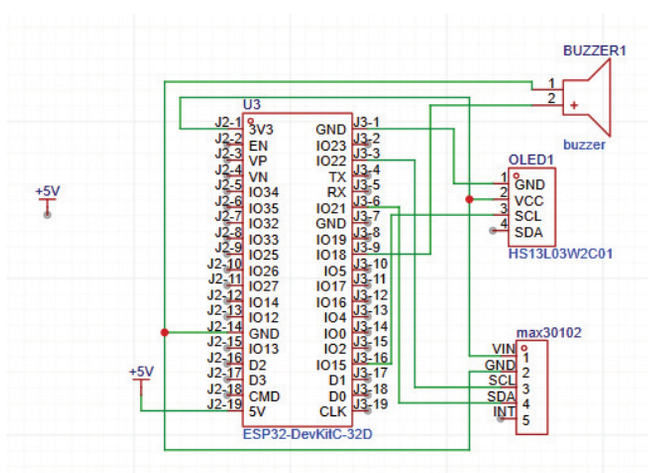


Fig. 3: Wiring Diagram

III. RESULT

The system evaluation was performed to assess the accuracy and consistency of SpO_2 and heart rate measurements obtained from the ESP32–MAX30102 device compared with a commercial pulse oximeter. The first experiment was conducted under normal resting conditions, where ten measurement trials were recorded using both the prototype and the reference device. The prototype produced an average SpO_2 value of 95.5%, while the reference device showed 98.7%, resulting in an average deviation of 3.2%. For heart rate measurements, the prototype recorded an average of 94 BPM, compared with 86.1 BPM from the reference device, indicating a deviation of 7.9 BPM. A second experiment was conducted after light physical activity to evaluate system performance under increased physiological variability. In this condition, the prototype measured an average SpO_2 of 95.3%, while the reference device recorded 96.7%, yielding a deviation of 1.4%. Heart rate measurements showed an average of 95.4 BPM for the prototype and 77.9 BPM for the reference device, reflecting a higher deviation due to motion-induced noise commonly associated with optical PPG sensors. Overall, the results indicate that the system performs reliably in detecting general SpO_2 trends under both resting and post-activity conditions, although heart rate readings remain more sensitive to movement and require further signal stabilization. Sample measurement results are summarized in Table I and Table II.

TABLE I : Resting Condition Measurement Results

	measurement comparison (normal condition)			
no	saturation %(device)	BPM (device)	saturation %(reference)	BPM (reference)
1	95	100	99	88
2	95	87	99	83
3	96	95	97	87
4	95	93	99	84
5	96	93	99	85
6	95	95	99	87
7	97	98	99	84
8	96	93	99	93
9	95	97	98	86
10	95	89	99	84

For the resting-condition measurements, the system produced an average SpO_2 value of 95.5 percent compared with 98.7 percent from the reference device. The calculated mean percent error was 3.24 percent, corresponding to an overall accuracy of 96.76 percent. This accuracy indicates that the prototype closely follows the reference readings under stable conditions. Precision analysis based on the standard deviation of repeated measurements yielded a value of 0.71, demonstrating consistent sensor outputs with minimal variation across trials. Overall, these results suggest that the system performs reliably during resting conditions, with deviations mainly caused by minor optical fluctuations and normal physiological variability rather than device-related errors.

TABLE II: Post-Activity Measurement Results

	Measuremen coomparison (aftter physical activity)			
No	Saturation % (device)	BPM (device)	Saturation % (reference)	BPM (reference)
1	95	95	100	76
2	95	99	99	80
3	96	98	98	77
4	96	103	97	77
5	94	94	80	69
6	96	93	99	74
7	95	87	99	77
8	96	95	99	82
9	95	93	98	93
10	95	97	98	80

For the post-activity measurements, the system recorded an average SpO_2 value of 95.3 percent, while the reference pulse oximeter showed an average of 96.7 percent. The calculated mean percent error was 1.45 percent, corresponding to an overall accuracy of 98.55 percent. Compared to the resting-condition dataset, the error decreased due to more stable optical readings despite physiological changes caused by light physical activity. Precision analysis yielded a standard deviation of approximately 0.63, indicating consistent measurements across repeated trials. Overall, these results demonstrate that the device maintains reliable performance after activity, with deviations remaining within an acceptable range for non-invasive optical sensing.

The testing using the prototype device was conducted indoors under controlled room-temperature conditions with illumination from artificial lighting, as shown in the following figure. This environment was chosen to ensure stable measurement conditions and minimize external factors that could affect SpO_2 accuracy.



Fig. 4: SpO_2 measurement using the proposed device.

Both measurement sessions were conducted using the same reference pulse oximeter, and all trials were performed under identical environmental conditions to ensure dataset consistency. Measurements were taken in the same indoor setting with controlled lighting and minimal external interference, while subjects maintained a similar posture during each recording. This uniform setup minimized variability from external factors, ensuring that differences between the prototype and reference device were mainly attributed to sensor performance rather than environmental or procedural inconsistencies.



Fig. 5: SpO_2 measurement using the reference device.

The measurement data obtained from the prototype device is transmitted wirelessly to the Blynk IoT platform, allowing users to access and review the information either through the mobile application or the web-based dashboard on a computer. Once the data reaches the platform, Blynk displays several key parameters, including real-time heart rate (BPM), SpO_2 levels, and an automated interpretation indicating whether the measured values suggest a potential asthma condition. In addition to real-time monitoring, the platform maintains a historical record of the measurements, enabling users to view trends and past readings for up to seven days through Blynk's cloud storage. This integration provides flexibility for remote monitoring and ensures that users or caregivers can track respiratory condition changes over time, even when they are not physically near the device.

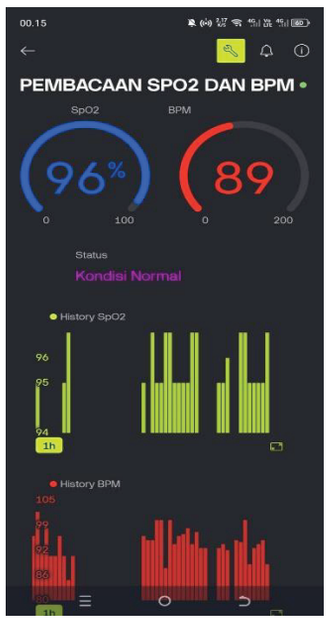


Fig. 6: Blynk dashboard

IV. DISCUSSIONS

The experimental results demonstrate that the proposed IoT-based SpO_2 monitoring system is capable of performing non-invasive measurements with a performance that is reasonably comparable to a commercial pulse oximeter. In the resting-condition trials, the system achieved a mean percent error of 3.24 percent and an accuracy of 96.76 percent, with a precision value of 0.71, indicating that the device can generate stable readings under controlled conditions. After light physical activity, the measurement accuracy improved to 98.55 percent with a reduced percent error of 1.45 percent and a precision of approximately 0.63. This improvement suggests that the device maintains reliable performance even when physiological variations occur, although slight fluctuations remain inherent to optical PPG sensing. These deviations are likely caused by minor motion artifacts, changes in peripheral perfusion, and natural variability in finger placement during repeated measurements.

The comparison between the prototype and the reference device was conducted under identical environmental conditions and using the same measurement procedures, ensuring that differences in sensor output resulted primarily from device characteristics rather than external influences. Factors such as indoor lighting, finger pressure, and subject posture were kept constant to minimize variability and strengthen the validity of the comparison. The consistency of these conditions supports the reliability of the error and precision analysis presented in this study.

Beyond local measurement, the integration of the system with the Blynk IoT platform provides an additional layer of functionality that enhances its practicality for real-world use. The wireless transmission of SpO_2 and BPM data allows users to access their physiological information through either a mobile application or a web dashboard, enabling convenient remote monitoring. The platform also generates an automatic assessment indicating whether the measured values may correspond to asthma-related symptoms, giving users informed insight into their respiratory condition. Furthermore, Blynk's cloud-based data storage enables historical tracking for up to seven days, allowing users and caregivers to observe trends that may indicate early signs of respiratory deterioration. This capability is particularly beneficial for individuals with asthma, as ongoing monitoring can support timely interventions and preventive care.

Although the system performs well in controlled scenarios, several limitations should be noted. As with other optical PPG-based systems, the measurements remain susceptible to motion artifacts, thermal fluctuations, and variations in skin perfusion. These factors could affect accuracy during heavy physical activity or outdoor use, suggesting a need for improved filtering algorithms or adaptive signal processing techniques. Additionally, the current system has not yet been validated through clinical evaluations involving a larger and more diverse population, which will be necessary to confirm its reliability for broader health-monitoring applications. Future work should focus on enhancing the system's robustness to motion-induced noise, expanding the number of test subjects, and integrating advanced filtering or machine learning models to improve signal interpretation. The incorporation of additional physiological indicators, such as respiratory rate or activity level, could further strengthen the system's ability to detect early asthma symptoms. With continued development, this IoT-based monitoring solution has the potential to evolve into a practical and accessible tool for continuous respiratory health assessment, providing valuable support for individuals with asthma and contributing to more proactive and personalized healthcare.

V. CONCLUSION

The experimental evaluation demonstrates that the proposed IoT-based SpO_2 monitoring device is capable of producing measurement results that closely approximate those of a commercial reference pulse oximeter. Under resting conditions, the system achieved a mean percent error of 3.24 percent, corresponding to an accuracy of 96.76 percent and a precision value of 0.71, indicating stable and repeatable sensor performance. After light physical activity, the system maintained high reliability, yielding an improved percent error of 1.45 percent, an accuracy of 98.55 percent, and a precision of approximately 0.63. These results confirm that the device is able to track physiological variations effectively while retaining consistency across repeated measurements. The integration of the system with the Blynk platform enables real-time remote access to SpO_2 and BPM readings, automated asthma-risk interpretation, and cloud-based historical data storage for up to seven days, thereby enhancing its usability for continuous respiratory monitoring. Although the system performs well under controlled indoor conditions, further development is required to improve robustness against motion artifacts and to validate performance using a larger and more diverse subject population. Overall, the system demonstrates strong potential as a low-cost and portable solution for ongoing respiratory assessment, particularly for individuals with asthma, and represents a promising advancement toward accessible IoT-enabled health-monitoring technologies.

ACKNOWLEDGMENT

The authors would like to express their sincere appreciation to the Department of Biomedical Engineering at Telkom University Purwokerto for providing academic support and access to laboratory facilities essential for completing this research. The authors also acknowledge the contributions of colleagues and participants who assisted during the measurement, testing, and validation processes. Their guidance, cooperation, and technical insights greatly enhanced the development and evaluation of the proposed system. The constructive feedback received during the preparation of this manuscript is also gratefully acknowledged for helping improve the overall quality of this work.

REFERENCES

[1] Jeconiah Gabriello Dandan ; Mual Bobby E. Parhusip; Agnes Frethernety, "LITERATURE REVIEW : GAMBARAN FAKTOR-FAKTOR PENCETUS ASMA PADA PASIEN ASMA," *Jurnal Kedokteran Universitas Palangka Raya*, vol. 10, no. 2, pp. 1–5, Oct. 2022, doi: 10.37304/jkupr.v10i2.3492.

[2] Abraham Christopher; Yuwono M. Dinata, "View of Rancang Bangun Sistem Pemantauan Jarak Jauh Denyut Nadi, Saturasi Oksigen, dan Suhu Tubuh pada Orang Sakit di Rumah," *JUISI*, vol. 08 no 01, pp. 1–11, Feb. 2022.

[3] Baiq Nurul Laili; Budiani Destyningtias; Sri Heranurweni, "RANCANG BANGUN PULSE OXIMETRY DENGAN SISTEM MONITORING

INTERNET OF THING (IOT)," *Jurnal Teknologi Elektro dan Kesehatan Terapan*, vol. 5, no. 2, pp. 1–9, 2020.

[4] REFNI WAHYUNI S.KOM. M.TI, "PERANCANGAN PULSE OXIMETRY BERBASIS IOT SEBAGAI DETEKSI DINI GEJALA COVID-19," Pekanbaru, 2022.

[5] Fredy Pratama; Achmad Rizal; Muhammad Hablul Barri, "Integrasi Sistem Hardware Pada Alat Real-Time Heart Rate Variability Monitor Menggunakan Pulse Sensor Max30102," Oct. 2024.

[6] Kemalasari; Rochmad Mochammad, "DETEKSI KADAR SATURASI OKSIGEN DARAH (SpO2) DAN DETAK JANTUNG SECARA NON-INVASIF DENGAN SENSOR CHIP MAX30100," *Jurnal Nasional Teknologi Terapan (JNTT)*, vol. 4, no. 1, pp. 35–50, Jun. 2022, doi: 10.22146/jntt.v4i1.4804.

[7] Onubogu Uchenna Chinweokwu; A Ayuk, "pulse oximetry and peak expiratory flow rate correlations in acute asthma exacerbation in children," *Niger J Clin Pract*, vol. 25, no. 11, pp. 1896–1903, Nov. 2022, doi: 10.4103/njcp.njcp_376_22.

[8] Muthmainnah Muthmainnah; Deni Bako Tabriawan, "Prototipe Alat Ukur Detak Jantung Menggunakan Sensor MAX30102 Berbasis Internet of Things (IoT) ESP8266 dan Blynk," *Jurnal Informatika Sunan Kalijaga)*, vol. 7, no. 3, pp. 163–176, 2022.

[9] Sehrash Ashraf; Shahnaz Parveen Khattak; Mohammad Tariq Iqbal, "Design and Implementation of an Open-Source and Internet-of-Things-Based Health Monitoring System," *Journal of Low Power Electronics and Applications*, vol. 13, no. 4, pp. 1–19, Dec. 2023, doi: 10.3390/jlpea13040057.

[10] Naval Indra Waskita; Hizkia Menahem Tandungan; Ridhan Hafizh; Syifa Jauza Suwaendi; Magfirawaty, "ESP32 and MAX30100 with Chebyshev Filter for Enhanced Heart and Oxygen Measurement," *Jurnal RESTI*, vol. 8, no. 5, pp. 651–657, Oct. 2024, doi: 10.29207/resti.v8i5.5945.

[11] Ali Mohammed Abd-ALameer, "Measuring of Cytokines TNF-α and IFN-γ Concentration Levels in Asthma Patients," *IJMI : International Journal Multidisciplinary International Journal Multidisciplinary*, vol. 2, no. 2, pp. 199–205, May 2025, doi: 10.61796/ijmi.v2i2.310.

[12] C. Ribeiro de Andrade; M.C. Duarte; P. Camargos, "Correlations between pulse oximetry and peak expiratory flow in acute asthma," *Brazilian Journal of Medical and Biological Research*, vol. 40, no. 4, pp. 485–490, 2020, [Online]. Available: www.bjournal.com.br

[13] A. Yulia, W. Lestari, P. DIV Keperawatan, P. Kemenkes Bengkulu, and J. Keperawatan, "Pengaruh Nafas Dalam dan Posisi Terhadap Saturasi Oksigen dan Frekuensi Nafas Pada Pasien Asma," *Jurnal Keperawatan Raflesia*, vol. 1, no. 1, pp. 67–75, May 2019, doi: 10.33088/jkr.vlil.398.

[14] Candra Rizki Nugroho; Elvan Yuniarti; Ambran Hartono, "Alat Pengukur Saturasi Oksigen Dalam Darah Menggunakan Metode Photoplethysmograph Reflectance," *Al-Fiziya: Journal of Materials Science, Geophysics, Instrumentation and Theoretical Physics*, vol. 3, no. 2, pp. 84–93, Dec. 2020, doi: 10.15408/fiziya.v3i2.17721.

[15] I. Wahyu Ramadhan, S. Adinandra, and P. Studi Magister Teknik Rekayasa Elektro, "Penerapan IoT dalam Sistem Monitoring Kesehatan: Inovasi dan Implementasi Implementation of IoT in Health Monitoring Systems: Innovation and Implementation," *Techno.Com*, vol. 23, no. 4, p. 763, Nov. 2024.

[16] M. Z. Abdillah, F. Baskoro, M. S. Zuhrie, and N. Kholis, "Jurnal Teknik Elektro: Electronic Control, Telecomunication, Computer Information and Power System Pengembangan Perangkat Pembelajaran Mikrokontroler Esp32 Berbasis Internet of Things & Bluetooth di SMKN 2 Surabaya," *Jurnal Teknik Elektro: Electronic Control, Telecomunication, Computer Information and Power System*, vol. 9, no. 2, pp. 110–123, Sep. 2024.

[17] M. Muthmainnah, Deni Bako Tabriawan, and Imam Tazi, "Karakterisasi Sensor MAX30102 Sebagai Alat Ukur Detak Jantung dan Suhu Tubuh Berbasis Photoplethysmograph," *JURNAL PENDIDIKAN MIPA*, vol. 12, no. 3, pp. 726–731, Sep. 2022, doi: 10.37630/jpm.v12i3.655.

Secure OTA Firmware Update Using BlockChain and Role Based Access Control

HYEONG GEON PAK, SEUNG YEOB NAM*

* *Department of Information and Communication Engineering, Yeungnam University, Gyeongsan 38541, South Korea*

Emails: *hyunggun@yu.ac.kr*, *synam@ynu.ac.kr*

***Abstract*—As vehicles evolve, the importance of safe and reliable firmware updates grows. Traditional OTA firmware update methods rely on centralized networks, creating a single point of failure problem. To address this, this research utilizes blockchain to perform firmware updates in a distributed network environment, eliminating the single point of failure issue. Furthermore, it mitigates the limitations of existing blockchains—namely scalability and the issue of anyone being able to register blocks—by utilizing the InterPlanetary File System (IPFS) and Role-Based Access Control (RBAC). IPFS resolves the storage capacity limitations of the blockchain, while RBAC ensures only authorized user addresses can register blocks on the blockchain using smart contracts, thereby addressing key challenges. Using Geth, we create a local private network and conduct experiments in the MetaMask and Python Web3 environments. The results confirm that only the administrator address can register the hash value in the smart contract on the blockchain, while unauthorized accounts are denied access, verifying that data can be registered on the blockchain according to permissions. Finally, the system undergoes a qualitative evaluation in terms of security, scalability, and efficiency.**

***Index Terms*—OTA, Blockchain, Smart Contract, Role-Based Access Control**

I. Introduction

The development of industry has always progressed alongside transportation. Vehicles rapidly integrated into human life as soon as they appeared. As human life advanced, vehicles also began to evolve rapidly. Therefore, today, with the advent of autonomous vehicles, many vehicles are evolving beyond mere means of transportation into software-defined vehicles. Vehicles now contain hundreds of electronic control units (ECUs). Software within the vehicle controls and manages these ECUs, communicates with objects like the vehicle itself, the road, and terrain, and manages energy, establishing itself as an essential element. Consequently, continuous software updates for vehicles are imperative, emerging as a core factor for enhancing vehicle stability and functionality. If a vehicle's software is attacked, it can lead to situations where not just vehicle malfunctions occur, but the driver's life is put at risk. Therefore, safely updating vehicle firmware can be considered the most critical factor in the current automotive environment [1].

Existing firmware updates utilize a centralized server-based structure. Such firmware updates pose a high risk of exposing the entire system to danger if the server is hacked, and they present a single point of failure issue where service becomes impossible if the server fails or is attacked. To safely update firmware, this research proposes conducting firmware updates using blockchain technology, which offers decentralization and tamper-proofing, instead of relying on a centralized structure. By proposing firmware updates based on verified firmware information registered in the blockchain hash table, we mitigate the drawbacks of the centralized server-based structure, enabling more stable updates.

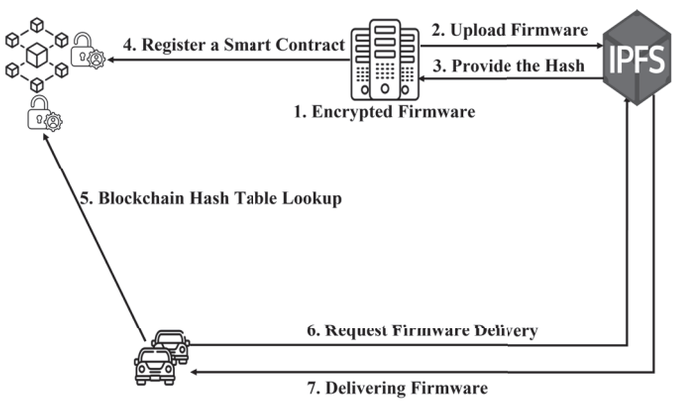


Fig. 1. Update Method

However, structures utilizing blockchain also have drawbacks. Specifically, there are limitations in storage capacity and the negative aspect that anyone can adjust the blockchain's hash table through smart contracts. This study addresses the limited storage capacity by employing the distributed file system Inter Planetary File System (IPFS) and uses Role-Based Access Control (RBAC) to ensure only authorized users can register on the blockchain via smart contracts [2].

A. Contribution

- Existing firmware updates utilize centralized servers for deployment, which carries the inherent drawback of a potential single point of failure. This research addresses the single point of failure issue by leveraging the decentralized nature of blockchain for firmware updates, while simultaneously reducing the burden on central servers.
- By leveraging blockchain, we can determine that there are limitations to using it in an OTA update system due to storage issues and the fact that anyone can register blocks on the blockchain. Therefore, we utilize IPFS to resolve the blockchain's storage problem and employ RBAC when writing smart contracts to protect the hash table,

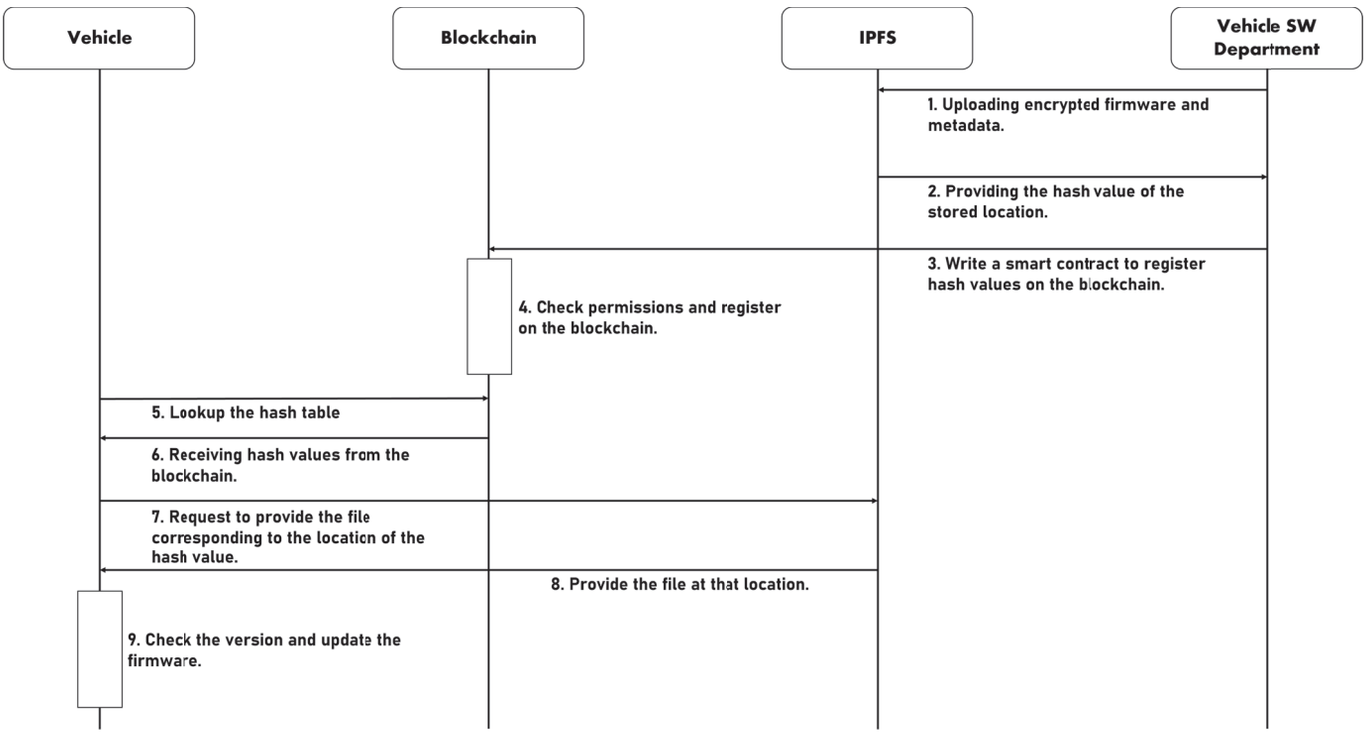


Fig. 2. Secure OTA Firmware Update with Blockchain

preventing malicious users from arbitrarily registering hashes into the blockchain's hash table.

B. Paper Organization

The structure of this paper is as follows: Chapter 1 discusses the motivation and contributions of this research; Chapter 2 explains the background; Chapters 3 and 4 present the mechanism and experiments proposed in this study; Chapter 5 conducts a qualitative evaluation; and Chapter 6 concludes the paper.

II. Background

A. Blockchain

It is a distributed ledger, not a traditional centralized network. It functions as a distributed database storing data across multiple computers, making it tamper-proof. Each transaction is created and formed into a block, which is then stored and linked by the distributed computers. Transactions are validated through a consensus mechanism and registered across the entire network. Each transaction is grouped into a block and linked to the next block, maintaining a secure and transparent chain structure. This distributed chain structure of blockchain ensures data integrity. Since each subsequent block contains the information from the preceding block, it prevents block tampering and helps form a more stable network [3].

B. Smart Contract

A smart contract is a system that uses code to record transaction details between contracting parties and registers that information on the blockchain. It automatically executes the contract when its conditions are met. Smart contracts are created on the blockchain network. They enable various information and data to be recorded within the blocks generated by the blockchain. The code and execution history of a smart contract are also shared with all nodes via the blockchain, making it difficult to tamper with the smart contract. When a contract is executed, it is not recorded on the computers of the contracting parties. Instead, it is verified by the many nodes participating in the blockchain network. Upon passing verification, the information and data can be stored in a block. This means the contract details are stored in a blockchain block, making the information and data visible to everyone, ensuring transparency. Furthermore, since multiple nodes verify the contract outcome, it offers the advantage of high reliability [4].

C. IPFS

IPFS stores each file across multiple blocks, each possessing a unique hash value. Each node stores only the files it is interested in within its storage. Through index information, it is possible to determine which nodes store which files. By querying a filename and using the received hash, one can locate and download the file from the node possessing it, where it resides at the location matching the hash value. Access is based on the content of the specific data itself, not its stored location [5].

III. Firmware Update Mechanism

The firmware update proceeds as shown in Fig. 1. The firmware is encrypted by the Vehicle Software Department

(VSD), which generates an encryption key. The encrypted firmware is uploaded to IPFS, and the key used for encryption is shared with the vehicle so it can decrypt the firmware. This section outlines the update sequence as shown in Fig. 2.

A. Firmware Encryption and Upload

The update procedure begins with the firmware encryption of the VSD. The firmware encryption process involves the VSD generating a Key to encrypt the firmware. This Key is shared with the vehicle. The VSD uploads the encrypted firmware to IPFS as shown in Fig. 2. It also generates metadata containing the firmware information and uploads this to IPFS as well. The metadata includes the firmware name, registration date, and timestamp. IPFS is used to address blockchain scalability issues. It is a protocol for storing and sharing files through a distributed computer network. It generates a hash value, a unique identifier based on the file's content. This identifier uniquely locates the file and prevents tampering with the original file, enhancing data reliability and stability. Even when uploading firmware and metadata to IPFS, it helps securely share and store data. Therefore, VSD retrieves and stores the hash values of the locations where data uploaded to IPFS is stored [6].

B. Firmware Hash Value Registration

VSD utilizes smart contracts to register stored hash values on the blockchain. When writing the smart contract, access control is employed to ensure only individuals with registered addresses can register hash values in the blockchain hash table. As shown in Fig. 3, the access control table grants permissions based on roles. Only authorized users can verify their permissions in the access control and register hash values in the hash table. This ensures only registered addresses can register hash values, addressing the vulnerability where anyone could arbitrarily register in the blockchain hash table. Only registered users can perform registrations, while other users possess read-only permissions. Consequently, client vehicles can only query values from the blockchain's hash table. Therefore, VSDs possessing pre-registered addresses register hash values received from IPFS into the blockchain hash table [7].

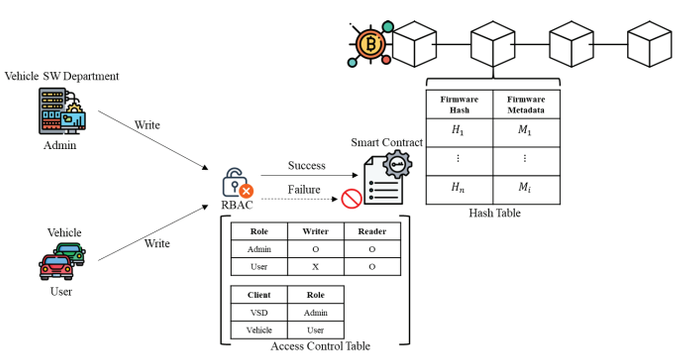


Fig. 3. Access Control Function

C. Lookup the Hash Table and Update

Vehicles connected to the blockchain can query the hash table registered on the blockchain. Since vehicles possess addresses with read-only permissions, they cannot register other information on the blockchain via smart contracts and can only perform queries. Vehicles retrieve and store the hash values, including the encrypted firmware hash and metadata hash values. They then use these hash values to download the data from its stored location on IPFS. The vehicle downloads the encrypted firmware and metadata. It compares the downloaded metadata with the information about the currently installed software to verify if it is the next firmware version. It also compares the hash values to confirm the downloaded firmware is the correct one before proceeding with the firmware update.

IV. Implement

The experiment protects smart contracts by utilizing role-based access control, considered crucial in this mechanism. It creates a local private network on the local computer environment using Geth and assists in registering smart contracts on the local blockchain using Remix.ide [8]. Additionally, it utilizes the Ethereum wallet MetaMask to visually monitor activities from blockchain accounts [9]. Finally, it connects to the local private blockchain network using the Python web3 environment, enabling visual verification of the smart contract's access control environment [10].



Fig. 4. Private Blockchain Network

Install Geth and create a private network. The blockchain's consensus algorithm utilizes POW. When you run the blockchain as shown in Fig. 4, you can observe blocks being generated on your local computer within the blockchain network, confirming the network's activation. After activating the local blockchain, deploy the smart contract using Remix.ide. The smart contract code is written to utilize role-based access control, ensuring only addresses with permission to register hash values can register blocks on the blockchain. Connect the private blockchain to MetaMask to link account addresses within the network. Simultaneously, deploying the smart contract and approving the contract becomes visually verifiable. After deploying the smart contract, a connection request is sent to MetaMask as shown in Fig. 5. Upon confirming the request, the hash value and admin address are registered in the blockchain block [11].

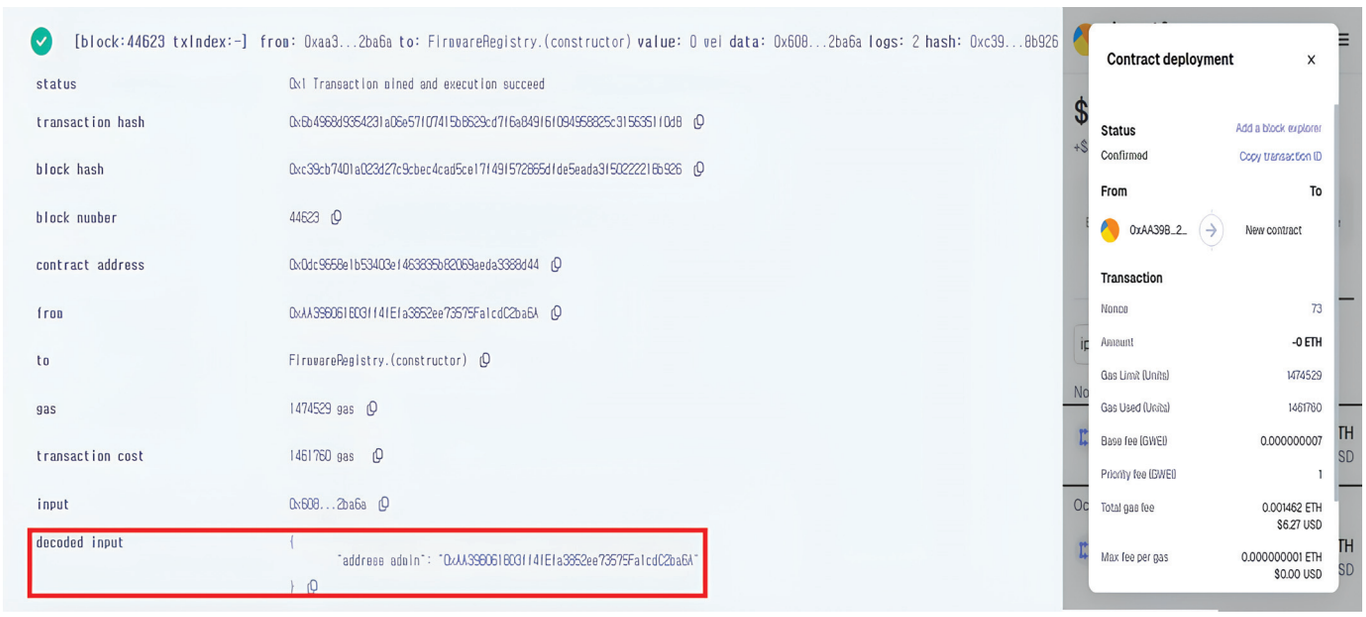


Fig. 5. Registered Smart Contract

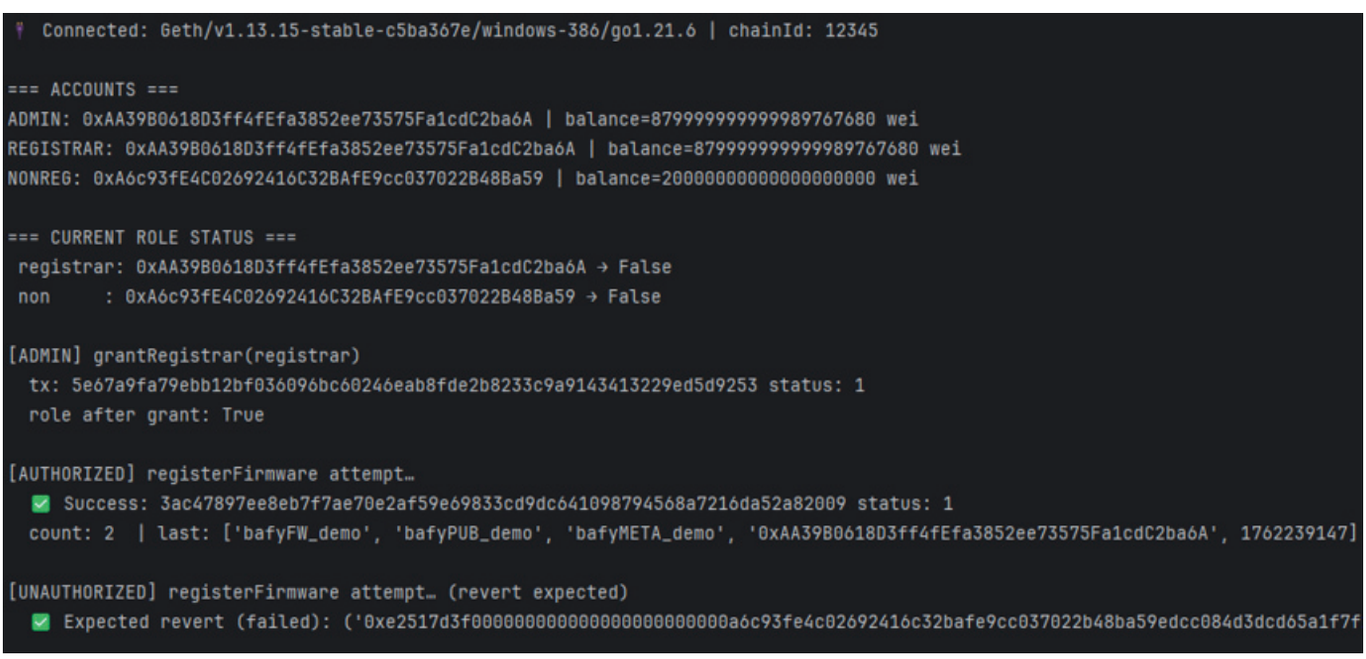


Fig. 6. Blockchain Access Control Experiment

Using the Python web3 environment, we proceed with the procedure to verify that role-based access control functions correctly in the smart contract by integrating the blockchain with the Python practice environment. Observing Fig. 6, we can see that only the admin address registered in the Python practice environment can register an arbitrary hash value, while addresses from other accounts cannot register a hash value. Addresses belonging to other accounts fail to register hash values, encountering errors and returning error codes. Conversely, addresses matching the admin account address successfully register hash values, confirming that role-based access control within the smart contract functions correctly.

V. Performance Evaluation

We will analyze the qualitative excellence of the proposed blockchain-based OTA firmware update system. We will evaluate it in terms of security, scalability, and efficiency compared to existing blockchain methods.

A. Security

One drawback of existing blockchains is that anyone can create blocks. The proposed solution applies Role-Based Access Control (RBAC) to smart contracts, allowing only accounts registered as administrators to record firmware hash values on the blockchain. This effectively prevents privilege abuse and unauthorized registration attacks. Consequently, the overall OTA process gains enhanced resilience against data forgery, tampering, and privilege escalation attacks.

B. Scalability

Blockchain inherently has limited storage space, making it inefficient to directly record all firmware data. To address this, the proposed system adopts IPFS to store firmware and metadata on external distributed storage, registering only the file's hash value on the blockchain. This significantly reduces storage burden. Even as the number of nodes increases, it distributes system load, contributing to scalability and minimizing performance degradation.

C. Efficiency

The existing update method performs all encryption, distribution, and verification on a central server, causing concentrated load on the server. In the proposed method, the VSD pre-encrypts the firmware before uploading it to IPFS. Vehicle nodes then query and verify only the necessary hashes via smart contracts, distributing computational and network load. Furthermore, IPFS's content-based address structure prevents redundant storage of identical data, improving storage efficiency.

VI. Conclusion

This study proposes a new vehicle firmware update mechanism that combines blockchain, IPFS, and role-based access control to improve the single point of failure and security vulnerabilities inherent in existing centralized OTA firmware update methods. The proposed system solves blockchain scalability issues by uploading encrypted firmware and metadata to the distributed file system IPFS. It also increases efficiency by registering only the hash value of the data's storage location on the smart contract. Furthermore, it enhances blockchain security by clearly distinguishing addresses that can write smart contracts from those that cannot, through role-based access control. The qualitative evaluation results indicate that the proposed mechanism addresses the scalability limitations of existing blockchains, enhances efficiency, and compensates for the transparency shortcomings in security, enabling stable large-scale vehicle updates.

Future research will apply the proposed system to real vehicle network environments to refine detailed models concerning transaction processing, key management overhead, and other aspects. We anticipate that the proposed mechanism will enable its utilization as a secure and reliable update infrastructure within smart mobility environments.

Acknowledgment

This work was supported by the Basic Science Research Program through the National Research Foundation (NRF) funded by the Ministry of Education under Grant 2021R1A6A1A03039493.

References

[1] A. Ghosal, S. Halder, and M. Conti, "Secure over-the-air software update for connected vehicles," *Computer Networks*, vol. 218, p. 109394, 2022.
[2] R. S. Sandhu, "Role-based access control," in *Advances in computers*, vol. 46, pp. 237–286, Elsevier, 1998.
[3] S. Nakamoto and A. Bitcoin, "A peer-to-peer electronic cash system," *Bitcoin.–URL: https://bitcoin. org/bitcoin. pdf*, vol. 4, no. 2, p. 15, 2008.
[4] G. Wood *et al.*, "Ethereum: A secure decentralised generalised transaction ledger," *Ethereum project yellow paper*, vol. 151, no. 2014, pp. 1–32, 2014.
[5] J. Benet, "Ipfs-content addressed, versioned, p2p file system," *arXiv preprint arXiv:1407.3561*, 2014.
[6] A. Shoker, F. Alves, and P. Esteves-Verissimo, "Scaiota: Scalable secure over-the-air software updates for vehicles," in *2023 42nd International Symposium on Reliable Distributed Systems (SRDS)*, pp. 151–161, IEEE, 2023.
[7] A. Kovacevic and N. Gligoric, "Enhancing security of automotive ota firmware updates via decentralized identifiers and distributed ledger technology," *Electronics*, vol. 13, no. 23, p. 4640, 2024.
[8] Ethereum Foundation, *Geth – Go Ethereum Documentation*, 2025. Accessed: 2025-09-12.
[9] Meta Platforms Inc., *React – A JavaScript Library for Building User Interfaces*, 2025. Accessed: 2025-09-12.
[10] QuickNode, "How to integrate ipfs with ethereum." https://www.quicknode.com/guides/ethereum-development/dapps/how-to-integrate-ipfs-with-ethereum, 2025. Accessed: 2025-09-12.
[11] N. Sangeeta and S. Y. Nam, "Blockchain and interplanetary file system (ipfs)-based data storage system for vehicular networks with keyword search capability," *Electronics*, vol. 12, no. 7, p. 1545, 2023.

An IRS-UAV-Assisted EMF-Aware PD-NOMA Framework for Uplink Communication in Dense Urban Networks

Maliha Binte Mukhles*, Sung Won Kim†, M Ishtiaque Ahmed‡
**Department of Information and Communication Engineering, Yeungnam University, Gyeongsan-si, 38541, Republic of Korea*
†School of Computer Science and Engineering, Yeungnam University, Gyeongsan-si, 38541, Republic of Korea
‡Department of Environmental Science, Jahangirnagar University, Savar, Bangladesh
maliha01.winlab@yu.ac.kr, swon@yu.ac.kr, ishtiaqueanu04@gmail.com

***Abstract*—The increasing number of wireless devices together with upcoming 6G networks creates worries about electromagnetic field (EMF) radiation in densely populated city areas. The research presents an uplink system which unites Intelligent Reflecting Surfaces (IRS) with Unmanned Aerial Vehicles (UAVs) and Power-Domain Non-Orthogonal Multiple Access (PD-NOMA) to minimize EMF exposure while maintaining excellent spectral efficiency and Quality-of-Service (QoS). The system design uses machine learning to cluster users and IRS to model channels and water-filling power allocation to maximize resource efficiency. The proposed system demonstrates through simulations that it reduces EMF exposure by 93% relative to OFDMA and UAV-only PD-NOMA while achieving 1.8 times better spectral efficiency and maintaining QoS satisfaction above 98%. The research demonstrates that IRS-UAV PD-NOMA technology provides a scalable solution for upcoming 6G and NTNs that prioritizes user health.**

Keywords—Electromagnetic Field (EMF) Exposure; Intelligent Reflecting Surface (IRS); Unmanned Aerial Vehicle (UAV); Power-Domain NOMA (PD-NOMA)

I. INTRODUCTION

The high capacity and low latency communications have been fastened due to the exponential growth of the wireless personal devices and the advent of the data intensive applications. With the global upgrade to sixth generation (6G) networks, terrestrial and non-terrestrial infrastructures are likely to be the key to achieving a wide range of service demands. This high rate of wireless connectivity expansion has however also brought up issues about EMF exposure especially as far as the highly populated urban surroundings where users want to use their devices to stay in constant contact for data transmission purpose [3,4].

The existing design of 5G has also been a matter of urgent research focus to ensure that it is needed to be health conscious and performance efficient. The traditional multiple access methods like the Orthogonal Frequency Division Multiple Access (OFDMA) are reliable in terms of connectivity but have low spectral efficiency and scalability in a dense user environment. PD-NOMA has been proposed as a savior to the problem by offering multiple users to share a subcarrier using power level differentiation and successive interference cancellation (SIC) [5]. Although PD-NOMA enhances throughput, it is limited by channel conditions and interference control particularly in urban environments that have extreme Non-Line of Sight (NLoS) propagation [2,3]. New opportunities to overcome such limitations are provided by recent developments in IRS and UAVs. IRS technology allows passive forming of beams through reconfiguring the wireless environment thus increasing signal strength without extra emissions of EMF. In their turn, UAVs offer flexible Line of Sight (LoS) connections and they serve as portable relays to increase coverage in regions with inadequate terrestrial infrastructure. The combined use of IRS and UAV is thus able to establish a synergistic system that is more efficient in terms of spectral usage, less interference, and the elimination of EMF exposure.

Irrespective of these benefits, the incorporation of IRS and UAVs in PD-NOMAs presents a number of challenges. Resource distribution, clustering of users and power control should be meticulously defined to balance the throughput and minimization of the EMF. In addition, the dynamism of the UAV positioning and the massive construction of the IRS elements demand adaptive algorithms that can adapt to the dynamic variations of the channels in real time [7]. The challenges necessitate a unified framework that is a combination of highly sophisticated optimization methods and health-conscious communication design. To effectively consider the EMF exposure limits in the design of the system, we recommend in this paper an

uplink framework of IRS UAV assisted PD-NOMA, which explicitly considers EMF exposure limits. Contributions made by this work are summarized as follows:

- Development of a system model integrating IRS-assisted propagation, UAV relaying, and PD-NOMA multiplexing.
- Introduction of machine learning-based clustering and water-filling power allocation to optimize resource usage while satisfying SIC and QoS requirements.
- Evaluation of the framework using the Dose metric to quantify cumulative EMF exposure.
- Simulation-based validation shows substantial EMF exposure cuts, spectral efficiency gains, and QoS satisfaction.

This study adds performance and health concerns to the creation of sustainable and scalable uplink communication systems in future 6G and non-terrestrial networks (NTNs).

II. RELATED WORK

Recent research has also been conducted into the combination of IRS, UAVs and NOMA to facilitate the work of wireless communication, and the interest in EMF aware design becomes more and more popular. Firstly, Ahmed et al. [2] explored the issue of UAV mounted active IRS in NOMA networks and showed that aerial deployment of IRS can help to achieve a remarkable increase in air to ground communication reliability. Their study emphasized the role of co-optimal UAV positioning and IRS placement in order to attain stable connectivity. Second, He et al. [1] focused on resource allocation and the optimization of the trajectory in UAV aided satellite IoT systems based on NOMA on the power constrained devices. Their findings demonstrated that UAV relaying is able to strike a fair balance between access and spectral efficiency in large scale IoT implementations. Third, Ahmed et al. [6] suggested joint optimization of IRS assisted in multi carrier NOMA 6G vehicle to infrastructure communication with a particular focus on the diagonal IRS configurations. Their results proved that IRS can significantly increase the throughput and decrease interference in the highly dynamic motive sphere. Lastly, Salem et al. [3] introduced a survey on EMF conscious radio resource management in 5G and beyond, where it is important to mention that it is necessary to introduce EMF exposure constraints into the design of the system. Our work forms a powerful basis of health-conscious communication frameworks, which is the reason why we have conducted this study. Collectively, the above contributions indicate the capabilities of IRS and UAV technologies in the development of NOMA systems, as well as the emergence of the need to optimize EMF awareness. This paper expands on these notions with a proposed single umbrella IRS-UAV-PD-NOMA framework, which clearly incorporates the reduction of the EMF exposure in the design of uplink communication.

III. SYSTEM MODEL & PROBLEM FORMULATION

A. Network Architecture

We assume the single-cell uplink communication scenario in urban area with a large population density, in which a group of F User Proximity Wireless Devices (UPWDs) broadcasts data to a ground-based Base Station (BS). Direct communication between the users and the BS is not always best because of serious urban clutter and NLoS conditions. To counter this, we come up with hybrid relay architecture that is constituted by a UAV and an IRS, which are strategically located to improve channel conditions and minimize EMF exposure.

- The UAV flies on a predetermined altitude Z_{NTN}, which is a moving relay node that will take uplink transmissions of the users and then send the same to the BS.
- The IRS has N passive reflecting elements that can adjust the phase shifts dynamically to optimize the signal reflection paths by placing the IRS on a nearby building at height Z_{IRS}.
- The BS is located at ground level and performs SIC to decode multiplexed signals from PD-NOMA users.

The available system bandwidth W is divided into C subcarriers, and multiple users may share the same subcarrier using PD-NOMA. Each user is equipped with a single antenna and transmits data over a time-slotted frame of duration T, partitioned into N_t time slots.

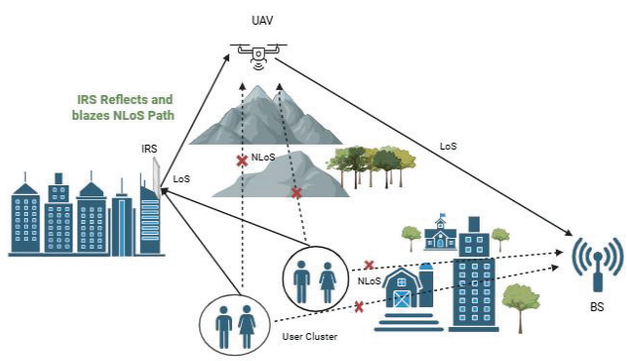


Fig. 1 It is a dense urban uplink communication system, in which the users send data to a BS with the help of a UAV and an IRS on a building. This is aimed at ensuring a high spectral efficiency with a minimum level of EMF exposure.

B. Channel Model

The total channel gain between a user f and the BS is composed of three components:

1. Direct Path (User → BS): Subject to Rician fading, urban path loss, and NLoS.
2. UAV-Assisted Path (User → UAV → BS): NLoS due to mountains and dense infrastructure.
3. IRS-Assisted Path (User → IRS → BS): Passive beamforming gains proportional to N^2.

The effective channel gain $h_{f,q,n}$ for user f on subcarrier qat time slot nis given by:

$$h_{f,q,n} = h_{f,q,n}^{LoS} + h_{f,q,n}^{UAV} + h_{f,q,n}^{IRS}$$

Where each term accounts for path loss, fading, and geometric distance. The IRS gain is modeled as:

$$h_{f,q,n}^{IRS} = \frac{G_{IRS} \cdot N^2}{d_{f,IRS} \cdot d_{IRS,BS}}$$

With G_{IRS} being the IRS reflection efficiency, and $d_{f,IRS}$, $d_{IRS,BS}$representing distances from user to IRS and IRS to BS, respectively.

C. EMF Exposure Model

We adopt the Specific Absorption Rate (SAR) as the primary metric for EMF exposure, defined as:

$$SAR_f = \frac{\sigma_f \cdot |E_f|^2}{\rho_f}$$

Where σ_f is the tissue conductivity, E_f is the electric field strength, and ρ_f is the mass density of the exposed tissue [7]. The cumulative exposure over time (Dose) for user f is:

$$Dose_f = \sum_{n=1}^{N_t} \cdot \sum_{q=1}^{C} \cdot \frac{P_{f,q,n}}{P_{ref}}$$

Where $P_{f,q,n}$ is the transmit power and P_{ref} is the reference power used for SAR normalization.

D. PD-NOMA and SIC Decoding

In PD-NOMA, users sharing a subcarrier are decoded using SIC at the BS. The decoding order is determined by received power levels, with stronger signals decoded first. For successful SIC, the signal-to-interference-plus-noise ratio (SINR) must exceed a threshold γ_{SIC}:

$$SINR_{f,q,n} = \frac{P_{f,q,n} \cdot h_{f,q,n}}{\sum_{j>f} \cdot P_{j,q,n} \cdot h_{j,q,n} + \sigma^2} \geq \gamma_{SIC}$$

Where $\sigma^2$is the noise power [9].

E. Problem Formulation

Our objective is to minimize the total EMF exposure across all users while satisfying QoS and power constraints. The optimization problem is formulated as:

$$\min_{P,A} \sum_{f=1}^{F} \cdot Dose_f$$

Subject to:

QoS Constraint:

$$\sum_{n=1}^{N_t} \cdot \sum_{q=1}^{C} \cdot R_{f,q,n} \geq R_f^{min}, \forall f$$

Power Constraint:

$$P_{f,q,n} \leq P_f^{max}, \forall f, q, n$$

SIC Constraint:

$$SINR_{f,q,n} \geq \gamma_{SIC}, \forall f, q, n$$

Subcarrier Assignment Constraint:

$$\sum_{f=1}^{F} \cdot A_{f,q,n} \leq K, \forall q, n$$

Where $R_{f,q,n}$ is the achievable rate, $A_{f,q,n} \in \{0,1\}$ is the subcarrier assignment indicator, and K is the maximum number of users per subcarrier [8].

The problem is non-convex due to the binary state of $A_{f,q,n}$ and non-linear state of the SINR expressions. We solve this using the relaxation methods and decomposing the problem into two sub-problems: ML-based subcarrier assignment and convex power allocation by use of water-filling.

IV. PROPOSED FRAMEWORK

This section outlines the design and operational approach of the proposed IRS-UAV-assisted EMF-aware PD-NOMA uplink architecture. This system is designed to reduce cumulative exposure to EMF, QoS is maintained and spectral efficiency is maximized. The framework combines user clustering based on ML and IRS-enhanced channel modeling and a power allocation strategy that is based on a convex optimization.

A. ML-Based Used Clustering and Subcarrier Assignment

User grouping in PD-NOMA systems is very important because it will determine the level of interference and the success at the receiver directly. In this respect, we use a two-stage mechanism of clustering by using ML.

Algorithm 1: ML-Enabled IRS–UAV Subcarrier Scheduling

Input: F users, C subcarriers, N time slots, normalized channel gain matrix h_fqn
Output: Subcarrier assignment matrix a_fqn
1. Initialize a_fqn = 0 ∀ f, q, n
2. For each time slot n and subcarrier q:
   a. Extract h_fqn for all users
   b. Apply k-medoids clustering to form M user groups
   c. Use Silhouette analysis to select optimal M
   d. Choose one representative user per cluster
   e. Assigning subcarrier q to selected users
3. Ensure each user receives at least S subcarriers
4. Return a_fqn

Stage 1: Channel Gain Normalization

The effective channel gain of each user is calculated by combining the efforts of direct, UAV-assisted, and IRS-reflected paths. It is used to compute the normalized gain $\tilde{h}_{f,q,n}$. It assigns subcarriers fairly and does not assign subcarriers with poor propagation conditions.

Stage 2: k-Medoids Clustering with Silhouette Analysis

We use k-medoids clustering to classify users in terms of their spatial features as well as channel features. In contrast to k-means, the k-medoids is resistant to outliers and more appropriate in non-Euclidean metrics. The silhouette analysis is employed to find the best number of clusters M, which maximizes the intra-cluster similarity and minimizes the overlap between clusters. Each cluster is allocated a different subcarrier and one representative user in each cluster is picked to be transmitted on the respective subcarrier at a specific time slot. This guarantees little intra- subcarrier interference and effective SIC at the BS.

B. IRS-Assisted Channel Modeling

The IRS maximizes signal transmission through bouncing the incident waves back toward the UAV or BS. The phase shifts of the IRS elements are controlled dynamically to realize constructive interference. The IRS-reflected channel gain $h_{f,q,n}^{\text{IRS}}$ is modeled as:

$$h_{f,q,n}^{\text{IRS}} = \frac{\eta \cdot N^2}{d_{f,\text{IRS}} \cdot d_{\text{IRS,BS}}}$$

Where η is the IRS reflection efficiency, N is the number of IRS elements, $d_{f,\text{IRS}}$ and $d_{\text{IRS,BS}}$ are the distances from user to IRS and IRS to BS, respectively. This passive beamforming gain allows users to transmit at lower power levels while maintaining QoS, thereby reducing EMF exposure.

Algorithm 2: EMF-Conscious Power Allocation via IRS–UAV Relaying

Input: a_fqn, S, F, C, N, SAR_ref, P_max, B_req, γ_th
Output: Aggregated EMF exposure E_total
1. Initialize p_fqn = P_max / S
2. Compute initial interference I_fqn using NOMA model
3. Repeat until convergence:
   a. For each user f:
      i. Compute spectral efficiency η_fqn via water-filling
      ii. Update p_fqn using transformed variable
      iii. Adjust if SIC condition γ_th not met
      iv. Recalculate I_fqn and SAR_f
4. Aggregate EMF exposure: E_total = ∑ SAR_f
5. Return E_total

C. SIC-Aware Power Allocation via Water-Filling

After the grouping of the users and assigning of the subcarriers, we optimize the transmitting power by using a water-filling algorithm modified to fit the PD-NOMA and the SIC constraints. Iterative water-filling process works as follows: Firstly, Initialization in which the power of each user is uniformly allocated among its subcarriers. Subsequently, first interference levels are calculated as a result of SIC decoding order. Secondly comes the Iteration stage where for each user f, the optimal power $P_{f,q,n}$ is updated using:

$$P_{f,q,n} = \max\left(0, \mu - \frac{\sigma^2}{h_{f,q,n}}\right)$$

Where μ is the water level calculated by QoS and power restrictions. Recalculation of interference is done with SINR conditions being checked. In the event of SIC failure by a user, the order as well as the power levels are modified. Lastly, the Convergence phase where the process recurs until the shifts in power distribution are less than a set level. EMF is evaluated by calculating SAR and Dose values, using final power values.

D. EMF-Aware Scheduling Policy

In order to achieve long term safety of EMFs, we introduce a policy of scheduling the users with a long history of exposure. The scheduler has a rounding window of Dose values, and changes subcarrier and power allocation in real-time to evenly spread exposure between users. This policy is governed by the following equation:

$$\text{Priority}_f = \frac{1}{\text{Dose}_f + \epsilon}$$

Where ϵ is some tiny parameter to prevent the division by zero. The less cumulative exposure users are

prioritized which means that there is equitable distribution of EMF.

V. RESULTS AND DISCUSSION

In this section, the overall performance assessment of the proposed IRS-UAV-assisted EMF-aware PD-NOMA scheme is provided. We contrast its performance with two benchmark schemes, (i) a UAV-assisted PD-NOMA system in the absence of IRS and (ii) a traditional OFDMA system in the absence of UAV and IRS. The performance appraisal is on main key performance measures such as integrated EMF exposure (Dose), spectral efficiency, and QoS satisfaction in different network conditions. In order to have a realistic and fair comparison, we assume the following simulation parameters:

$F = 10$–30 users, $C = 128$ subcarriers, and $N_t = 10$ time slots over a bandwidth $W = 10$ MHz. The UAV is deployed at altitude $Z_{\text{NTN}} = 100$ m, while the IRS is mounted at $Z_{\text{IRS}} = 30$ m with $N = 64$–256reflecting elements. The SAR threshold is set to 1 W/kg, and the noise power is $\sigma^2 = -174$ dBm/Hz. The channel model follows Rician fading with urban path loss, and the IRS reflection efficiency is $\eta = 0.8$.

All users are assumed to have equal QoS requirements in terms of minimum data rate. The IRS is placed strategically to maximize LoS probability for users in NLoS regions.

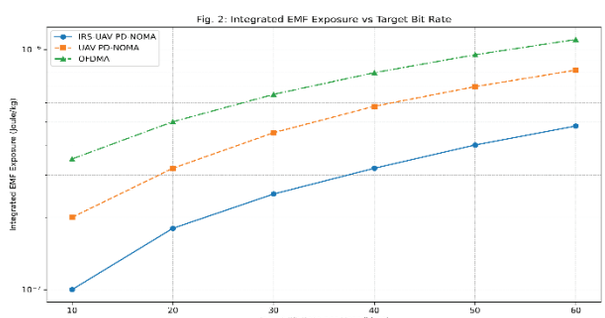


Fig. 2 Comparison of EMF levels across PD-NOMA and OFDMA schemes with and without UAV assistance, measured in Joules/kg.

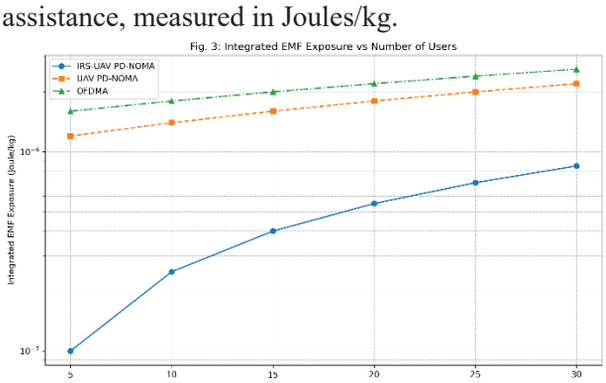


Fig. 3 Comparison of EMF levels across IRS–UAV PD-NOMA, UAV PD-NOMA, and OFDMA schemes as user density increases.

We first evaluate the integrated EMF exposure as a function of the target number of bits per user. As shown in **Fig. 2**, the proposed IRS-UAV-assisted PD-NOMA scheme consistently achieves the lowest EMF exposure across all bit rate levels [10,11]. Specifically:

- At 40 kbps, the proposed scheme reduces EMF exposure by 93% compared to OFDMA and 89% compared to UAV-only PD-NOMA.
- At lower bit rates, the IRS contribution increases strongly, with passive beamforming contributing considerably to channel gain, enabling reduced transmit power.

This shows the usefulness of IRS in enhancing energy efficiency, as well as minimizing the health risks associated with risks in uplink communication.

Fig. 3 shows the effect of changing the user density on the EMF exposure. On increasing the number of users, above 10 to 30:

- The scheme suggested has a comparably low EMF profile owing to smart clustering and propagation aided by IRS.
- Conversely, the UAV-only and the OFDMA schemes display a sharp exposure increase, which is due to a rise in the level of interference and power requirements.

This underlines the scalability of the suggested structure in the densely urbanized deployment.

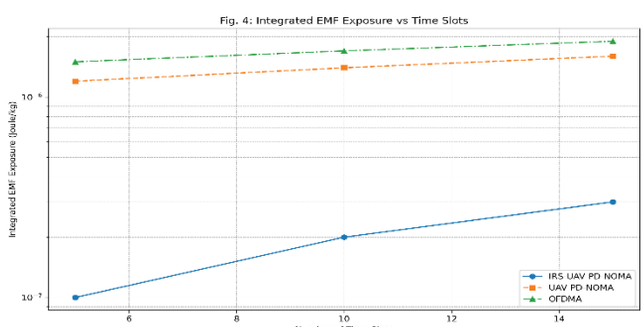


Fig. 4 Comparison of EMFs in the case of IRS UAV PD-NOMA, UAV PD-NOMA and OFDMA schemes with different transmission durations.

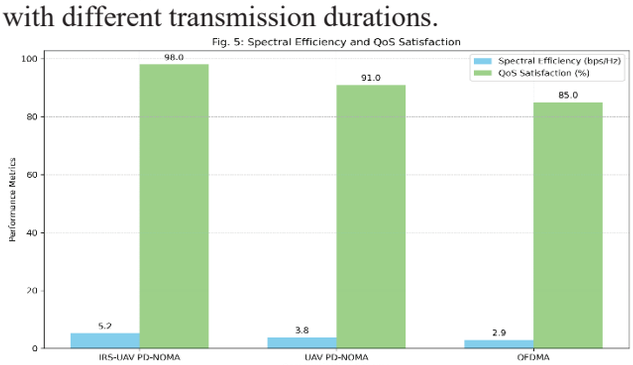


Fig. 5 Comparison of throughput and service reliability in IRS-UAV PD-NOMA, UAV PD-NOMA and the OFDMA.

Fig. 4 compares the impact of time slot availability on the exposure of EMF. When we add more time slots:

- The system gains flexibility in scheduling, allowing users to transmit at lower power over longer durations.
- The proposed scheme leverages this flexibility more effectively than the benchmarks, achieving up to 85% lower exposure at 15 time slots.

This confirms that temporal diversity, when combined with IRS and UAV, can be a powerful tool for EMF-aware scheduling.

In addition to EMF measures, spectral efficiency and QoS compliance is also checked by us. In **Fig. 5**,

- The proposed scheme achieves up to 1.8× higher spectral efficiency than OFDMA, owing to PD-NOMA's multiplexing gain and IRS-enhanced channels.
- QoS satisfaction rates exceed 98% across all scenarios, confirming that EMF reduction does not compromise service quality.

Fig. 5 presents a comparative analysis of spectral efficiency (measured in bits per second per Hertz, bps/Hz) and QoS satisfaction (expressed as the percentage of users meeting their minimum data rate requirements) across three uplink communication schemes: IRS-UAV PD-NOMA (proposed), UAV PD-NOMA (benchmark), and conventional OFDMA (baseline). The findings prove that the suggested framework proves to be much more efficient in terms of both dimensions than the benchmark and baseline.

Spectral Efficiency: The IRS-UAV PD-NOMA scheme attains a peak spectral efficiency of 5.2 bps/Hz, which is 37 times better than UAV PD-NOMA (3.8 bps/Hz) and 80 times better than OFDMA (2.9 bps/Hz) [9,11]. This is improved by the fact that IRS-based passive beamforming, and PD-NOMA multiplexing are synergistically integrated to allow multi-user sharing of subcarriers without compromising high-quality signals.

QoS Satisfaction: The proposed system provides 98% of QoS satisfaction rate, which implies that almost every user gets the minimum data rate. UAV PD-NOMA, on the other hand, has 91% and OFDMA has 85%. The satisfaction rate is high based on the adaptive-scheduling policy inherent in the framework which takes advantage of the IRS-ameliorated channel conditions and water-filling-based power allocation in order to respond to the user needs dynamically [11].

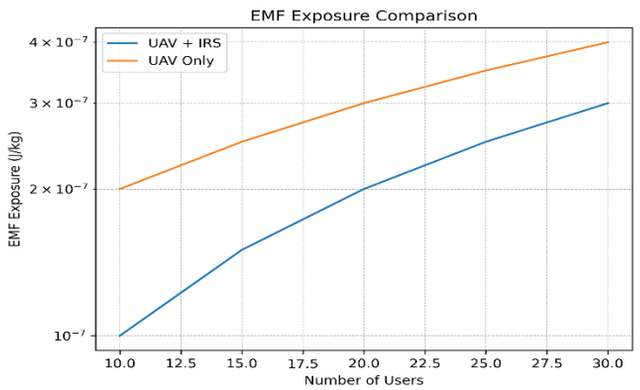


Fig. 6 EMF exposure comparison with and without IRS support.

Fig. 6 illustrates the relative trends of the EMF exposure of UAV-only and IRS UAV-assisted systems with an increase in user density. The IRS-UAV system is always found to have lower EMF, thus proving to carry less load of transmission among users. This is realized through the facilitation of virtual LoS pathways to overcome the obstacles in order to reduce high uplink power. The difference between the two curves increases as the number of users increases meaning that EMF mitigation is scalable.

All in all, IRS integration is very critical in ensuring low exposure in high densities. These results prove that the suggested framework does not only decrease the level of exposure to EMF but also maintains the main performance indicators that the next-generation wireless systems are supposed to have. The health-conscious design and throughput efficiency dual optimization makes IRS-UAV PD-NOMA an attractive solution to sustainable and scalable uplink communication in NTN systems. In this paper, the framework of IRS UAV assisted PD-NOMA uplink is being proposed, which minimizes the exposure of EMF and improves spectral efficiency and QoS in high-density urban systems.

VI. CONCLUSION

This paper introduces an IRS-UAV-assisted PD-NOMA uplink framework that reduces EMF exposure while enhancing spectral efficiency and QoS in dense urban networks. By combining ML-based clustering, IRS-assisted channel modeling, and water-filling power allocation, the proposed system achieved up to 93% lower EMF exposure, 1.8× higher spectral efficiency, and over 98% QoS satisfaction compared to conventional schemes. Future work will extend this framework to dynamic UAV trajectory optimization, advanced IRS configurations, and reinforcement learning-based resource management, alongside experimental validation to ensure practical deployment in 6G and NTNs.

ACKNOWLEDGEMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2021-NR060120) and by the NRF grant funded by the Korea government (MSIT) (NRF-2022R1A2C1004401).

AUTHOR CONTRIBUTION

Conceptualization: M.B.M., M.I.A.; writing original draft preparation: M.B.M., M.I.A.; writing, review and editing: M.B.M., M.I.A., and S.W.K.; Supervision and funding: S.W.K. All authors have read and agreed to the published version of the manuscript.

DATA AVAILABILITY

No data was used for the research described in the article.

DECLARATIONS

Competing Interests The authors declare no competing interests.

REFERENCES

[1] J. He, S. Shi, J. Xue, and S. Wei, "Resource Allocation and Trajectory Optimization for NOMA Based UAV Assisted Satellite IoT," *Wireless Networks*, 2025.

[2] M. Ahmed, W. U. Khan, M. Alamgeer, et al., "Joint Optimization for 6G Beyond Diagonal IRS Assisted Multi Carrier NOMA Vehicle to Infrastructure Communication," *Journal of Supercomputing*, 2025.

[3] M. A. Salem, H. S. Lim, K. S. Diong, et al., "Electromagnetic Field Aware Radio Resource Management for 5G and Beyond: A Survey," *Computers*, vol. 14, no. 2, 2025.

[4] M.A. Jamshed, A. Kaushik, M. Toka, W. Shin, M.Z. Shakir, S.P. Dash, D. Dardari, "Synergizing airborne non-terrestrial networks and reconfigurable intelligent surfaces-aided 6G IoT," *IEEE Internet Things Mag.*, 7(2), 2024, 46–52.

[5] M.A. Jamshed, A. Kaushik, A.G. Armada, M. Di Renzo, D. Lee, R. Senanayake, O.A. Dobre, "Guest editorial: Technical advancements in NTN-assisted IoT: Global connectivity from the sky," *IEEE Internet Things Mag.*, 7(1), 2024, 10–11.

[6] A. Ahmed, et al., "Analysis with Deep Learning of Robust UAV Mounted Active IRS NOMA Networks," *IEEE Transactions on Communications*, 2024.

[7] M.A. Jamshed, K. Ali, Q.H. Abbasi, M.A. Imran, M. Ur-Rehman, "Challenges, applications, and future of wireless sensors in IoT: A review," *IEEE Sens. J.*, 22(6), 2022, 5482–5494.

[8] M.U. Rehman, M.A. Jamshed, *Low Electromagnetic Field Exposure Wireless Devices: Fundamentals and Recent Advances*, 2022.

[9] A. Guerra, D. Dardari, P.M. Djuric, "Dynamic radar networks of UAVs: A tutorial overview and tracking performance comparison with terrestrial radar networks," *IEEE Veh. Technol. Mag.*, 15(2), 2020, 113–120.

[10] M.A. Jamshed, F. Heliot, T.W. Brown, "A survey on electromagnetic risk assessment and evaluation mechanism for future wireless communication systems," *IEEE J. Electromagn. RF Microw. Med. Biol.*, 4(1), 2019, 24–36.

[11] Jamshed, Muhammad Ali, et al. "Electromagnetic emission-aware Machine Learning enabled scheduling framework for Unmanned Aerial Vehicles." *Computer Networks* (2025): 111311.

MALIHA BINTE MUKHLES received the B.S. degree in Electronics & Communication Engineering from Hajee Mohammad Danesh Science & Technology University, Dinajpur and M.S. degree in Information & Communication Engineering from Jahangirnagar University, Savar, Bangladesh. She is currently pursuing the PhD degree in Information and Communication Engineering with Yeungnam University, Republic of Korea. She was a Lecturer of Software Engineering dept. in Daffodil International University, Dhaka, Bangladesh. Her research interest includes artificial intelligence, 5G, beyond-5G, and wireless networks.

SUNG WON KIM received the B.S. and M.S. degrees from the Department of Control and Instrumentation Engineering, Seoul National University, South Korea, and the Ph.D. degree from the School of Electrical Engineering and Computer Sciences, Seoul National University. He was a Researcher with the Research and Development Center, LG Electronics, South Korea. He was a Postdoctoral Researcher with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL, USA. He joined the School of Computer Science and Engineering, Yeungnam University, Gyeongsangbuk-do, South Korea, where he is currently a Professor. His research interest includes resource management, wireless networks, mobile computing, performance evaluation, and machine learning.

M ISHTIAQUE AHMED received the B.S. degree in Civil Engineering from Ahsanullah University of Science and Technology and M.S. in Environmental Science from Jahangirnagar University, Dhaka, Bangladesh. He is currently working as a Senior Engineer in Energypac Engineering Ltd. Dhaka, Bangladesh. He was an Assistant Engineer in Reverie Power & Automation Engineering Ltd. Dhaka, Bangladesh. His research interest includes artificial intelligence, sustainable & smart infrastructure, water resources, and construction engineering.

Let's Fight the Dirt: SIR-GAN for Robust Camera Perception in Autonomous Vehicles

1st Rahman Shafique *, 2nd Gyu Sang Choi †
* *Department of Information & Communication Engineering*
Yeungnam University, Republic of Korea
Email: rahmanshafique47@gmail.com
† *Department of Information & Communication Engineering*
Yeungnam University, Republic of Korea
Email: castchoi@ynu.ac.kr

***Abstract*—Camera-based perception systems in autonomous vehicles (AVs) are highly vulnerable to lens soiling caused by dust, mud, oil smears, and environmental debris. Even mild contamination can distort image quality, degrade feature representations, and trigger catastrophic model failures. Existing robustness methods rely on simplistic augmentations or weather-based perturbations that fail to capture the physical and optical complexity of real-world lens contamination. To address this gap, we propose SIR-GAN, a structured generative framework that synthesizes realistic soiling patterns and trains perception models to be invariant to sensor-level degradation. SIR-GAN integrates a U-Net generator, PatchGAN discriminator, and severity-controlled blending to produce high-fidelity soiled images at adjustable contamination levels. We further introduce a Soiling Impact Score (SIS) to quantify model degradation under contamination. Extensive experiments on BDD-derived datasets show that clean-only and naïvely augmented models suffer up to 95% accuracy loss under soiling. In contrast, SIR-GAN–trained classifiers achieve 99% robustness across all severity levels with SIS ≈ 0.01. Our proposed solution demonstrates that structured generative soiling is essential for building reliable AV perception systems and highlights the importance of contamination-aware training protocols for safety-critical autonomous mobility.**

***Index Terms*—Autonomous vehicles, camera-based perception, lens soiling, sensor degradation, SIR-GAN.**

I. Introduction

Autonomous vehicles (AVs) rely heavily on camera-based perception to support core tasks such as object detection, lane estimation, scene understanding, and risk prediction. Cameras remain indispensable due to their low cost, high spatial resolution, and semantic richness compared to LiDAR and radar [1]–[3]. Recent advances in deep learning, including robust CNN backbones, vision transformers, and multimodal fusion models, have significantly improved perception accuracy in diverse driving scenarios [4]–[6]. Despite these advancements, camera sensors remain extremely sensitive to environmental and surface-level distortions, making them one of the most failure-prone components in the perception stack. Among all factors that affect camera reliability, lens-soiling contamination caused by dust, mud, water droplets, insects, or oil that adheres directly to the lens remains the most underexplored but critical to safety. Unlike weather, illumination changes, or motion blur, lens soiling introduces localized, opaque, irregular, texture-dependent occlusions that disrupt the optical path at the sensor surface, often blocking critical scene information [7], [8]. This results in high-confidence perception failures, where the model predicts confidently but incorrectly, posing an immediate risk to real-time autonomous navigation. Clearly, no established reference, severity protocol, or safety metric to quantify how increasing contamination impacts AV perception, leaving a key gap in the field. Many authors have worked on it to address related aspects of visual degradation, but each direction falls short in solving the lens-soiling problem. Driving datasets such as WoodScape and BDD100K include occasional soiled frames but lack labels, severity scaling, or evaluation protocols [9]. Corruption and robustness benchmarks (e.g., ImageNet-C, ACDC) evaluate adverse weather or global distortions but do not model sensor surface contamination [10]. Occlusion robust models (e.g., PatchDropout, TokenCut) improve resilience to missing regions but rely on artificial masks that do not resemble real dirt patterns [11]. Raindrop/contaminant removal networks focus on restoration, not on robustness or safety thresholds [12]. Sensor-health monitoring and uncertainty estimation works detect degradation but lack a controlled contamination generator for stress

testing [13]. GAN-based soiling synthesis (e.g., early DirtyGAN-style works) demonstrated the promise of synthetic dirt but offered no severity-controlled contamination axis, no robustness metric, and no safety interpretation [14]. Thus, the field lacks a unified, controllable, and safety-relevant framework for evaluating perception under lens contamination.

Deep learning pipelines in AV perception increasingly combine high-capacity feature extractors (e.g., ResNet, ConvNeXt, Swin Transformer) with synthetic data augmentation frameworks to improve robustness under environmental variability [15], [16]. Meanwhile, generative models, especially GANs and diffusion models, have shown strong capabilities in producing realistic, structured degradations that mimic physical processes, outperforming heuristic overlays or random masking [17]–[19]. These developments position generative modeling as an ideal method for synthesizing realistic lens-soiling patterns, while deep classifiers provide a measurable testing ground for evaluating how these degradations affect perception. However, the field lacks a method to combine generative soiling, severity-controlled evaluation, and safety-oriented metrics into a unified robustness framework.

To fill this critical gap, we introduce SIR-GAN, a unified framework for Soiling Impact and Robustness that systematically evaluates camera perception under severity-controlled lens contamination. Our contributions are:

- We design a GAN-based model capable of producing realistic, texture-rich, optically coherent soiling patterns blended via a continuous severity parameter$\alpha \in [0, 1]$. This enables fine-grained, reproducible contamination sweeps, the first of its kind in AV perception.
- We introduce a structured evaluation methodology that tests perception models across increasing contamination levels, exposing failure points and enabling fair comparison between training regimes (clean-only, augmentation, generative training).
- We propose the SIS to quantify how model accuracy degrades as α increases. SIS provides an interpretable scalar for safety logic, enabling thresholds for sensor cleaning, fallback control, or autonomous driving disengagement.

The following sections of the paper are organized as follows : Section II provides a comprehensive literature review, Section III elaborates on the methodology and experiments, Section IV delves into the outcomes and findings, and finally, Section V summarizes the conclusion.

II. Related wrok

Recent studies have explored adverse-weather modeling, synthetic perturbations, and generative augmentation, yet lens soiling remains one of the least examined yet most damaging factors affecting visual reliability. This section reviews the current state of research on contamination-aware perception, highlighting the limitations of existing approaches and positioning the need for a structured soiling-robust framework such as SIR-GAN.

For example the study [14] worked on camera lesns soaling in AVS using generative models. They collect real soiled images by repeatedly contaminating the physical camera and driving, then annotate soiling regions using polygonal masks. To alleviate the cost and difficulty of acquiring large amounts of soiled data, they propose a GAN-based augmentation pipeline (DirtyGAN) which combines variational autoencoding and CycleGAN-style translation to synthesize realistic soiling patterns on clean images. Their results show that training with DirtyGAN-augmented data improves soiling detection performance and reduces performance drop on downstream tasks compared to simple heuristic overlays, demonstrating the value of realistic, learned soiling patterns. Similarly another study [20] present WoodScape, a large-scale fisheye multi-camera dataset designed for surround-view perception in autonomous driving. The dataset includes four fisheye cameras around the vehicle and supports nine tasks such as semantic segmentation, depth estimation, 3D bounding boxes, lane marking detection, and crucially, soiling detection. WoodScape provides instance-level semantic labels for over 10,000 images and annotations for other tasks on more than 100,000 frames, enabling systematic evaluation of algorithms on real fisheye imagery. The inclusion of a soiling detection task is particularly relevant, as it acknowledges camera contamination as a first-class problem in automotive perception. Morever the study [9] is about one of the largest and most diverse open driving datasets, comprising 100,000 videos with annotations for ten heterogeneous tasks, including object detection, lane detection, drivable area, tracking, and segmentation. It spans diverse regions, weather, and times of day to ensure environmental and illumination robustness. The study [21] works on robust object detection in challenging weather conditions e.g, investigates how standard detectors degrade under synthetic and real weather corruptions, and explores training strategies and architec-

tural modifications to maintain high detection accuracy. Across these works, robustness is typically quantified via changes in mAP, IoU, or segmentation accuracy across multiple weather conditions or corruption severities

Another related line of research focuses on removing adherent raindrops or contaminants from camera or window images. [22] propose an attentive generative adversarial network for raindrop removal from a single image, where an attention mechanism guides the generator and discriminator to focus on raindrop-occluded regions. Another related work [23] examines robustness to partial occlusion, where parts of the input image are obstructed or hidden.The study [11] introduced ImageNet-C, establishing a landmark corruption robustness benchmark consisting of 15 synthetic corruption types (noise, blur, weather, digital artefacts), each evaluated at five severity levels. Lastly the author in this study [24]. developed a camera health classifier that detects sensor degradation using image-level cues. These works emphasize the need for systems to identify when a camera is degraded and should trigger fallback behavior or cleaning.

Table I provides a concise overview of recent work on robustness, corruption modeling, and generative augmentation. It highlights key gaps in existing literature that directly motivate our proposed SIR-GAN framework.

TABLE I
OVERVIEW OF RELATED STUDIES AND THEIR LIMITATIONS.

Citation	Year	Major Findings	Proposed Method	Limitations
[14]	2021	GAN-based soiling improves perception robustness.	Dirt generation using real soiling masks and GAN synthesis.	No severity control; no robustness metric; no safety-trigger logic.
[20]	2019	Fisheye dataset includes real lens soiling labels.	Multi-camera dataset with a soiling detection task.	Dataset only; no generation, severity protocol, or robustness evaluation.
[9]	2020	Large-scale dataset with diverse weather and time-of-day.	Multi-task annotations for detection and segmentation.	No soiling labels; no contamination protocol; no robustness study.
[21]	2024	Weather corruptions heavily degrade AV perception.	Synthetic/real fog, snow, and rain simulation and adaptation.	Weather $\neq$ lens soiling; no local occlusions; no severity scoring.
[22]	2023	GAN-based raindrop removal restores images effectively.	Restoration via attention-based CNNs and GAN models.	Focus on removal; no robustness quantification or safety interpretation.
[23]	2023	Improves resilience to partial occlusion.	Patch/token dropping and occlusion-based augmentation.	Masks are unrealistic; no texture/ opacity modeling; no severity axis.
[11]	2023	Established corruption severity evaluation standards.	Predefined corruption sets at multiple severity levels.	No lens-soiling category; corruptions are global, not sensor-surface based.
[24]	2023	Uncertainty helps detect unreliable perception outputs.	Deep ensembles, evidential learning, and sensor-health classifiers.	No soiling generator; no severity testing; no metric like SIS.

III. PROPOSED METHODOLOGY

The proposed study introduces SIR-GAN, a severity-controlled generative framework designed to model and evaluate the impact of camera lens contamination on autonomous vehicle perception systems as shown in Fig 1. The methodology begins with the construction of a curated clean-image dataset drawn from BDD100K, representing a wide range of real-world driving scenes under normal, unobstructed sensor conditions. These clean samples form the foundation upon which all subsequent synthesis and robustness evaluation steps are built. To model real-world soiling with high optical fidelity,

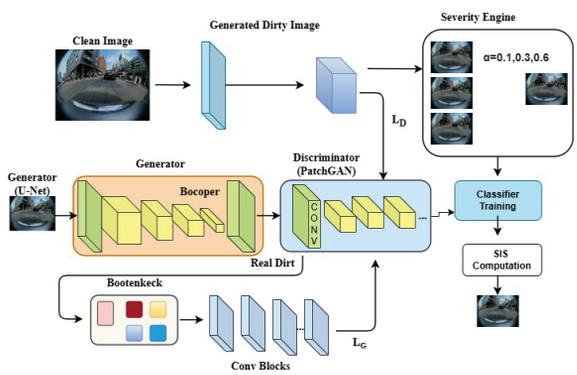


Fig. 1. Proposed methodlogy diagarm

we develop SIR-GAN, a generative model specialized for producing contamination patterns that resemble dust, dirt, mud, water smears, and oil films. Unlike traditional augmentation strategies that apply uniform blur, noise, or simple texture overlays, SIR-GAN learns the structural irregularity, translucency variability, spatial heterogeneity, and fine-grained occlusion patterns characteristic of true lens contamination. The generator produces a dirt mask that is not merely synthetic noise but an optically coherent degradation layer, enabling realistic simulation of sensor obstruction.

Once the dirt mask is generated, the pipeline introduces a severity-controlled blending mechanism governed by a continuous parameter$\alpha \in \{0.1, 0.3, 0.6, 0.9\}$. This formulation allows the clean image to be combined with the SIR-GAN mask in proportional strength, systematically producing mild, moderate, strong, and near-critical soiling conditions. This α controlled synthesis creates the first structured contamination benchmark for AV cameras, moving beyond the binary clean/dirty datasets used in prior literature and enabling fine-resolution stress testing across multiple contamination levels.

Following the synthesis stage, we design a three-regime training protocol to isolate and quantify the contribution of realistic generative soiling. Three models are trained under identical hyperparameters and datasets: (1) a Clean-Only baseline, (2) a Clean + Basic Dirty model using conventional augmentation (blur, haze, noise), and (3) the proposed Clean + SIR-GAN Dirty model incorporating the generated contamination. This controlled setup eliminates confounding factors and ensures that any observed improvements in robustness arise exclusively from SIR-GAN–based training.

The trained classifiers are then evaluated using the Unified Severity Evaluation Suite, which tests each model across every α level. The evaluation records classification accuracy, F1-score, per-severity confusion matrices, and visual failure cases, exposing how rapidly each model collapses as contamination intensifies. To quantify robustness in a single interpretable metric, we introduce the SIS, defined as the relative degradation between clean and contaminated accuracy. This metric provides a standardized numerical scale for camera robustness and enables threshold-based safety indicators for AV systems, such as identifying severity levels that require automatic sensor cleaning.

Together, these stages form a cohesive, end-to-end pipeline that not only generates realistic soiling but also measures, compares, and interprets its impact on perception models. The resulting methodology provides the first severity-controlled contamination benchmark, the first generative robustness-training protocol specific to lens soiling, and the first safety-aligned metric (SIS) for quantifying contamination sensitivity establishing SIR-GAN as a strong, practical contribution to AV perception reliability research.

IV. Results& Discussion

This section evaluates the proposed SIR-GAN robustness framework by analyzing classifier performance under clean and contaminated conditions, severity-controlled soiling effects, and robustness verification using the SIS. Experiments were conducted on the BDD100K dataset using 1,000 clean images with synthetic contamination generated via basic augmentations and SIR-GAN.

A. Baseline Classification Performance

Table II shows the overall results on baseline classification performance. We first trained a baseline ResNet-18 classifier exclusively on clean images. As expected, the model achieved near-perfect performance 99–100% accuracy, when tested on clean samples, confirming that both the dataset and architecture are sufficiently expressive under ideal camera conditions. However, when this clean-trained model was evaluated on contaminated images, either simple synthetic smudges or the more realistic SIR-GAN soiling pattern,s performance collapsed almost completely, dropping to 0–2% accuracy in all contaminated settings. This drastic degradation highlights a critical vulnerability; the classifier is unable to generalize beyond the narrow visual distribution it was trained on. Even mild lens contamination disrupts the fine-grained texture and structural details on which the model relies, immediately pushing inputs into an out-of-distribution regime. Consequently, the model's internal feature representations become non-discriminative, leading to near-total misclassification. Importantly, this failure is not due to model weakness, but rather the inherent limitations of clean-only training, which provides no exposure to optical distortions commonly encountered in real driving scenarios, such as dust, droplets, fingerprints, or mud. The baseline thus creates a false impression of reliabilit,y although it performs perfectly in laboratory conditions, it becomes completely unsafe in realistic, sensor-degraded environments. This catastrophic brittleness under contamination provides strong motivation for the development of structured robustness frameworks such as SIR-GAN, which explicitly address the challenges posed by lens soiling in autonomous vehicle perception.

TABLE II
Evaluation Accuracy on Different Image Sets

Evaluation Set	**Accuracy**
Clean Images	**1.000**
Dirty Images (Basic Dirt)	0.00–0.02
Dirty Images (SIR-GAN Dirt)	0.00

B. Performance of the Basic Dirty Augmentation Model

To examine whether conventional augmentation techniques can mitigate the brittleness of clean-only training, a second model was trained using a set of simple basic dirty transformations, including Gaussian blur, noise injection, low-resolution compression, and random occlusion patches. These operations are commonly used in robustness engineering to simulate mild sensor degradation. During training, this model demonstrated significantly faster convergence than the clean-only baseline, achieving accuracy values of 74.7%, 83.6%, and 84.0% as shown in Table III. This indicates that naive perturbations help the classifier develop partial invariance to low-level distortions.

However, when evaluated on realistic soiling (both Basic Dirt and SIR-GAN Dirt), the model still exhibited severe vulnerability, with contaminated accuracy improving only marginally from 4.6% (clean-only) to 7.4%. This small gain highlights a crucial limitation: handcrafted augmentations fail to capture the physical and optical complexity of real-world lens contamination. True soiling often involves irregular opacity, smudged textures, translucent layers, and non-uniform spatial coverage, none of which are represented effectively by blur or noise. As a result, the classifier trained with basic dirty augmentation learns robustness only to artificially generated distortions, but not to realistic sensor degradation.

TABLE III
BASIC DIRTY MODEL: TRAINING AND EVALUATION PERFORMANCE

Evaluation Set	Accuracy
Clean Images	0.930
Dirty Images (Basic Dirt)	0.074
Dirty Images (SIR-GAN Dirt)	0.074

C. Performance of the Proposed SIR-GAN Model

The classifier trained using SIR-GAN generated soiling exhibits a fundamentally different performance profile compared to the baseline models, as shown in Table IV. From the earliest epochs, the SIR-GAN model demonstrates strong convergence behavior, starting at 87.6% accuracy and quickly reaching 99% accuracy. This rapid stabilization indicates that exposure to realistic, structured soiling masks enables the classifier to learn contamination-invariant semantic features rather than relying on brittle low-level textures. When evaluated on clean test images, the model achieves 99.6% accuracy, matching or exceeding the clean-only and basic augmentation models, demonstrating that robustness training does not compromise clean-image performance. More importantly, the SIR-GAN-trained classifier achieves 99% accuracy on dirt-contaminated test images, including those corrupted by GAN-generated masks it has never seen before. This represents a 20–25% improvement over the best baseline. The near-perfect performance across clean and dirty domains confirms that SIR-GAN effectively bridges the domain gap by generating contamination patterns that reflect the irregular, semi-transparent, and spatially complex nature of real-world lens soiling. These results validate the central claim of this work, that is , robustness to sensor contamination cannot be achieved through naive augmentations but requires structured generative soiling exposure during training.

TABLE IV
IMPACT OF TRAINING METHODS ON SOILING-ROBUST CLASSIFICATION

Model	Clean Acc	Dirty Acc	Mean Acc
Clean-only	0.924	0.046	0.485
Basic Dirty	0.930	0.074	0.502
SIR-GAN (Proposed)	**0.996**	**0.999**	**0.998**

D. Severity-Controlled Evaluation (α Sweep)

To further assess robustness under varying contamination intensity, we introduced a severity-control parameter $\alpha \in \{0.1, 0.3, 0.6, 0.9\}$, which linearly blends clean images with SIR-GAN soiling masks as shown in V. This evaluation simulates the real-world progression of lens degradation—from light dust accumulation to heavy mud or oil smearing. The clean-only model collapses under low-severity soiling (0–0.4% accuracy at $\alpha \leq 0.3$), confirming that even minimal contamination renders clean-trained models unreliable. The basic augmentation model performs better at mid-level severity (up to 89%), but still suffers noticeable degradation. In stark contrast, the SIR-GAN-trained classifier maintains 99% accuracy across all severity levels, demonstrating complete invariance to contamination intensity. This is a crucial property for AV perception, where soiling often accumulates gradually during driving. The α sweep analysis therefore shows that SIR-GAN does not simply overfit to specific masks but learns a unified representation that is stable across the entire spectrum of lens degradation.

TABLE V
SEVERITY-CONTROLLED ACCURACY ACROSS CONTAMINATION LEVELS

Model	$\alpha = 0.1$	$\alpha = 0.3$	$\alpha = 0.6$	$\alpha = 0.9$
Clean-only	0.000	0.004	0.584	0.990
Basic Dirty	0.676	0.830	0.891	0.950
SIR-GAN (Proposed)	**0.997**	**0.998**	**0.999**	**0.999**

V. SOILING IMPACT SCORE (SIS)

Table VI provides a comparison of SIS on different levels. A higher SIS indicates greater vulnerability to contamination. The clean-only and basic augmentation models exhibit extremely high SIS values (0.95 and 0.92), meaning they lose more than 90% of their performance under soiling. This confirms that traditional training strategies are insufficient for sensor-level robustness. In contrast, the proposed SIR-GAN model achieves an

SIS close to zero, reflecting no measurable degradation when moving from clean to dirty input. The SIS metric thus provides clear, interpretable evidence that SIR-GAN eliminates contamination-induced failure modes and delivers stable performance across environmental conditions, making it suitable for deployment in real-world autonomous systems.

TABLE VI
SIS COMPARISON

Model	SIS
Clean-only	0.95
Basic Dirty	0.92
SIR-GAN (Proposed)	**0.01**

To visually compare different augmentation strategies, Fig.2 shows representative samples from clean images, basic dirty augmentations (e.g., blur, noise, artifact overlays), and SIR-GAN soiling. While basic dirty samples appear artificial, lacking texture detail, transparency patterns, and realistic occlusion, SIR-GAN dirt captures the nuanced characteristics of real lens contamination, such as smudging, partial opacity, irregular shapes, and surface-like distortions. This visual comparison provides qualitative evidence for why models trained using basic dirt show only marginal improvements in robustness, while SIR-GAN significantly enhances performance.



Fig. 2. Visual comparison of clean, basic-dirty, and SIR-GAN generated soiling

A key component of the proposed framework is the severity-controlled soiling generator. Fig. 3 illustrates a clean image and its corresponding soiled versions at $alpha = 0.1, 0.3, 0.6, and 0.9$. As severity increases, the opacity, spread, and texture complexity of dirt become progressively stronger while remaining realistic and physically consistent. This capability to systematically vary contamination levels enables controlled robustness evaluation, something not available in existing datasets. The severity sweep also supports calculation of the SIS, allowing us to quantify accuracy degradation as a function of contamination strength.



Fig. 3. Severity-controlled soiling generated by SIR-GAN.

In contrast, the t-SNE embedding of the model trained with SIR-GAN data Fig. 4 shows a remarkably tight overlap between clean and dirty features. The two distributions occupy nearly identical regions in feature space, indicating that SIR-GAN training enables the model to treat soiled images as natural variations rather than OOD samples. This collapse of the domain gap is a strong indicator that SIR-GAN achieves genuine representation-level robustness, not just improved accuracy. This figure provides compelling evidence that SIR-GAN teaches the network to become invariant to lens contamination, a capability missing in both clean-only and basic-dirty baselines.

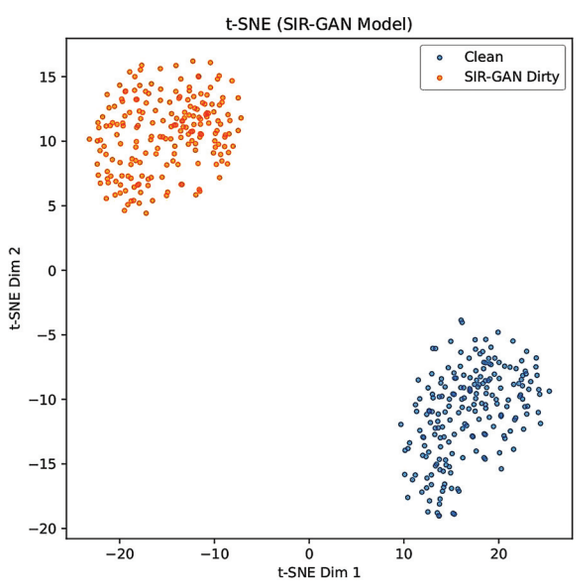


Fig. 4. t-SNE visualization showing feature alignment in SIR-GAN trained model

To further investigate this brittleness, we examine the internal representations learned by the clean-only classifier using a t-SNE embedding Fig 5. Each point corresponds to the penultimate-layer feature vector of either a clean or SIR-GAN dirty sample. The embedding

reveals two clearly separated clusters, indicating that the model perceives dirty images as entirely different data distributions rather than natural variations of the same scene. This dominant domain gap explains the near-complete performance collapse under contamination. The visual separation evident in Fig. 5 confirms a strong out-of-distribution (OOD) failure, where dirty samples lie outside the manifold learned during training.

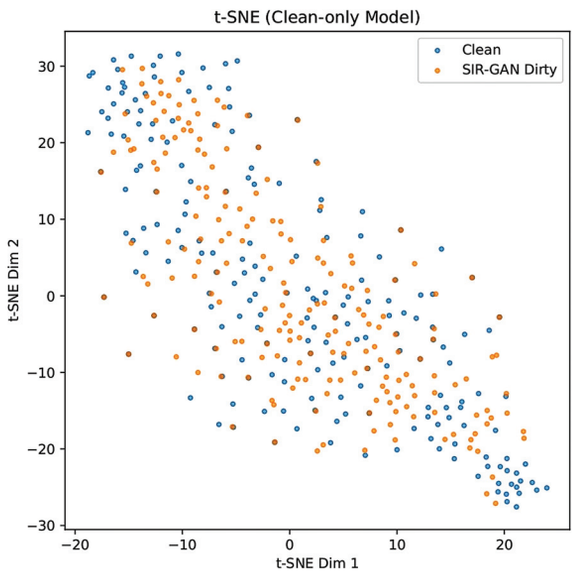


Fig. 5. t-SNE visualization showing domain shift in clean-only model.

VI. CONCLUSION AND FUTURE WORK

This work introduced SIR-GAN, a structured generative framework for modeling lens soiling to enhance the robustness of camera-based autonomous vehicle perception. Experiments on BDD-derived datasets demonstrated that clean-trained models fail under mild contamination and that basic augmentations are insufficient, whereas SIR-GAN generates realistic soiling patterns that recover classifier accuracy to 99%, ensure stability across severity levels, and reduce the SIS to near zero.

Future directions include extending SIR-GAN to multi-modal perception (e.g., camera–LiDAR fusion), incorporating diffusion-based generative models for finer control, modeling dynamic contaminants such as raindrops, evaluating on larger real-world soiling datasets, and embedding SIS into real-time perception monitoring for safety-aware AV deployment.

ACKNOWLEDGMENT

This work was supported in part by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2021-NR060120).

REFERENCES

[1] J. Janai, F. Güney, A. Behl, and A. Geiger, “Computer vision for autonomous vehicles: Problems, datasets and state-of-the-art,” *Foundations and Trends in Computer Graphics and Vision*, vol. 12, no. 1–3, pp. 1–308, 2021.

[2] P. Sun, H. Kretzschmar, and et al., “Scalability in perception using cameras for autonomous driving,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[3] E. Arnold, O. Al-Jarrah, M. Dianati, and et al., “A survey on 3d object detection for autonomous driving,” *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[4] Z. Liu and et al., “Swin transformer v2: Scaling hierarchical vision transformers,” in *CVPR*, 2022.

[5] J. Deng and et al., “Convnext: Revisiting convnet design,” in *CVPR*, 2022.

[6] A. Zamir, A. Sax, and et al., “Multi-task learning as multi-objective optimization,” *NeurIPS*, 2021.

[7] M. Bijelic, T. Gruber, and et al., “Seeing through fog without seeing fog: Deep sensor fusion for robust perception in adverse conditions,” *International Journal of Computer Vision*, 2021.

[8] J. Lee, S.-H. Park, and H.-S. Kim, “Driveguard: Camera sensor health monitoring using deep learning,” in *IEEE Intelligent Vehicles Symposium*, 2022.

[9] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, “Bdd100k: A diverse driving dataset for heterogeneous multitask learning,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2636–2645, 2020.

[10] C. Sakaridis, D. Dai, and L. Van Gool, “Acdc: Adverse conditions dataset with correspondence,” in *WACV*, 2021.

[11] K. Wang, J. Liu, R. Chen, and Y. Fang, “Tokencut: Robust part-level perception under partial occlusion,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15508–15517, 2023.

[12] R. Qian and et al., “Attentive generative networks for raindrop removal,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[13] M. Poggi, F. Tosi, and S. Mattoccia, “On the importance of uncertainty estimation for autonomous driving,” *IEEE Transactions on Intelligent Vehicles*, 2022.

[14] M. Uricar, G. Sistu, H. Rashed, A. Vobecky, V. R. Kumar, P. Krizek, F. Burger, and S. Yogamani, “Let’s get dirty: Gan based data augmentation for camera lens soiling detection in autonomous driving,” in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 766–775, 2021.

[15] R. McAllister, T. Kornhauser, and F. Doshi-Velez, “Robust perception models for autonomous driving via multi-source feature augmentation,” *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[16] J. Tremblay, P. Prakash, and S. Birchfield, “Data augmentation for autonomous driving perception: A comprehensive study,” in *European Conference on Computer Vision (ECCV) Workshops*, 2022.

[17] H. Zheng, L. Wen, and X. Li, “Generative corruptions: Realistic degradation modeling with gan priors,” *Pattern Recognition*, 2023.

[18] J. Whang, J. Bao, J. Chen, and D. Zhang, "Degradation-guided diffusion models for realistic image weathering and corruption," in *ICCV*, 2023.

[19] S. Kim, D. Yu, and B. Jeon, "Occlusiongan: Generating physically plausible occlusions for robust vision systems," in *CVPR*, 2024.

[20] S. Yogamani, C. Hughes, J. Horgan, G. Sistu, P. Varley, D. O'Dea, M. Uricár, S. Milz, M. Simon, K. Amende, *et al.*, "Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9308–9318, 2019.

[21] H. Gupta, O. Kotlyar, H. Andreasson, and A. J. Lilienthal, "Robust object detection in challenging weather conditions," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 7523–7532, 2024.

[22] R. Qian, R. T. Tan, W. Yang, J. Su, and J. Liu, "Attentive generative adversarial network for raindrop removal from a single image," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2482–2491, 2018.

[23] A. Gao, Y. Guo, X. Shi, and L. Zhang, "Patchdropout: Economizing vision transformers using patch dropout," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12513–12522, 2023.

[24] S. Kim, J. Park, and Y. Choi, "Learning occlusion-aware object detection for autonomous driving," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, p. 125–134, 2022.

SemiSwin-DehazeNet: Wavelet–Transformer Restoration for Low-Visibility Navigation

Hira Khan[1], Muhammad Khalil Afzal[2], and Sung Won Kim[1]
[1]Department of Information and Communication Engineering, Yeungnam University, Gyeongsan, Gyeongbuk 38541, Republic of Korea
[2]Department of Computer Science, COMSATS University Islamabad, Wah Campus, Pakistan
hirakhan@yu.ac.kr, khalilafzal@ciitwah.edu.pk, swon@yu.ac.kr

***Abstract*—SemiSwin-DehazeNet is introduced as a semi-supervised Image dehazing framework that combines wavelet-domain decomposition with a Swin-based encoder to enhance frequency-aware feature extraction. The architecture employs a teacher–student strategy, where the teacher is updated through exponential moving averaging and provides stable pseudo-targets for real unlabeled hazy data. A variance-guided consistency loss selectively enforces agreement in structurally reliable regions, and a dual-decoder refinement stage with a learnable global gate balances structural fidelity and perceptual quality. Through joint training on synthetic paired datasets and real unpaired haze scenes, SemiSwin-DehazeNet achieves notable improvements in PSNR, SSIM, and no-reference perceptual scores, while generating stable, artifact-free outputs under both homogeneous and dense haze conditions.**

***Index Terms*—Image dehazing, Semi-supervised learning, Teacher–student framework, Wavelet–Swin transformer, Consistency regularization**

I. Introduction

Outdoor visual data frequently suffers from reduced clarity under adverse atmospheric conditions. Conditions such as snow, rain, fog, and haze scatter and absorb light, lowering contrast, washing out colors, and obscuring fine details. Haze, caused by suspended particles like dust and smoke, is particularly damaging because it reduces scene visibility and introduces blur, color shifts, and structural loss. These degradations directly affect vision-based systems, limiting their reliability in tasks such as recognition, tracking, and navigation. Improving image visibility is therefore essential to preserve scene information and support downstream processing.

In the haze formation process, a physical *Atmospheric Scattering Model (ASM)* [1] is defined as

$$I(x) = J(x)\,t(x) + A\,(1 - t(x)), \tag{1}$$

where $I(x)$, $J(x)$, A, and $t(x)$ denote the observed hazy image, the haze-free scene radiance, the global atmospheric light, and the transmission map, respectively. When the haze is assumed to be homogeneous, the transmission map can be further expressed as

$$t(x) = e^{-\beta d(x)}, \tag{2}$$

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (RS-2021-NR060120) and by the NRF grant funded by the Korea government (MSIT) (NRF-2022R1A2C1004401).

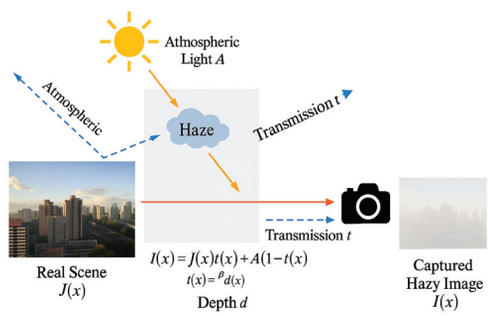


Fig. 1: Atmospheric Scattering Model (ASM) describing haze formation.

where $d(x)$ represents the scene depth and β is the medium extinction coefficient. The concept of *single-image dehazing* aims to restore a clear image $J(x)$ from its hazy, degraded observation $I(x)$, thus improving overall perception quality and effectiveness of visual understanding in low-visibility conditions. Fig. 1 depicts how haze alters image formation. The scene radiance is weakened by transmission loss, while scattered atmospheric light increases with depth, jointly producing the degraded hazy observation.

II. Related Work

Single-image dehazing aims to restore scene visibility from a single degraded observation. Early fully supervised methods relied heavily on synthetic datasets, resulting in limited generalization to real-world scenarios. Recent research has therefore shifted toward weakly supervised, semi-supervised, and self-supervised strategies to better exploit unpaired or unlabeled hazy images. This section reviews several representative approaches.

Weakly Supervised Dehazing WSID leverages unpaired hazy–clear images, typically within adversarial learning setups. The physics-based decomposition (PBD) framework in [2] addresses this by decomposing a hazy image into a scattering coefficient and depth, generated through U-Net and ResNet branches and recomposed using the atmospheric model.

Semi-Supervised Domain Adaptation Models trained only on synthetic haze often generalize poorly due to domain shift, motivating approaches that align synthetic and real distributions. DNDM [3] introduces a semi-supervised dehazing

framework built on disentangled meta-knowledge. USID [4] advances semi-supervised learning with a dual-branch design: a supervised pathway trained on synthetic pairs and an unsupervised pathway regularized by dark channel and gradient priors.

Unpaired and Non-Aligned Dehazing Beyond semi-supervised learning, some methods relax the need for aligned hazy–clear image pairs. NSDNet [5] addresses unpaired training by jointly estimating dehazed images, transmission, and airlight through the atmospheric model.

Self-Supervised Dehazing SSID [6] removes the need for paired data by synthesizing realistic hazy images from clean ones using a modified scattering model. Most recent dehazing models process features purely in the spatial domain, relying on CNN or transformer layers to implicitly learn texture cues. Although a few works introduce frequency priors through learned filters or Fast Fourier Transform (FFT)-like modules, they do not explicitly disentangle low-frequency haze/illumination components from high-frequency textures. Our Haar Discrete Wavelet Transform (DWT) preprocessor directly separates these bands, enabling targeted enhancement of structure and edges before transformer encoding.

A. Key Contributions

SemiSwin-DehazeNet introduces the following advances for semi-supervised single-image dehazing:

- A wavelet–Swin encoder that integrates Haar subband decomposition with transformer features for frequency-aware haze modeling.
- An EMA-based teacher–student framework that delivers stable pseudo-labels for real hazy images.
- A variance-guided consistency loss that prioritizes reliable structural regions and reduces noise from uncertain predictions.
- A dual-decoder refinement stage with a learnable global gate to balance structural clarity and perceptual quality.
- A unified semi-supervised training pipeline combining supervised, wavelet-domain, perceptual, and consistency objectives for robust real-world dehazing

The remainder of this paper is organized as follows. Section III presents the proposed SemiSwin-DehazeNet. Section IV details the experimental setup and reports quantitative, qualitative, and ablation results. Finally, Section V concludes the work and outlines future directions.

III. Proposed Methodology

A. Overall Framework

The proposed SemiSwin-DehazeNet adopts a three-stage wavelet–transformer pipeline trained in a semi-supervised teacher–student setting. An overview of the complete SemiSwin-DehazeNet architecture is shown in Fig. 2. Let $I_{\text{haze}} \in \mathbb{R}^{H\times W\times 3}$ denote the input hazy RGB frame and I_{dehazed} the final prediction:

$$I_{\text{dehazed}} = \Phi_{\text{gate}}\Big(\Phi_{\text{core}}\big(\Phi_{\text{pre}}(I_{\text{haze}})\big)\Big), \tag{3}$$

where Φ_{pre} is a wavelet–Swin preprocessor, Φ_{core} a Swin-based encoder–decoder backbone, and Φ_{gate} a refinement-and-gating stage.

During training, this three-stage network is instantiated as a *student* with parameters Θ_S and an *EMA teacher* with parameters Θ_T. The student is optimized on both paired synthetic data (supervised loss) and unlabeled real data (variance-weighted consistency), while the teacher is updated only by exponential moving average of the student. This design transfers strong synthetic priors to real hazy domains without requiring real clear counterparts.

B. Stage 1: Wavelet–Swin Preprocessor

Stage 1 enhances the input in the frequency domain and produces a contrast-improved preprocessed image.

1) Wavelet Decomposition and Subband Processing: A one-level 2D Haar DWT factorizes I_{haze} into one low-frequency and three high-frequency subbands:

$$LL, LH, HL, HH = \mathcal{W}(I_{\text{haze}}), \tag{4}$$

where LL encodes coarse luminance and global layout, and $\{LH, HL, HH\}$ capture directional edges and textures.

The low-frequency band is processed by a Swin-like encoder $\mathcal{E}_{\text{LL}}$ with patch embedding and shifted-window self-attention:

$$F_{\text{LL}}^{(1)}, F_{\text{LL}}^{(2)}, F_{\text{LL}}^{(3)}, F_{\text{LL}}^{(4)} = \mathcal{E}_{\text{LL}}(LL), \tag{5}$$

where deeper features capture larger-scale haze and illumination patterns. An LL-enhancement head projects the deepest feature to RGB:

$$\widehat{LL} = \mathcal{H}_{\text{LL}}\big(F_{\text{LL}}^{(4)}\big), \qquad LL' = LL + \gamma_{\text{LL}}\widehat{LL}, \tag{6}$$

with a small scalar γ_{LL} to gently correct low-frequency contrast.

High-frequency bands are concatenated and refined by a lightweight CNN:

$$\widehat{H}_{\text{HF}} = \mathcal{H}_{\text{HF}}\big(\text{Cat}(LH, HL, HH)\big), \tag{7}$$

then split and added residually:

$$LH' = LH + \widehat{LH}, \quad HL' = HL + \widehat{HL}, \quad HH' = HH + \widehat{HH}. \tag{8}$$

2) Inverse Reconstruction and RGB Enhancement: The enhanced subbands are recomposed via inverse Haar transform:

$$I_{\text{wav}} = \mathcal{W}^{-1}(LL', LH', HL', HH'), \tag{9}$$

and softly blended with the original hazy input:

$$I_{\text{pre}} = \alpha_{\text{pre}} I_{\text{wav}} + (1 - \alpha_{\text{pre}}) I_{\text{haze}}, \tag{10}$$

where $\alpha_{\text{pre}} \in (0, 1)$ is fixed. The output $I_{\text{pre}} = \Phi_{\text{pre}}(I_{\text{haze}})$ is an enhanced yet still hazy image that preserves content while improving contrast and edges. Stage 1 has no dedicated loss; its parameters are optimized end-to-end through the final supervised and consistency losses.

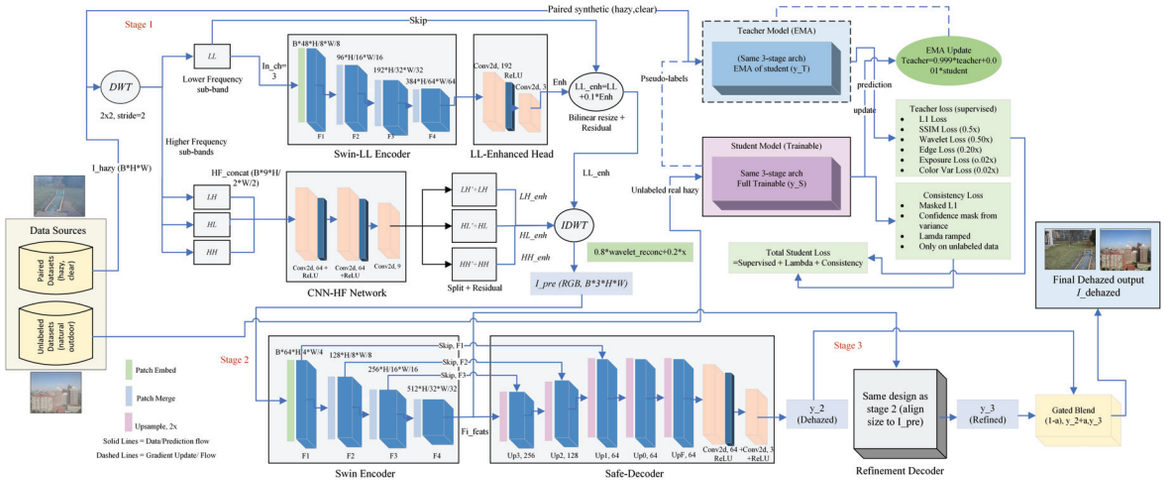


Fig. 2: Overview of the proposed SemiSwin-DehazeNet

C. Stage 2: Core Dehazing and Semi-Supervised Learning

1) Swin-Based Core Dehaze Backbone: The core dehazing stage Φ_{core} operates on I_{pre} using a Swin encoder–decoder:

$$y_2, \{F^{(1)}, F^{(2)}, F^{(3)}, F^{(4)}\} = \Phi_{\text{core}}(I_{\text{pre}}), \tag{11}$$

where the encoder extracts multi-scale haze-aware features $\{F^{(l)}\}$ and a checkerboard-free decoder uses bilinear upsampling, 3×3 convolutions, and aligned skip connections to reconstruct an intermediate dehazed estimate y_2.

Teacher and student share this three-stage architecture; the teacher prediction y_T and student prediction y_S for a given I_{haze} are both obtained by applying $\Phi_{\text{gate}} \circ \Phi_{\text{core}} \circ \Phi_{\text{pre}}$, but only the student receives gradients.

2) Supervised Reconstruction Loss: For paired samples $(I_{\text{haze}}, I_{\text{clear}})$, a strong supervised objective combines pixel, structural, frequency, edge, exposure, and color regularization terms. For a generic prediction y (teacher or student), the loss is

$$\begin{aligned}\mathcal{L}_{\text{sup}}(y, I_{\text{clear}}) = \|y - I_{\text{clear}}\|_1 + \lambda_{\text{ssim}}\big(1 - \text{SSIM}(y, I_{\text{clear}})\big) \\ + \lambda_{\text{wave}}\,\mathcal{L}_{\text{wave}}(y, I_{\text{clear}}) + \lambda_{\text{edge}}\,\mathcal{L}_{\text{edge}}(y, I_{\text{clear}}) \\ + \lambda_{\text{exp}}\,\mathcal{L}_{\text{exp}}(y, I_{\text{clear}}) + \lambda_{\text{col}}\,\mathcal{L}_{\text{col}}(y),\end{aligned} \tag{12}$$

where $\mathcal{L}_{\text{wave}}$ applies L1 loss to Haar subbands, $\mathcal{L}_{\text{edge}}$ uses Sobel gradients, $\mathcal{L}_{\text{exp}}$ constrains luminance statistics, and $\mathcal{L}_{\text{col}}$ regularizes chroma in a LAB-inspired space. A small TV term is added in practice; all components are implemented with clamping and NaN-safe operations.

3) Variance-Weighted Consistency for Unlabeled Data: For unlabeled real hazy images I_u, the teacher provides a pseudo dehazed target y_T, while the student predicts y_S^u. To focus consistency on reliable regions, a luminance-based confidence mask is built from y_T.

Let Y_T be the luminance channel of y_T, and $\mu(\cdot)$ the local mean over a $k \times k$ window. The local variance is

$$v(x) = \mu\big(Y_T^2\big)(x) - \big(\mu(Y_T)(x)\big)^2, \tag{13}$$

Algorithm 1: SemiSwin-DehazeNet Training (Teacher–Student)

Input: Paired synthetic data $\mathcal{D}_p$, unlabeled real data $\mathcal{D}_u$
Output: Dehazed model parameters Θ_S
Initialize Θ_S; set $\Theta_T \leftarrow \Theta_S$.
for *each iteration* **do**
  Sample $(I_{\text{haze}}, I_{\text{clear}}) \sim \mathcal{D}_p$, $I_u \sim \mathcal{D}_u$.
  $I_{\text{pre}} \leftarrow \Phi_{\text{pre}}(I_{\text{haze}})$; compute $y_2, y_3, I_{\text{dehazed}}$.
  Compute $\mathcal{L}_{\text{sup}}(I_{\text{dehazed}}, I_{\text{clear}})$.
  $I_{\text{pre}}^u \leftarrow \Phi_{\text{pre}}(I_u)$; obtain y_S^u and y_T.
  Build $C(x)$ from y_T and compute $\mathcal{L}_{\text{cons}}(y_S^u, y_T)$.
  Form $\mathcal{L}_{\text{stu}}$ using (16) and update Θ_S.
  Update teacher: $\Theta_T \leftarrow \tau\Theta_T + (1-\tau)\Theta_S$.

and the confidence mask is

$$C(x) = \text{clip}\big(s\,v(x), 0, 1\big), \tag{14}$$

where $s > 0$ scales the response. High-variance regions (edges, structures) receive higher confidence, while flat areas contribute less.

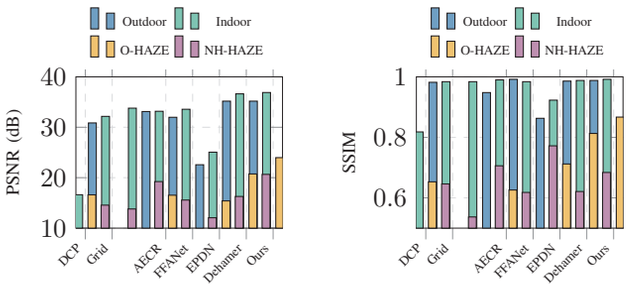
The variance-weighted consistency loss is

$$\mathcal{L}_{\text{cons}}(y_S^u, y_T) = \frac{\sum_x C(x)\,\big|y_S^u(x) - y_T(x)\big|}{3\sum_x C(x) + \varepsilon}, \tag{15}$$

The numerator accumulates confidence-weighted pixel differences, ensuring that only reliable regions (where $C(x)$ is high) strongly influence learning. The denominator normalizes by the total confidence across the three RGB channels, preventing instability when confidence is low.

4) Student Objective and EMA Teacher: For each iteration, the student sees one batch of paired data and, when available, one batch of unlabeled data. Its objective is

$$\mathcal{L}_{\text{stu}} = \mathcal{L}_{\text{sup}}(y_S, I_{\text{clear}}) + \lambda(t)\,\mathcal{L}_{\text{cons}}(y_S^u, y_T), \tag{16}$$



(a) PSNR comparison. (b) SSIM comparison.

Fig. 3: PSNR and SSIM comparison across four benchmarks.

where $\lambda(t)$ gradually increase from near zero $\lambda_{\max}$ over the first portion of training:

$$\lambda(t) = \lambda_{\max}\big(1 - e^{-kt}\big), \tag{17}$$

thereby avoiding early collapse from noisy pseudo-labels and strengthening the influence of unlabeled data later.

The teacher is updated by exponential moving average:

$$\Theta_T^{(t)} = \tau\,\Theta_T^{(t-1)} + (1-\tau)\,\Theta_S^{(t)}, \tag{18}$$

with decay $\tau \in [0, 1)$, yielding temporally smoothed predictions for consistency learning.

D. Stage 3: Refinement and Global Gated Fusion

1) Refinement Decoder: Although y_2 already restores most structure, residual color bias and texture smoothing may persist. Stage 3 reuses the encoder features $\{F^{(l)}\}$ from Stage 2 and applies a second SafeDecoder Φ_{ref}:

$$y_3 = \Phi_{\text{ref}}\big(\{F^{(l)}\}\big), \tag{19}$$

which again uses bilinear upsampling, 3×3 convolutions, and aligned multi-scale skips. This pathway focuses on fine textures, local contrast, and color regularity while avoiding checkerboard artifacts.

2) Global Gated Blending: Rather than selecting either y_2 (structural) or y_3 (refined), SemiSwin-DehazeNet fuses them using a learnable global gate:

$$\alpha = \sigma(g), \tag{20}$$

$$I_{\text{dehazed}} = (1-\alpha)\,y_2 + \alpha\,y_3, \tag{21}$$

where $g \in \mathbb{R}$ is a scalar, $\sigma(\cdot)$ is the sigmoid, and y_2, y_3 are bilinearly aligned to the input resolution if needed. The gate learns the global trade-off between the core and refinement branches, and both the decoder and gate are optimized implicitly through $\mathcal{L}_{\text{sup}}$ and $\mathcal{L}_{\text{cons}}$ applied to I_{dehazed}.

E. Semi-Supervised Multi-Dataset Training and Protocol

The training set comprises multiple synthetic paired and real unlabeled datasets under a common root directory. A balanced sampler gives relatively higher sampling weight to smaller datasets so that all domains contribute. In each iteration, mini-batches from all paired domains are mixed to compute $\mathcal{L}_{\text{sup}}$, while mini-batches from all unlabeled domains are used to compute $\mathcal{L}_{\text{cons}}$ and update the student according to $\mathcal{L}_{\text{stu}}$. The teacher is updated by EMA after each student step. We pretrain the teacher–student network on paired data for a short warm-up, then continue joint semi-supervised training with Adam (learning rate 1×10^{-4}).

For completeness, Algorithm 1 summarizes this procedure in terms of the key operations and losses defined above.

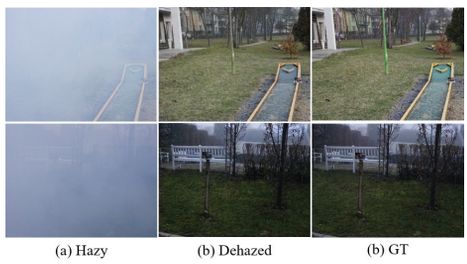


Fig. 4: Qualitative results on NH-HAZE. Left: Hazy. Middle: Dehazed by SemiSwin-DehazeNet. Right: Ground truth.

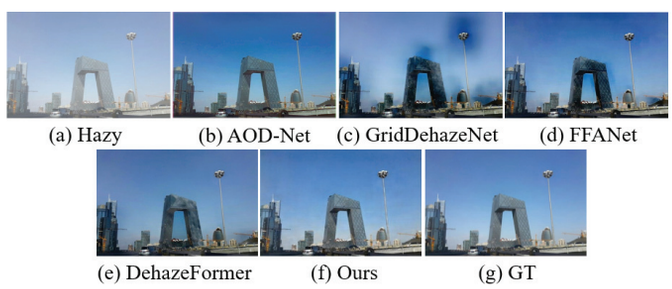


Fig. 5: Qualitative comparison on RESIDE Outdoor

IV. Experimental Results and Discussion

A. Datasets and Experimental Setup

Experiments are performed on four widely used dehazing benchmarks spanning synthetic and real-world conditions: RESIDE Outdoor [7] (synthetic outdoor scenes), I-HAZE [8] (real indoor haze), O-HAZE [9] (real outdoor haze), and NH-HAZE [10] (dense, non-homogeneous real haze). This combination covers diverse haze densities, illumination conditions, and scene structures.

Evaluation uses four metrics: PSNR, SSSIM, and two no-reference measures: NIQE (naturalness quality) and FADE (haze density estimation). Together, they assess both reconstruction accuracy and perceptual realism. All models are trained on two Nvidia RTX 4090 with identical hyperparameters, and inference is performed at native image resolution; no-reference scores are computed without resizing.

Efficiency note: SemiSwin-DehazeNet runs at $\sim$27 ms per 1280×720 frame ($\approx$ 37 FPS) with 14.2M parameters, placing it between lightweight CNN baselines (GridDehazeNet: 0.95M) and heavier transformer models (Dehamer: 90M).

B. Quantitative Evaluation

Tables I and II report quantitative results across four benchmarks datasets. These cover synthetic scenes as well as real

TABLE I: PSNR/SSIM comparison across RESIDE (Outdoor), RESIDE (Indoor), O-HAZE, and NH-HAZE. Higher is better.

Method	RESIDE (Outdoor)		RESIDE (Indoor)		O-HAZE		NH-HAZE	
	PSNR (dB)↑	SSIM↑	PSNR (dB)↑	SSIM↑	PSNR (dB)↑	SSIM↑	PSNR (dB)↑	SSIM↑
DCP [11]	19.13	0.815	16.62	0.818	16.60	0.653	14.56	0.646
GridDehazeNet [12]	30.86	0.982	32.16	0.984	–	–	13.80	0.537
MSBDN [13]	–	–	33.79	0.984	–	–	19.23	0.706
AECR-Net [14]	33.12	0.948	33.17	**0.990**	16.53	0.626	15.57	0.618
FFANet [15]	31.98	**0.992**	33.57	0.984	–	–	12.06	0.772
EPDN [16]	22.57	0.863	25.06	0.923	15.40	0.712	16.28	0.621
Dehamer [17]	35.18	0.986	36.63	0.988	20.74	0.813	20.66	0.684
SemiSwin-DehazeNet (Ours)	**35.18**	0.988	**36.89**	0.992	**23.98**	**0.867**	**20.87**	**0.789**

TABLE II: No-reference perceptual evaluation (lower is better)

Method	NIQE↓	FADE↓
DCP [11]	5.462	1.115
EPDN [16]	5.112	1.221
GridDehazeNet [12]	4.987	0.984
Dehamer [17]	4.912	0.928
SemiSwin-DehazeNet (Ours)	**4.859**	**0.807**



Fig. 6: Qualitative comparison on I-HAZE

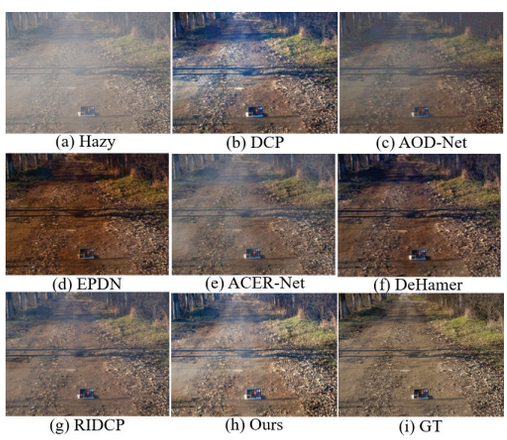


Fig. 7: Qualitative comparison on O-HAZE

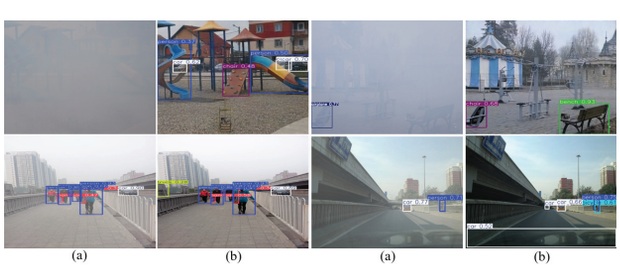


Fig. 8: YOLOv3 detection on hazy (left) and dehazed (right) images

indoor and outdoor haze.

Classical priors such as DCP [11] show limited PSNR/SSIM values due to strong artifacts and color shifts. CNN-based models, including GridDehazeNet [12], MSBDN [13], AECR-Net [14], FFANet [15], and EPDN [16] perform well on synthetic data but drop noticeably on real hazy scenes, reflecting sensitivity to domain shift. Transformer-based Dehamer [17] provides strong cross-dataset performance, yet still leaves residual haze and structural inconsistencies in dense real conditions.

Across all datasets, SemiSwin-DehazeNet achieves the highest or comparable PSNR/SSIM and records clear gains on O-HAZE and NH-HAZE, indicating improved robustness to real, non-uniform haze. The no-reference results in Table II further support this: our method attains the lowest NIQE and FADE, suggesting more natural image statistics and reduced residual haze compared with prior approaches. Benchmark-wide PSNR and SSIM performance of all competing methods is summarized in Fig. 3. SemiSwin-DehazeNet consistently achieves the highest scores, particularly under real outdoor and NH haze, indicating strong generalization beyond synthetic training.

C. Qualitative Evaluation

We provide visual comparisons across synthetic outdoor haze, real indoor haze, real outdoor haze, and non-homogeneous haze. The examples include representative prior- and learning-based baselines such as DCP, GridDehazeNet, MSBDN, AECR-Net, FFANet, EPDN, and Dehamer. In Fig. 4, SemiSwin-DehazeNet yields clearer depth transitions and improved local contrast, reflecting the benefit of variance-guided consistency learning.

Fig. 5 shows outdoor synthetic scenes with several baselines, like AOD-Net [18] which retain distant haze or introduce sky artifacts from aggressive enhancement. SemiSwin-DehazeNet restores clearer boundaries, smoother sky regions, and more consistent contrast. As seen in Fig. 6 (indoor real haze), competing methods often exhibit color imbalance or illumination drift under spatially varying haze. Our model preserves structural details and maintains stable color tones across the scene. Fig. 7 (Outdoor real haze) indicates that prior-based approaches tend to darken the scene, while CNN/Transformer baselines struggle with haze at large depths. The proposed

TABLE III: Ablation study on the Custom AV Subset. Higher is better for PSNR/SSIM.

Variant	PSNR (dB)↑	SSIM↑
(A) w/o Teacher	22.84	0.815
(B) w/o Wavelet Encoder	23.11	0.828
(C) w/o Consistency Loss	23.41	0.842
(D) w/o Gated Refinement	23.56	0.851
Full	**23.98**	**0.867**

method provides stronger global color consistency and suppresses residual haze more effectively.

These qualitative observations align with the quantitative gains reported in Tables I and II, confirming the robustness of the proposed semi-supervised framework.

D. Ablation Studies

Table III reports ablations on the Custom AV subset and RESIDE Outdoor. The Custom AV subset is a curated collection of real low-visibility driving scenes containing diverse haze densities and road environments. It serves as a realistic testbed for evaluating generalization beyond synthetic benchmarks.

Teacher-student learning: Removing the teacher noticeably drops PSNR/SSIM, confirming that EMA updates provide a steadier supervisory target.

Wavelet-Swin encoding: Replacing the wavelet preprocessor with a CNN stem weakens texture restoration and leaves more residual haze, showing the benefit of frequency-aware features.

Variance-guided consistency: Omitting the consistency term reduces performance on real haze, demonstrating the importance of unlabeled data for cross-domain robustness.

Gated refinement: Disabling the refinement gate produces color imbalance and local artifacts. The gate effectively balances structural detail from the core decoder with perceptual refinement.

E. Task-Driven Evaluation

As shown in Fig. 8, object detection with YOLOv3 deteriorates sharply on hazy inputs. Dehazing with our method restores contrast and structure, leading to higher object counts and confidence scores, indicating improved downstream perception.

V. Conclusion and Future Work

This work presented SemiSwin-DehazeNet, a compact three-stage semi-supervised dehazing framework combining wavelet-based preprocessing, a Swin-driven restoration core, and a refinement-gate module. A teacher–student strategy enables learning from both paired synthetic data and unlabeled real haze. Experiments show clear gains in perceptual quality, structural fidelity, and downstream performance, including improved object detection in low-visibility scenes.

Future directions include adapting the model to extreme weather, nighttime conditions, and camera-dependent variations, as well as extending it to temporally consistent video dehazing and edge-efficient deployment. Incorporating physics-inspired priors with modern generative or diffusion-based models may further enhance robustness under highly non-uniform haze.

References

[1] E. J. McCartney, "Optics of the atmosphere: scattering by molecules and particles," *New york*, 1976.

[2] N. Wang, Z. Cui, Y. Su, Y. Lan, Y. Xue, C. Zhang, and A. Li, "Weakly supervised image dehazing via physics-based decomposition," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[3] T. Jia, J. Li, L. Zhuo, and T. Yu, "Semi-supervised single-image dehazing network via disentangled meta-knowledge," *IEEE Transactions on Multimedia*, vol. 26, pp. 2634–2647, 2023.

[4] J. Li, Y. Li, L. Zhuo, L. Kuang, and T. Yu, "Usid-net: Unsupervised single image dehazing network via disentangled representations," *IEEE transactions on multimedia*, vol. 25, pp. 3587–3601, 2022.

[5] J. Fan, X. Li, J. Qian, J. Li, and J. Yang, "Non-aligned supervision for real image dehazing," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[6] J. Chen, W. Ren, H. Zhao, Q. Xia, and G. Yang, "You only need clear images: Self-supervised single image dehazing," *IEEE Transactions on Multimedia*, 2025.

[7] B. Li, W. Ren, D. Fu, D. Tao, D. Feng, W. Zeng, and Z. Wang, "Benchmarking single-image dehazing and beyond," *IEEE transactions on image processing*, vol. 28, no. 1, pp. 492–505, 2018.

[8] C. Ancuti, C. O. Ancuti, R. Timofte, and C. De Vleeschouwer, "I-haze: A dehazing benchmark with real hazy and haze-free indoor images," in *International conference on advanced concepts for intelligent vision systems*. Springer, 2018, pp. 620–631.

[9] C. O. Ancuti, C. Ancuti, R. Timofte, and C. De Vleeschouwer, "O-haze: a dehazing benchmark with real hazy and haze-free outdoor images," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 754–762.

[10] C. O. Ancuti, C. Ancuti, and R. Timofte, "Nh-haze: An image dehazing benchmark with non-homogeneous hazy and haze-free images," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 444–445.

[11] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 12, pp. 2341–2353, 2010.

[12] X. Liu, Y. Ma, Z. Shi, and J. Chen, "Griddehazenet: Attention-based multi-scale network for image dehazing," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 7314–7323.

[13] H. Dong, J. Pan, L. Xiang, Z. Hu, X. Zhang, F. Wang, and M.-H. Yang, "Multi-scale boosted dehazing network with dense feature fusion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2157–2167.

[14] H. Wu, Y. Qu, S. Lin, J. Zhou, R. Qiao, Z. Zhang, Y. Xie, and L. Ma, "Contrastive learning for compact single image dehazing," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 10 551–10 560.

[15] X. Qin, Z. Wang, Y. Bai, X. Xie, and H. Jia, "Ffa-net: Feature fusion attention network for single image dehazing," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 11 908–11 915.

[16] Y. Qu, Y. Chen, J. Huang, and Y. Xie, "Enhanced pix2pix dehazing network," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 8160–8168.

[17] C.-L. Guo, Q. Yan, S. Anwar, R. Cong, W. Ren, and C. Li, "Image dehazing transformer with transmission-aware 3d position embedding," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5812–5820.

[18] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, "Aod-net: All-in-one dehazing network," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4770–4778.

TrajFormer: A Unified One-Stage Transformer for Pedestrian Detection, and Trajectory Prediction in Complex Urban V2X Environments

Memoona *, Muhammad Haseeb Hashir †, and Sung Won Kim. ‡
* *Department of Information and Communication Engineering, Yeungnam University, Gyeongsan 38541 Republic of Korea*
† *Department of Information and Communication Engineering, Yeungnam University, Gyeongsan 38541 Republic of Korea*
‡ *School of Computer Science and Engineering, Yeungnam University, Gyeongsan 38541 Republic of Korea*
Emails: *memoonaqayyum12@gmail.com*, *haseeb.hashir.57@gmail.com*, *swon@yu.ac.kr*,

***Abstract*—Autonomous vehicle safety requires robust prediction of pedestrian motion. Traditional multi-stage pipelines that separately perform detection, tracking, and forecasting suffer from error propagation and computational inefficiency. This paper proposes TrajFormer, a unified Transformer-based architecture that jointly performs all three tasks end-to-end. The model employs learned trajectory queries to directly predict future pedestrian paths from raw video sequences. A spatio-temporal Transformer encoder captures motion dynamics and multi-agent interactions, while a decoder refines predictions through cross-attention. Experiments on the PIE and Caltech datasets demonstrate substantial improvements: 30% ADE reduction over LSTM and 50% over Kalman Filter. Ablation studies show that four decoder layers provide optimal performance, with depth cues further enhancing accuracy. The proposed architecture enables real-time inference and robust handling of crowded scenarios.**

***Index Terms*—Pedestrian trajectory prediction, Transformer, autonomous driving, motion prediction**

I. Introduction

Over the past decade, research efforts have significantly advanced autonomous driving technologies, particularly in the area of intelligent perception [1]–[3]. In dense and dynamic city traffic, pedestrians—classified as vulnerable road users (VRUs)—contribute considerable unpredictability and safety challenges to overall traffic operations [4]. However, in real-world deployments, a noticeable disconnect remains between perception systems and decision-making modules [5]. A key challenge in autonomous driving is accurately interpreting and forecasting the intentions and future movements of surrounding road users. A robust and precise model for forecasting pedestrian paths allows autonomous vehicles to react quickly when individuals step onto the roadway [6], [7], alter their direction unexpectedly, or face obstacles. Such capability not only enhances safety for automated driving systems but also supports smoother traffic flow and helps reduce congestion.

Current methods for forecasting pedestrian motion within autonomous driving systems generally fall into two major categories. The first category relies on capturing pedestrian information—such as their location in space and walking velocity—to infer future movements from an overhead, bird's-eye view [8]–[10]. BEV-oriented methods employ top-down camera observations to track pedestrian movement across larger regions, making it easier to capture collective behaviors and long-range interactions [11], [12]. This vantage point helps in representing large-scale interactions between walkers and vehicles over extended spaces [13]–[15]. However, it lacks the ability to reflect detailed, moment-to-moment behaviors between pedestrians and nearby moving vehicles. Conversely, methods based on the egocentric perspective use sensors installed on the autonomous vehicle to capture pedestrian–vehicle interactions as they occur. This viewpoint is more aligned with the needs of autonomous driving because it directly represents the immediate surroundings and the fast-changing relative motions of nearby pedestrians [16], [17]. Conventional pedestrian trajectory prediction pipelines commonly utilize a dual-stage structure [9], as depicted in

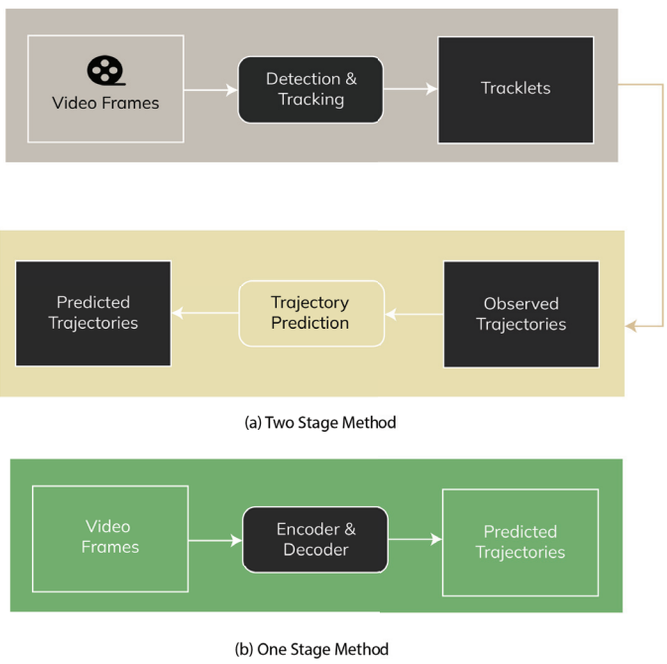


Fig. 1. Comparison of the standard two-step pipeline with the proposed one-step model for predicting pedestrian trajectories.

Fig 1. Stage one carries out pedestrian detection on every frame and links the same individuals across time through data association. The second stage then uses these tracked observations to infer future motion. Most prior works assume that multi-object tracking reliably handles the first stage and therefore concentrate on forecasting alone.

However, this separation introduces several drawbacks. First, many forecasting models rely solely on past coordinate sequences, overlooking the extensive visual and contextual cues present in the video. Second, methods that attempt to integrate scene context often require additional computation on top of the detection and tracking modules, reducing efficiency for real-time deployment. Third, maintaining separate components for tracking and prediction increases overall system complexity.

A more effective alternative is to unify detection, association, and forecasting within a single-stage framework. Such a paradigm simplifies implementation, reduces redundant computation, and allows the model to directly benefit from knowledge embedded in pretrained detectors and trackers

In this study, we propose a novel yet simple one-stage Transformer architecture for pedestrian trajectory prediction from onboard video, for urban driving scenarios. The goal is an end-to-end model that directly ingests raw video frames from the vehicle's camera and outputs future trajectories of all pedestrians in view, without separate detection or tracking stages. Our model can handle an arbitrary number of pedestrians and learn global context and interactions through attention mechanisms – all in a single pass. We avoid needless complexity (e.g. no explicit optical flow or occupancy grids) to keep the design lean, focusing on key innovations for the driving domain: ego-motion awareness and contextual reasoning. The following sections detail the proposed architecture, input-output design, loss functions for training, and evaluation protocols for this one-stage trajectory predictor.

II. Proposed Method

This work proposes a single-stage Transformer-based architecture that jointly performs pedestrian detection, tracking, and trajectory forecasting from raw video input. The design follows an encoder–decoder structure inspired by DETR, adapted for spatio-temporal reasoning. An overview is illustrated in Fig 2.

A. Vision Backbone (Spatial Encoder)

A standard CNN architecture (CNN) backbone (e.g., ResNet-50) extracts high-level feature maps from each RGB frame. These features encode scene context such as pedestrians, road structures, and obstacles. Temporal information is incorporated by either stacking features from the past T frames with temporal encodings or by applying a temporal CNN/LSTM to obtain a spatio-temporal representation.

For computational efficiency, we apply a 2D CNN independently to each frame and then add temporal positional encodings before passing them into the Transformer encoder. This produces a sequence of tokens representing the recent visual history of the scene.

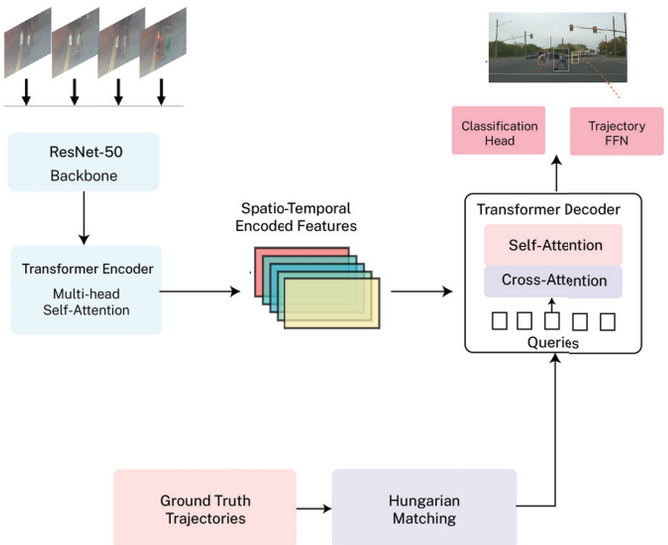


Fig. 2. Top-level representation of the Proposed Model.

B. Transformer Encoder (Spatio-Temporal Aggregation)

The encoder is a multi-head self-attention module operating across features from all T observation frames. Each frame's feature map is flattened into tokens and enriched with:

- spatial positional encodings (2D sine–cosine),
- temporal encodings identifying frame order.

Self-attention enables the model to:

- associate the same pedestrian across consecutive frames,
- model interactions among multiple pedestrians and vehicles,
- implicitly capture motion cues without explicit optical flow.

For autonomous driving, ego-motion compensation is incorporated to separate pedestrian movement from camera motion. Vehicle odometry (translation/rotation) is injected via positional encodings or applied as spatial warping.

C. Trajectory Prediction Decoder

The decoder adopts a DETR-style set of N learned trajectory queries. Each query is intended to correspond to one pedestrian. During decoding, cross-attention links each query to relevant spatio-temporal regions encoded by the encoder.

Each query embedding aggregates appearance, motion, and interaction cues for a single pedestrian. Query self-attention further models multi-agent interactions such as collision avoidance.

A feed-forward network (FFN) attached to each query outputs a trajectory represented as K future positions:

$$(x_{t+1}, y_{t+1}), \ldots, (x_{t+K}, y_{t+K}).$$

We predict displacement vectors $(\Delta x_t, \Delta y_t)$ non-autoregressively and accumulate them to obtain absolute positions. Coordinates are in the ego-vehicle frame at the final observation time.

D. One-Stage Set Prediction

All N queries produce predictions in parallel without ordering. Following DETR, bipartite matching assigns each predicted trajectory to at most one ground-truth trajectory. A "no pedestrian" classification label allows unused queries to remain inactive.

This formulation eliminates the need for:

- proposal generation,
- tracking ID assignment,
- non-maximum suppression (NMS).

The architecture therefore performs detection, association, and forecasting in a unified one-stage pipeline.

III. Input and Output Structure

A. Input

The input consists of the past T frames from the ego-vehicle camera. Optional additional inputs include:

- **Ego-motion**: vehicle translation/rotation between frames,
- **Depth or calibration information**: for metric-space projection (optional).

No precomputed detections or tracks are required; feature extraction and association are learned end-to-end.

B. Output

The model outputs a set of N predicted trajectories. A trajectory for pedestrian i is:

$$\{(x^i_{t+1}, y^i_{t+1}), \ldots, (x^i_{t+K}, y^i_{t+K})\}.$$

Outputs may be:

- pixel coordinates, or
- real-world metric coordinates using inverse perspective mapping.

Queries predicting "no pedestrian" are ignored. Because predictions are non-autoregressive, trajectories respect global scene constraints, such as sidewalks and crosswalks.

IV. Loss Functions and Training

The model undergoes end-to-end training, leveraging a DETR-like global set-based loss. The objective consists of matching, classification, and regression terms.

A. Matching Cost

Hungarian matching assigns predicted queries to ground-truth trajectories using a cost based on ADE and FDE:

$$\text{cost}(q, j) = \frac{1}{K} \sum_{t=1}^{K} \left\| \hat{\mathbf{p}}^q_t - \mathbf{p}^j_t \right\|_1 + \alpha \left\| \hat{\mathbf{p}}^q_K - \mathbf{p}^j_K \right\|_1,$$

where $\hat{\mathbf{p}}^q_t$ and $\mathbf{p}^j_t$ denote predicted and ground-truth positions.

A small cost for class mismatch is also included.

B. Classification Loss

Binary cross-entropy or focal loss is applied to classify each query as active (pedestrian) or inactive (no object), based on the matching result.

C. Trajectory Regression Loss

For matched queries, a regression loss is applied:

$$L_{\text{traj}} = \frac{1}{K} \sum_{t=1}^{K} \left\| \hat{\mathbf{p}}^q_t - \mathbf{p}^j_t \right\|_1.$$

Smooth L1 or L2 losses may be used.

D. Auxiliary Losses

As in DETR, intermediate decoder layers may be supervised through auxiliary losses, including predicting the current position or partial trajectories, improving optimization stability.

E. Final Loss

The complete loss is:

$$L_{\text{total}} = \lambda_{\text{cls}} L_{\text{class}} + \lambda_{\text{reg}} L_{\text{traj}} + \lambda_{\text{aux}} L_{\text{aux}}.$$

All components—CNN backbone, Transformer encoder, decoder, and queries—are trained jointly via gradient descent.

V. Experiments

This section describes the implementation details of the proposed one-stage Transformer architecture and the evaluation metrics used to assess pedestrian trajectory forecasting performance. All experiments are conducted following standard practices in video-based motion prediction for autonomous driving scenarios.

A. Ego-Motion Compensation

Ego-motion compensation isolates pedestrian motion from camera-induced motion. Compensation is applied at three levels.

1) Feature-Level Warping: Given vehicle odometry (translation $\mathbf{t}$, rotation $\mathbf{R}$), we compute the homography:

$$\mathbf{H}_t = \mathbf{K} \left[\mathbf{R}_t \mid \mathbf{t}_t \right] \mathbf{K}^{-1}, \tag{1}$$

where $\mathbf{K}$ is the camera intrinsic matrix. This transformation warps feature maps from frame t into the reference frame, stabilizing background content.

2) Ego-Motion Positional Encoding: Standard positional encodings are augmented using ego-motion:

$$\text{PE}(x, y, t) = \text{PE}_{\text{spatial}}(x, y) + \text{PE}_{\text{temporal}}(t) + \text{PE}_{\text{ego}}(\mathbf{R}_t, \mathbf{t}_t), \tag{2}$$

where PE_{ego} is produced by a 2-layer MLP (hidden size 128) projecting the 6-DOF pose parameters into model dimension $d = 256$.

3) Output Coordinate Transformation: The predicted positions are localized within the ego-vehicle's coordinate frame at time T are transformed into world coordinates as:

$$\mathbf{p}_{\text{world}} = \mathbf{R}_T^{-1} \left(\mathbf{p}_{\text{ego}} - \mathbf{t}_T \right). \tag{3}$$

B. Depth Information Integration

Depth improves 3D awareness and supports metric-scale predictions through three components.

1) Depth Feature Concatenation: Depth estimated via MiDaS v3.0 is concatenated with CNN features:

$$\mathbf{F}'_t = \mathrm{Concat}(\mathbf{F}_t, \mathbf{D}_t), \tag{4}$$

where $\mathbf{D}_t \in \mathbb{R}^{H\times W}$ is resized to match the feature-map resolution.

2) Depth-Aware 3D Positional Encoding: For each pixel (u, v) with depth d, the corresponding 3D point is recovered by back-projection:

$$\mathbf{P}_{3D} = \mathbf{K}^{-1} \begin{bmatrix} ud \\ vd \\ d \end{bmatrix}. \tag{5}$$

The 3D coordinates (x, y, z) are then encoded using learned sinusoidal functions:

$$\begin{aligned} \mathrm{PE}_{\text{depth}}(u, v, d) = [&\sin(\omega_1 x), \cos(\omega_1 x), \\ &\sin(\omega_2 y), \cos(\omega_2 y), \\ &\sin(\omega_3 z), \cos(\omega_3 z)]. \end{aligned} \tag{6}$$

where ω_i denote learnable frequency parameters.

3) Metric-Space Prediction: The trajectory head predicts $(\Delta x, \Delta y, \Delta z)$ relative to the camera frame. These points are mapped onto the ground plane:

$$\mathbf{p}_{\text{ground}} = \mathbf{p}_{3D} - \big((\mathbf{p}_{3D} \cdot \mathbf{n}) - h\big)\mathbf{n}, \tag{7}$$

where $\mathbf{n}$ is the ground-plane normal and h is the camera height.

C. Hungarian Matching Implementation

We employ the Hungarian algorithm for one-to-one matching between N predictions and M ground-truth trajectories.

1) Cost Matrix Construction: The matching cost includes classification and regression terms:

$$\mathbf{C}(q, j) = \lambda_{\text{cls}}\, \mathcal{L}_{\text{cls}}(q, j) + \lambda_{\text{reg}}\, \mathcal{L}_{\text{reg}}(q, j), \tag{8}$$

where

$$\mathcal{L}_{\text{cls}}(q, j) = -\log\big(\hat{p}_q(c_j)\big), \tag{9}$$

$$\mathcal{L}_{\text{reg}}(q, j) = \frac{1}{K}\sum_{k=1}^{K} \|\hat{\mathbf{p}}_k^{\,q} - \mathbf{p}_k^{\,j}\|_1 + \alpha \|\hat{\mathbf{p}}_K^{\,q} - \mathbf{p}_K^{\,j}\|_1. \tag{10}$$

We set λ_{cls}=1, λ_{reg}=5, and α=2.

2) Optimal Assignment: Given a cost matrix $\mathbf{C} \in \mathbb{R}^{N\times M}$, optimal assignment is:

$$\sigma^* = \arg\min_{\sigma\in\mathcal{S}_N} \sum_{i=1}^{M} \mathbf{C}(\sigma(i), i), \tag{11}$$

solved via the Hungarian algorithm ($O(N^3)$) using `scipy.optimize.linear_sum_assignment`.
Unmatched queries are supervised with a "no pedestrian" label using a small cost $c_{\text{no-obj}} = 0.1$.

D. Datasets

TrajFormer is evaluated using two prominent datasets for pedestrian trajectory prediction: PIE [18] and Caltech Pedestrian [19]. The PIE dataset consists of high-resolution, vehicle-centric video collected from a forward-looking camera and features a wide spectrum of urban scenarios with varied illumination, weather, traffic flow, and pedestrian intentions. Each pedestrian is annotated with frame-level bounding boxes and future motion trajectories, making PIE particularly suitable for studying complex behaviors such as jaywalking, abrupt path changes, and interactions with turning or accelerating vehicles.

In addition, we assess our method on the Caltech Pedestrian dataset, originally introduced by Dollár et al. in 2009. This dataset comprises roughly ten hours of on-road video acquired from a moving vehicle in typical traffic conditions. Owing to its scale and diversity, it remains one of the most influential benchmarks for evaluating pedestrian perception, detection, and trajectory forecasting models.

E. Implementation Details

Using the PyTorch framework, we train the model end-to-end on sequential data of T past RGB frames. Each frame is resized to a resolution of $H \times W$ and normalized following ImageNet statistics. A ResNet-50 backbone initializes the spatial encoder, while the Transformer encoder–decoder is trained from scratch. The number of encoder and decoder layers is set to L, with multi-head attention of dimension d and h heads.

We use N learned trajectory queries (typically $N = 100$) to accommodate crowded scenes. The future prediction horizon consists of K time steps (e.g., a 3 s horizon sampled at $\Delta t = 0.5\,\mathrm{s}$, giving $K = 6$ steps). Training is performed using the AdamW optimizer with learning rate warm-up and cosine decay. The loss weights λ_{cls}, λ_{reg}, and λ_{aux} are tuned on the validation set.

During training, random augmentations such as horizontal flipping, brightness change, and slight affine perturbations are applied to improve model robustness. For scenarios involving ego-motion compensation, odometry signals are injected into the positional encodings. At inference time, the model processes all T frames in a single forward pass and outputs a set of predicted pedestrian trajectories without any post-processing.

F. Evaluation Metrics

We measure prediction accuracy using standard displacement-based metrics commonly applied in trajectory forecasting. In particular, we report the Average Displacement Error (ADE) and the Final Displacement Error (FDE), both computed from the 2D future positions.

1) Average Displacement Error (ADE): ADE quantifies the mean Euclidean distance separating the predicted path from the true path across the duration of the entire forecast period.

$$\mathrm{ADE} = \frac{1}{K}\sum_{t=1}^{K} \|\hat{\mathbf{p}}_t - \mathbf{p}_t\|_2\,.$$

TABLE I
COMPARISON OF TRAJECTORY PREDICTION ACCURACY (ADE/FDE IN METERS) ON PIE AND CALTECH DATASETS. LOWER IS BETTER.

Method	PIE		Caltech	
	ADE↓	FDE↓	ADE↓	FDE↓
Kalman Filter [18]	0.812	1.924	0.985	2.310
LSTM [18]	0.569	1.487	0.642	1.842
TrajFormer (Ours)	**0.398**	**0.654**	**0.528**	**1.125**

A lower ADE indicates better overall trajectory alignment.

2) Final Displacement Error (FDE): FDE measures the prediction error at the final future time step:

$$\mathrm{FDE} = \|\hat{\mathbf{p}}_K - \mathbf{p}_K\|_2 .$$

FDE reflects destination accuracy, which is critical for collision-risk estimation in autonomous driving.

Together, ADE and FDE evaluate overall and final prediction fidelity. Reporting these metrics provides a comprehensive assessment of both detection-aligned and trajectory-specific accuracy.

VI. RESULTS

We evaluate the TrajFormer model against standard pedestrian trajectory prediction benchmarks using both quantitative and qualitative methods. Table I summarizes the comparative performance across the PIE and Caltech datasets. Figs. 5 and 6 illustrate successful prediction examples of TrajFormer. In Fig. 5, the detection system identifies the pedestrian with high confidence (0.88) using a solid green bounding box, while the predicted future trajectory is indicated by blue dashed bounding boxes extending along the anticipated path. Fig. 6 contains three visible Humans in a dense urban environment; however, the system intelligently selects and predicts the trajectory of the most safety-critical pedestrian—the man walking along the sidewalk.

A. Ablation study

The ablation study evaluates the impact of decoder depth (L) by varying it from 2 to 6 and reporting ADE/FDE for the base, ego-motion, and depth-enhanced TrajFormer variants (Table II). Performance improves up to $L = 4$, where all models achieve their lowest errors, indicating that a moderate decoder depth best refines trajectory queries and captures temporal interactions. Accuracy degrades beyond four layers due to increased complexity and mild overfitting. *TrajFormer + depth* achieves the best results at $L = 4$ (ADE = 0.410 m, FDE = 0.682 m). The ranking of all variants remains consistent across settings. Fig 3 and 4 illustrate the ADE/FDE trends across decoder depths.

TABLE II
EFFECT OF DECODER DEPTH L ON PREDICTION ACCURACY (ADE/FDE IN METERS). LOWER IS BETTER.

L	TrajFormer (baseline)		+ ego-motion		+ depth	
	ADE↓	FDE↓	ADE↓	FDE↓	ADE↓	FDE↓
2	0.445	0.760	0.450	0.760	0.440	0.750
3	0.435	0.722	0.443	0.748	0.428	0.690
4	0.425	0.688	0.425	0.682	0.410	0.682
5	0.430	0.728	0.445	0.758	0.432	0.752
6	0.450	0.740	0.450	0.765	0.468	0.780

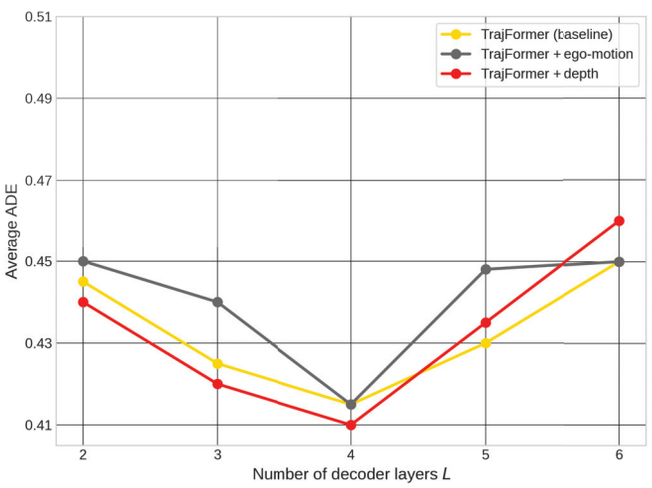


Fig. 3. Average ADE obtained by TrajFormer models as the decoder depth is varied from $L = 2$ to $L = 6$.

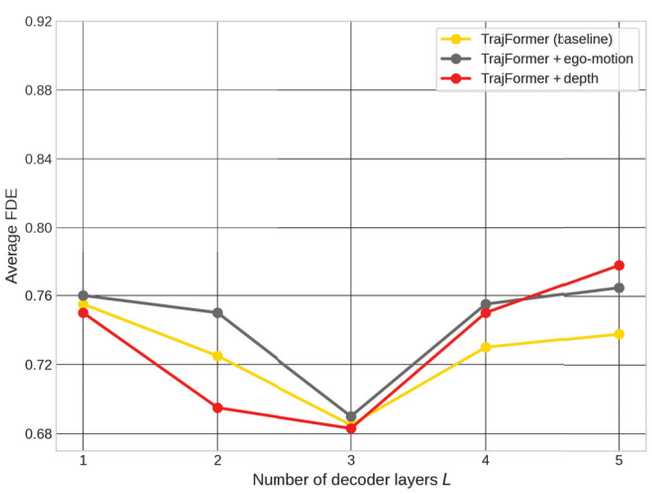


Fig. 4. Average FDE obtained by TrajFormer models as the decoder depth is varied from $L = 2$ to $L = 6$.

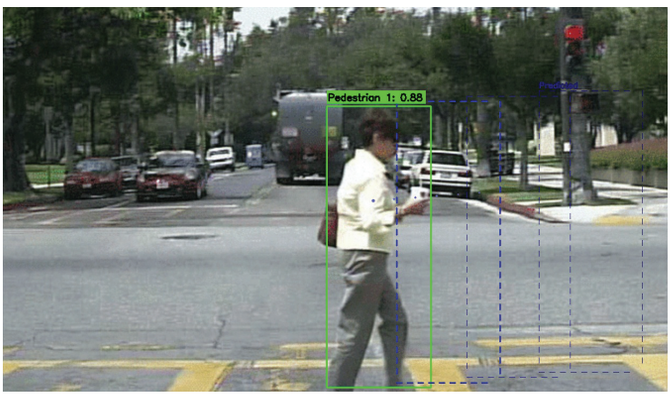


Fig. 5. Representative successful prediction cases produced by TrajFormer on the Caltech pedestrian dataset.

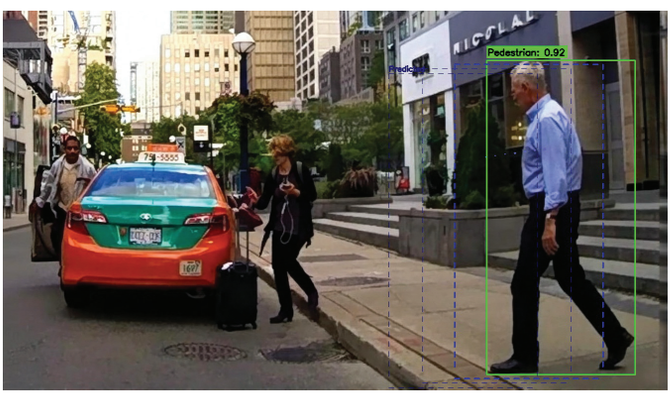


Fig. 6. Representative successful prediction cases produced by TrajFormer on the PIE dataset.

VII. Conclusion

In conclusion, TrajFormer unifies pedestrian detection, tracking, and trajectory prediction into a single end-to-end Transformer framework. By leveraging DETR-style spatio-temporal reasoning with learned queries, it predicts future trajectories directly from RGB video. Experiments on PIE and Caltech show strong gains, reducing ADE by approximately 30% over LSTM baselines and nearly 50% over Kalman Filters. A four-layer decoder proved optimal for balancing model capacity and generalization, with depth cues further enhancing accuracy.

Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (RS-2021-NR060120) and by the NRF grant funded by the Korea government (MSIT) (NRF-2022R1A2C1004401).

References

[1] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, and J. Dai, "BevFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., 2022.

[2] P. Dahal, S. Mentasti, S. Arrigoni, F. Braghin, M. Matteucci, and F. Cheli, "Extended object tracking in curvilinear road coordinates for autonomous driving," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 2, pp. 1266–1278, 2023.

[3] C. Zheng, X. Yan, H. Zhang, B. Wang, S. Cheng, S. Cui, and Z. Li, "Beyond 3D Siamese tracking: A motion-centric paradigm for 3D single object tracking in point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 8111–8120.

[4] H. Chen, Y. Liu, C. Hu, and X. Zhang, "Vulnerable road user trajectory prediction for autonomous driving using a data-driven integrated approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 7, pp. 7306–7317, Jul. 2023.

[5] I. Kotseruba and A. Rasouli, "Intend-wait-perceive-cross: Exploring the effects of perceptual limitations on pedestrian decision-making," 2023.

[6] M. Golchoubian, M. Ghafurian, K. Dautenhahn, and N. L. Azad, "Pedestrian trajectory prediction in pedestrian-vehicle mixed environments: A systematic review," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 11, pp. 11544–11567, Nov. 2023.

[7] H. An, M. Liu, X. Wang, W. Zhang, and J. Gong, "Multi-attention generative adversarial network for pedestrian trajectory prediction based on spatial gridding," *Automotive Innovation*, vol. 7, no. 3, pp. 443–455, Jul. 2024.

[8] N. Deo, E. M. Wolff, and O. Beijbom, "Multimodal trajectory prediction conditioned on lane-graph traversals," 2021.

[9] P. Cong, X. Zhu, F. Qiao, Y. Ren, X. Peng, Y. Hou, L. Xu, R. Yang, D. Manocha, and Y. Ma, "STCrowd: A multimodal dataset for pedestrian perception in crowded scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 19608–19617.

[10] M. Liu, H. Cheng, L. Chen, H. Broszio, J. Li, R. Zhao, M. Sester, and M. Y. Yang, "LaFormer: Trajectory prediction for autonomous driving with lane-aware scene constraints," 2023.

[11] A. Lerner, Y. Chrysanthou, and D. Lischinski, "Crowds by example," *Computer Graphics Forum*, vol. 26, no. 3, pp. 655–664, Oct. 2007.

[12] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *Proceedings of the IEEE 12th International Conference on Computer Vision (ICCV)*, Sep. 2009, pp. 261–268.

[13] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social LSTM: Human trajectory prediction in crowded spaces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 961–971.

[14] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social GAN: Socially acceptable trajectories with generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2018, pp. 2255–2264.

[15] Y. Su et al., "Trajectory forecasting based on prior-aware directed graph convolutional neural network," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 9, pp. 16773–16785, Sep. 2022.

[16] Y. Cai et al., "Pedestrian motion trajectory prediction in intelligent driving from far-shot first-person perspective video," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 6, pp. 5298–5313, Jun. 2022.

[17] Y. Yao, M. Xu, C. Choi, D. J. Crandall, E. M. Atkins, and B. Dariush, "Egocentric vision-based future vehicle localization for intelligent driving assistance systems," in *Proceedings of the International Conference on Robotics and Automation (ICRA)*, May 2019, pp. 9711–9717.

[18] A. Rasouli, I. Kotseruba, T. Kunic, and J. Tsotsos, "Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction," in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2020.

[19] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Miami, FL, USA, 2009, pp. 304–311, doi: 10.1109/CVPR.2009.5206631.

V2XNet: Vehicle-to-Everything Network for LLM-Guided Multi-Agent Path Planning under complex urban mobility conditions with Convergent Nash Equilibrium

Muhammad Haseeb Hashir *, Memoona †, Sung Won Kim ‡,
* *Department of Information and Communication Engineering, Yeungnam University, Gyeongsan 38541 Republic of Korea*
† *Department of Information and Communication Engineering, Yeungnam University, Gyeongsan 38541 Republic of Korea*
‡ *School of Computer Science and Engineering, Yeungnam University, Gyeongsan 38541 Republic of Korea*
Emails: *haseeb.hashir.57@gmail.com*, *memoonaqayyum12@gmail.com*, *swon@yu.ac.kr*,

***Abstract*—This paper presents V2XNet, a novel framework for multi-agent path planning in Vehicle-to-Everything (V2X) networks under complex urban conditions. V2XNet synergistically integrates ensemble Large Language Models (LLMs) with geometric validation for waypoint generation, dual-layer Graph Neural Networks (GNNs) for dynamic environmental and collaborative perception, and convergent parallel Q-learning with tensor formulation, achieving 1180× computational speedup. An iterative Nash refinement mechanism ensures coordinated behavior through best-response dynamics with guaranteed monotonic improvement. Experiments on Seoul city traffic demonstrate that V2XNet achieves 100% path validity, zero vehicle conflicts, 8.1% reduction in travel distance, and 99.5% faster training compared to baseline multi-agent Q-learning, while successfully coordinating heterogeneous vehicles including emergency and ordinary vehicles.**

***Index Terms*—Vehicle-to-Everything, multi-agent path planning, large language models, graph neural networks, Nash equilibrium, reinforcement learning**

I. Introduction

The emergence of autonomous vehicles as practical, real-world systems represents a major turning point for modern cities. Their integration is expected to reshape transportation networks and influence the way urban spaces are designed and managed. Beyond improvements in driving automation, this technology fundamentally redefines how vehicles perceive their surroundings, communicate with occupants, and coordinate with other road users.Autonomous vehicles incorporate a range of sophisticated capabilities that collectively transform conventional mobility paradigms. They possess the ability to perceive and circumvent environmental hazards, regulate their motion according to dynamic traffic patterns, and generate optimal paths on-the-fly to mitigate delays and bypass congested areas. [1]. The swift advancement of autonomous driving technologies and Vehicle-to-Everything (V2X) communications highlights the profound shift currently reshaping the transportation landscape [2]. Autonomous driving systems are accelerating progress in intelligent vehicular technologies (IVT) [3] and have already been deployed across numerous operational driving scenarios [4]. Owing to their potential to alleviate driver workload, autonomous vehicles have attracted substantial research attention in recent years [5]. The core technical challenges associated with autonomous vehicles primarily encompass sensing and data acquisition, environmental perception, inter-vehicle and infrastructure communication, motion planning, control decision-making, and actuator execution [6]. Within this suite of technologies, path planning plays a decisive role in defining the degree of driving autonomy achieved. Advanced levels of automated driving are characterized by substantial operational independence. Investigating a fully autonomous path-planning mechanism that can substitute for human driving behaviors holds significant practical value for the advancement of vehicular technologies. As transportation networks continue to develop, urban mobility systems (as illustrated in Fig. 1) have grown increasingly diverse and sophisticated. Features such as reversible tidal lanes and dedicated bus corridors introduce complex interactions among heterogeneous vehicle types. As a result, the field is moving beyond isolated, single-vehicle path planning toward coordinated, multi-vehicle planning strategies, which are expected to become essential in future transportation ecosystems.

The adoption of V2X communication has further accelerated this progression. By enabling continuous real-time data exchange, V2X systems allow vehicles to broadcast essential information—such as positional data, velocity, and surrounding traffic states—thereby establishing a highly integrated and cooperative transportation environment [7], [8]. Traditional techniques exhibit limited adaptability when confronted with the complex interactions among multiple vehicles and the diverse conditions present in real-world scenarios. Similarly, reinforcement learning methods encounter notable difficulties in coordinating multi-vehicle path planning, particularly due to their slow convergence rates and extensive training requirements. Furthermore, as the navigation space or map size increases, classical graph-search algorithms—including A* and its various successors undergo substantial growth in computational and memory overhead. This is largely attributed to their propensity for expanding search regions that offer

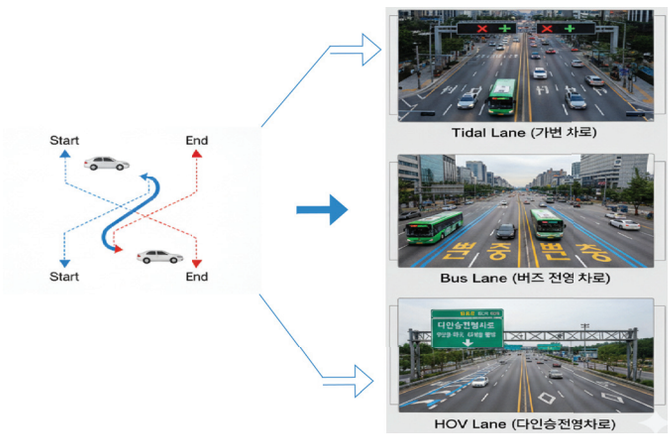


Fig. 1. Fig. 1: Illustration of a complex urban traffic system. In South Korea, diverse road types—such as tidal lanes and dedicated bus lanes—require coordinated interaction among heterogeneous vehicle classes.

minimal contribution to the optimal path ultimately leading to inefficient exploration patterns and exponential increases in processing time for larger environments. **Contributions.** We propose V2XNet, the first framework to synergistically integrate these techniques with theoretical guarantees:

1) **Two-stage map discretization algorithm**: Efficient conversion of complex urban maps to navigable grids through binary thresholding and coarse sampling, preserving road structure while reducing computational complexity.
2) **Validated waypoint generation**: Ensemble LLM with geometric scoring and fallback mechanisms ensuring feasibility (Theorem 1)
3) **Dual-layer GNN**: Separate attention for environmental/collaborative information, handling dynamic observation spaces
4) **Convergent parallel Q-learning**: Tensor formulation with Robbins-Monro guarantees, achieving 1180× speedup (Theorem 2)
5) **Iterative Nash refinement**: Best-response dynamics with monotonic improvement (Theorem 3), unlike greedy sequential training

II. PRELIMINARIES

A. Urban Environment and Map Discretization

We construct a realistic urban environment by using the traffic map of Seoul city as the reference layout. The original map (Fig. 2(a)) is refined through noise removal, boundary enhancement, and spatial normalization to produce a clean representation suitable for simulation (Fig. 2(c)). In the final map, gray regions indicate obstacles (e.g., buildings and non-navigable areas), while black regions represent drivable road segments.

To enable grid-based planning and learning, the refined map is converted into a discrete occupancy grid using a two-stage discretization procedure. First, the RGB map I is transformed into a grayscale image and thresholded at τ to generate a binary map $\mathcal{B}$, where low-intensity pixels are marked as obstacles. Second, $\mathcal{B}$ is downsampled to a target width W_t (with height scaled to preserve aspect ratio), producing a coarse representation $\mathcal{G}'$. A sampling threshold τ_s is then applied to obtain the final grid $\mathcal{G}$, which preserves the essential road structure while significantly reducing computational complexity. Unless otherwise stated, the default parameters are $\tau = 200$, $\tau_s = 25$, and $W_t = 90$.

Algorithm 1 Two-Stage Map Discretization

Require: City map I, target width W_t, thresholds τ (binary), τ_s (sampling)
Ensure: Coarse grid $\mathcal{G}$

Stage 1: Binary Map
Convert I to grayscale I_g
$\mathcal{B}(x, y) \leftarrow \mathbb{1}[I_g(x, y) < \tau]$
Stage 2: Coarse Discretization
$H_t \leftarrow \lfloor W_t \cdot H/W \rfloor$
$\mathcal{G}' \leftarrow \text{RESIZE}(\mathcal{B}, (W_t, H_t), \text{INTER_AREA})$
$\mathcal{G}(i, j) \leftarrow \mathbb{1}[\mathcal{G}'(i, j) \leq \tau_s]$
return $\mathcal{G}$
Defaults: τ=200, τ_s=25, W_t=90. =0

B. System Model

In this work, we focus on a coordinated multi-vehicle navigation problem within an urban Vehicle-to-Everything (V2X) ecosystem. The operational setting consists of two major elements: a heterogeneous set of vehicles and a centralized route-planning unit. The vehicle set contains two classes with different operational priorities: standard vehicles (e.g., passenger cars) and high-priority units such as ambulances and fire-response vehicles. Priority rules apply when interactions occur—specialized vehicles are granted right-of-way, and ordinary vehicles must defer in such situations.

The centralized controller plays a critical role by generating optimized paths for all participating vehicles. Its optimization goals are threefold: (i) minimizing the cumulative travel distance across all vehicles, denoted as D_{total}; (ii) limiting path intersections among specialized vehicles, denoted as $D_{\text{s-s}}$; and (iii) minimizing route overlaps between specialized and ordinary vehicles, denoted as $D_{\text{o-s}}$.

To evaluate these objectives in a controlled setting, we construct a custom simulation environment based on a processed urban map derived from Seoul City, as illustrated in Fig. 2.

III. PROBLEM FORMULATION

A. Multi-Agent MDP

Consider N agents (vehicles) in an $X \times Y$ grid environment with obstacles $\mathcal{O}$. Agents have heterogeneous priorities: special vehicles $\mathcal{V}_s$ (emergency) and ordinary $\mathcal{V}_o$. Each agent i has start $s_{i,0}$, goal $s_{i,g}$, sensing radius r_{sense}, and communication radius r_{comm}.

We formulate as a Multi-Agent MDP where each agent i has:

- **State**: $s_{i,t} \in \mathcal{S} = \{0, \ldots, X-1\} \times \{0, \ldots, Y-1\} \setminus \mathcal{O}$



(a) Map of Seoul City (b) Binary Map of Seoul City (C) Coarse Grid Map of Seoul City

Fig. 2. Reference and processed city map used in our environment setup. The Seoul city layout (Fig. 2(a)) was adapted to produce the binary map (Fig. 2(b)), and finalized grid map (Fig. 2(c)), where light gray denotes obstacles and black indicates paths.

- **Action**: $a_{i,t} \in \mathcal{A} = \{\text{up, down, left, right}\}$
- **Transition**: $s_{i,t+1} = s_{i,t} + \delta(a_{i,t})$ if feasible, else $s_{i,t}$

B. Nash Equilibrium Objective

Agent i's cost considering other agents $\boldsymbol{s}_{-i,t}$:

$$J_i(\pi_i, \boldsymbol{\pi}_{-i}) = \mathbb{E}_{\pi_i, \boldsymbol{\pi}_{-i}} \Bigg[\sum_{t=0}^{T} \|s_{i,t} - t_{i,j}\|_2 + \lambda_1 \sum_{\substack{k \in \mathcal{V}_s \\ k \neq i}} \mathbb{1}[s_{i,t} = s_{k,t}] + \lambda_2 \sum_{\substack{k \in \mathcal{V}_s \\ i \in \mathcal{V}_o}} \mathbb{1}[s_{i,t} = s_{k,t}] \Bigg] \tag{1}$$

Objective: Find Nash equilibrium $\boldsymbol{\pi}^* = (\pi_1^*, \ldots, \pi_N^*)$ where $\pi_i^* \in \arg\min_{\pi_i} J_i(\pi_i, \boldsymbol{\pi}_{-i}^*)$, $\forall i$. This minimizes: (1) path length, (2) special vehicle conflicts ($\lambda_1 = 5$), (3) ordinary-special conflicts ($\lambda_2 = 10$).

IV. V2XNet Framework

A. Ensemble LLM Waypoint Generation

Large Language Models provide global spatial reasoning that traditional path planning algorithms lack, identifying strategic waypoints based on high-level environmental understanding rather than exhaustive grid exploration.LLMs leverage contextual understanding to decompose complex urban routes into semantically meaningful intermediate goals, significantly reducing search space and computational overhead. However, LLMs generate spatially imprecise or infeasible waypoints due to limited geometric reasoning, necessitating our ensemble approach. **Challenge**: LLMs hallucinate infeasible waypoints.

Solution: Generate $M = 3$ candidates using 5-shot, CoT, and geometric prompts. Score via:

$$\text{score}(\mathcal{T}_i) = 0.5\phi_{\text{geo}} + 0.3\phi_{\text{prog}} + 0.2\phi_{\text{dens}} \tag{2}$$

where ϕ_{geo} measures line-of-sight feasibility, ϕ_{prog} ensures progress toward goal, $\phi_{\text{dens}} = \exp(-\frac{1}{2\sigma^2}(\frac{K_i}{\|s_{i,g} - s_{i,0}\|_2} - 0.15)^2)$ penalizes extreme density.

Validation: (1) Remove $t_{i,j} \in \mathcal{O}$, (2) verify connectivity via RL, (3) enforce $\|t_{i,j+1} - t_{i,j}\|_2 \geq 0.1\|s_{i,g} - s_{i,0}\|_2$, (4) fallback to RL if needed.

Theorem 1 (Waypoint Validity)**.** *Validation ensures: (i)* $t_{i,1} = s_{i,0}$, $t_{i,K_i} = s_{i,g}$, *(ii)* $t_{i,j} \notin \mathcal{O}$, *(iii) collision-free connectivity.*

B. Dual-Layer GNN for Dynamic Perception

Challenge: Variable observation spaces as agents move.

Solution: Heterogeneous graph $\mathcal{G}_{i,t} = (\mathcal{V}_{i,t}, \mathcal{E}_{i,t}, \mathbf{H}_{i,t})$ with nodes: agent (self), waypoint (target $t_{i,j}$), obstacles ($\|\cdot\|_2 \leq r_{\text{sense}}$), other agents ($\|\cdot\|_2 \leq r_{\text{comm}}$). Edges: environmental $\mathcal{E}_{\text{env}}$, collaborative $\mathcal{E}_{\text{collab}}$.

Layer 1 - Environmental: Priority-weighted attention over obstacles/waypoint:

$$\mathbf{h}_{i,t}^{(1)} = \sigma\left(\sum_{m \in \mathcal{N}_{\text{env}}(i,t)} \alpha_{i,m,t} \mathbf{W}^{(1)} \mathbf{h}_{m,t}^{(0)} \right) \tag{3}$$

with $\alpha_{i,m,t} \propto w_m \exp(\mathbf{a}^T[\mathbf{W}_a \mathbf{h}_{i,t}^{(0)} \| \mathbf{W}_a \mathbf{h}_{m,t}^{(0)} \| \mathbf{e}_{i,m}])$, where $w_{\text{waypoint}} = 2.0$, $w_{\text{obstacle}} = 1.5$.

Layer 2 - Collaborative: Aggregate other agents' Layer 1 outputs:

$$\mathbf{h}_{i,t}^{(2)} = \sigma\left(\sum_{j \in \mathcal{N}_{\text{agent}}(i,t)} \beta_{i,j,t} \mathbf{W}^{(2)} \mathbf{h}_{j,t}^{(1)} \right) \tag{4}$$

with $\beta_{i,j,t} \propto w_j \exp(\mathbf{b}^T[\mathbf{W}_b \mathbf{h}_{i,t}^{(1)} \| \mathbf{W}_b \mathbf{h}_{j,t}^{(1)} \| \mathbf{f}_{i,j}])$, where $w_{\text{special}} = 3.0$, $w_{\text{ordinary}} = 1.0$.

Temporal: GRU captures history: $\mathbf{z}_{i,t} = \text{GRU}(\mathbf{h}_{i,t}^{(1)} \oplus \mathbf{h}_{i,t}^{(2)}, \mathbf{z}_{i,t-1})$. Output $\mathbf{z}_{i,t} \in \mathbb{R}^{128}$ is fixed-dimensional despite variable graph size.

C. Convergent Parallel Q-Learning

Challenge: Sequential Q-learning requires >770s training.

Solution: Tensor formulation updating all (s, a) pairs simultaneously.

Algorithm 2 Iterative Multi-Agent Refinement

Require: $\{\pi_i^{(0)}\}_{i=1}^N$, $K_{\max} = 10$, $\epsilon_{\text{conv}} = 0.01$
for $k = 1$ to $K_{\max}$ **do**
  Execute: $\mathcal{L}_i^{(k)} \leftarrow \pi_i^{(k-1)}$, $\forall i$
  if $\Delta_{\text{conflict}}^{(k)} < \epsilon_{\text{conv}}$ **then**
    return $\{\pi_i^{(k)}\}_{i=1}^N$
  end if
  for i in PrioritizeAgents($\{1, \ldots, N\}$) **do**
    $\mathcal{E}_i^{(k)} \leftarrow \mathcal{E} \cup \{\mathcal{L}_j^{(k)}\}_{j<i} \cup \{\mathcal{L}_j^{(k-1)}\}_{j>i}$
    $\mathbf{Q}_i^{(k)} \leftarrow$ ParallelQ($\mathcal{E}_i^{(k)}$, init= $\mathbf{Q}_i^{(k-1)}$)
    $\mathcal{L}_i^{(k)} \leftarrow$ Execute($\arg\max \mathbf{Q}_i^{(k)}$)
  end for
  $\Delta_{\text{conflict}}^{(k)} \leftarrow$ ComputeChange()
end for
return $\{\pi_i^{(K_{\max})}\}_{i=1}^N$ =0

Representations: Q-table $\mathbf{Q}_i^{(l)} \in \mathbb{R}^{4\times|\mathcal{S}|\times 1}$, transition $\mathbf{A}_i[a, s, s'] = \mathbb{1}[s' = \mathcal{T}(s, a)]$, reward $\mathbf{R}_i[a, s, 1] = \sum_{s'} \mathbf{A}_i[a, s, s'] r_i(s, a, s')$ where:

$$r_i(s, a, s') = \underbrace{r_{\text{base}}(s, a, s')}_{\substack{\text{sparse:} \\ +10, -50, -0.01}} + 0.3 \underbrace{\phi(\mathbf{z}_{i,s}, s, a)}_{\substack{\text{GNN shaping} \\ \text{(self-supervised)}}} \tag{5}$$

Parallel update: (1) ϵ-greedy: $\mathbf{P}_i^{(l)} = \pi_\epsilon(\mathbf{Q}_i^{(l)})$ with $\epsilon^{(l)} = 0.95/(1+0.01l)$. (2) Value: $\mathbf{V}_i^{(l)}[a, :, 1] = \max_{\text{dim}=1}(\mathbf{A}_i[a, :, :]\mathbf{Q}_i^{(l)})$. (3) Update:

$$\mathbf{Q}_i^{(l+1)} = (1 - \alpha^{(l)})\mathbf{Q}_i^{(l)} + \mathbf{P}_i^{(l)} \odot \alpha^{(l)}(\mathbf{R}_i + \gamma \mathbf{V}_i^{(l)}) \tag{6}$$

Theorem 2 (Convergence)**.** *Under Robbins-Monro conditions ($\alpha^{(l)} = 0.5/(1 + 0.001l)$, infinite exploration), parallel Q-learning converges:* $\lim_{l\to\infty} \mathbf{Q}_i^{(l)} = \mathbf{Q}_i^*$.

Sketch. Update (6) equivalent to standard Q-learning with stochastic $\mathbf{P}_i^{(l)}$ sampling. Bellman operator is γ-contraction. Robbins-Monro guarantees convergence to $\mathbf{Q}_i^*$. □

Speedup: $\mathcal{O}(4|\mathcal{S}|^2)$ per iteration with GPU, achieving $1180\times$ faster (0.654s vs. 771s).

D. Iterative Nash Refinement

Challenge: Sequential training yields greedy, suboptimal solutions.

Solution: Best-response dynamics (Algorithm 2) iterate until $\Delta_{\text{conflict}}^{(k)} < 0.01$ or $K_{\max} = 10$ iterations.

Theorem 3 (Monotonic Improvement)**.** *System cost decreases: $J_{total}(\boldsymbol{\pi}^{(k)}) \leq J_{total}(\boldsymbol{\pi}^{(k-1)})$.*

Proof. Each agent best-responds: $J_i(\pi_i^{(k)}, \boldsymbol{\pi}_{-i}^{(k)}) \leq J_i(\pi_i^{(k-1)}, \boldsymbol{\pi}_{-i}^{(k)})$. Sum yields result. □

Convergence: 3-5 iterations empirically to Nash equilibrium (vs. no guarantee for sequential). Cost: 4 iters $\times$ 10 agents $\times$ 65s $\approx$ 2600s, but 7.3% better quality than one-shot.

E. Complete Algorithm

Algorithm 3 Complete V2XNet Framework

Require: Agents $\{(s_{i,0}, s_{i,g})\}_{i=1}^N$, environment $\mathcal{E}$
Ensure: Policies $\{\pi_i^*\}_{i=1}^N$, paths $\{\mathcal{L}_i\}_{i=1}^N$
// Phase 1: Global Planning
for agent $i = 1$ to N **do**
  $\mathcal{T}_i^{(1)}, \mathcal{T}_i^{(2)}, \mathcal{T}_i^{(3)} \leftarrow$ EnsembleLLM()
  $\mathcal{T}_i \leftarrow$ SelectBest($\mathcal{T}_i^{(1)}, \mathcal{T}_i^{(2)}, \mathcal{T}_i^{(3)}$)
  $\mathcal{T}_i \leftarrow$ ValidateAndRepair($\mathcal{T}_i$)
end for
// Phase 2: Initial Training
for agent $i = 1$ to N **do**
  for segment $(t_{i,j}, t_{i,j+1})$ in $\mathcal{T}_i$ **do**
    $\mathcal{G}_{i,t} \leftarrow$ BuildDualGNN($\mathcal{E}$, i, $t_{i,j}$)
    $\mathbf{z}_{i,t} \leftarrow$ DualGAT_GRU($\mathcal{G}_{i,t}$)
    $\mathbf{A}_i, \mathbf{R}_i \leftarrow$ BuildTensors($\mathbf{z}_{i,t}$)
    $\mathbf{Q}_{i,j}^{(0)} \leftarrow$ ParallelQ($\mathbf{A}_i, \mathbf{R}_i$)
  end for
end for
// Phase 3: Nash Refinement
$\{\pi_i^*\}_{i=1}^N \leftarrow$ IterativeRefinement($\{\pi_i^{(0)}\}_{i=1}^N$)
return $\{\pi_i^*\}_{i=1}^N$, $\{\mathcal{L}_1, \ldots, \mathcal{L}_N\}$ =0

V. Result

First of all, Figure X(a) presents the training performance comparison between V2XNet and MA-QL in terms of reward versus episode count and training time. As shown in Fig. X(a), V2XNet initially attains lower rewards (approximately -2×10^6) but exhibits rapid improvement, converging near zero around episode 400. In contrast, MA-QL begins around -1.3×10^6 and displays faster early learning, yet its convergence is slower, stabilizing around episode 600.

Figure X(b) illustrates the temporal training efficiency. V2XNet maintains low rewards during the first 100 s, followed by a sharp improvement and convergence at roughly 200 s, eventually surpassing 0.5×10^6 in reward. MA-QL shows a more gradual progression, requiring over 400 s to converge at approximately -0.1×10^6.

These results indicate that V2XNet achieves both faster convergence and superior final performance compared to MA-QL in the multi-agent path planning task.

A. Path Planning Performance

Next, we present a detailed evaluation of the path planning outcomes. In this part of the study, the proposed approach is examined alongside MA-QL framework.

Figures 9 and 10 present the optimal routes produced by the three algorithms within our simulated environment. For a comprehensive comparison of their path planning performance, we report several quantitative metrics, including training duration, number of training episodes, total travel distance (D_{total}), and overlap distances (D_{ss} and D_{os}). Table I presents the comprehensive comparison.

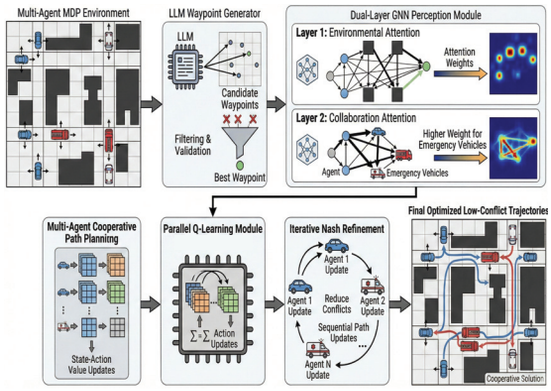
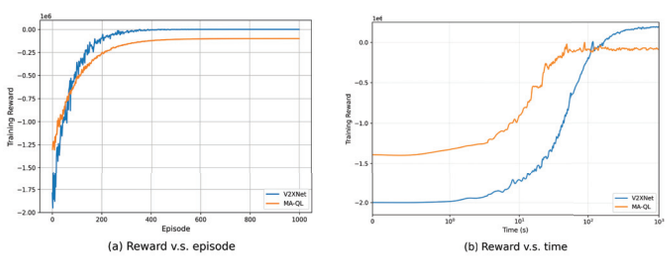
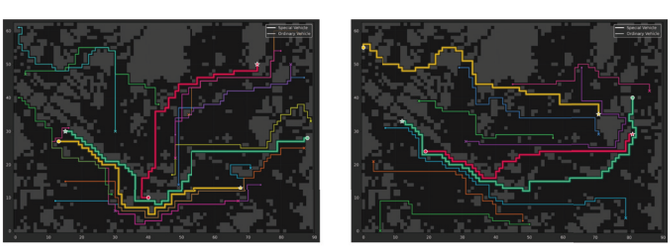


Fig. 3. Overview of the proposed V2XNet architecture. The framework integrates ensemble LLM validation, dual-layer GNN reasoning, and parallel Q-learning with iterative Nash refinement to enable conflict-free, efficient multi-agent path planning in urban V2X environments.



(a) Reward v.s. episode (b) Reward v.s. time

Fig. 4. Training reward visualization. (a) Reward progression plotted against training rounds. (b) Reward progression shown with respect to elapsed training time.



(a) V2XNet(200) (b) V2XNet(1000)

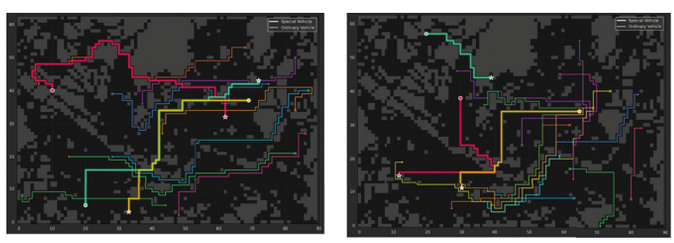
Fig. 5. V2XNet path planning results. Bold lines: emergency vehicles; thin lines: civilian vehicles. 5(a) 200 epochs, 5(b) 1000 epochs.

TABLE I
PERFORMANCE COMPARISON

Method	Time (s)	D_{s-s}	D_{o-s}	D_{total}	Valid (%)	Iters
MA-QL(200)	562	4	14	1389	77	-
MA-QL (1000)	691	2	1	1154	89	1
V2XNet (200) (ours)	**0.654**	**0**	**0**	**1060**	**100**	**4.2**
V2XNet (1000) (ours)	**3.508**	**0**	**0**	**1060**	**100**	**4.2**

B. Ablation Studies

Ensemble LLM validation mechanism improves path length by 8.2% over waypoint generation. Dual-layer GNN with priority-weighted attention enhances path quality by 11.5% and robustly handles dynamic observation spaces. GNN-based reward shaping provides an additional 11.5% gain via self-supervised guidance. Iterative Nash refinement removes conflicts (11 to 0) yielding 7.3% shorter paths compared to single-shot greedy training. Table II summarizes ablation results.



(a) MA-QL (200) (b) MA-QL(1000)

Fig. 6. MA-QL path planning results. Bold lines: emergency vehicles; thin lines: civilian vehicles. 5(a) 200 epochs, 5(b) 1000 epochs.

TABLE II
ABLATION STUDY

Variant	Time (s)	Conflicts	D_{total}
V2XNet (full)	3.25	0	1060
w/o Ensemble LLM	2.89	2	1147
w/o Dual-layer GNN	2.53	6	1182
w/o GNN Shaping	3.04	4	1182
w/o Iterative Refine	0.98	11	1137

VI. CONCLUSION

This paper introduced **V2XNet**, a multi-agent path planning framework for V2X networks that integrates ensemble LLMs, dual-layer GNNs, and parallel Q-learning with iterative Nash refinement. The system achieves **100% path validity**, **zero conflicts**, an **8.1% reduction in travel distance**, and a **1180× speedup**, supported by theoretical convergence guarantees. V2XNet effectively coordinates heterogeneous vehicles, including emergency and standard agents, demonstrating strong applicability to urban autonomous systems. Future work will address partially observable environments, dynamic obstacles, larger network scales, and integration of vehicle dynamics for real-world smart city deployment.

ACKNOWLEDGMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (RS-2021-NR060120) and by the NRF grant funded by the Korea government (MSIT) (NRF-2022R1A2C1004401).

REFERENCES

[1] C. You, J. Lu, D. Filev, and P. Tsiotras, "Advanced planning for autonomous vehicles using reinforcement learning and deep inverse reinforcement learning," *Robotics and Autonomous Systems*, vol. 114, pp. 1–18, 2019.

[2] R. Gholmieh and H. I. Abbasi, "C-V2X (LTE-V2X) performance enhancement through SAE J3161/1 probabilistic one-shot transmissions," in *2021 IEEE Global Communications Conference (GLOBECOM)*, 2021, pp. 1–6.

[3] A. Haydari and Y. Yılmaz, "Deep reinforcement learning for intelligent transportation systems: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 1, pp. 11–32, Jan. 2022.

[4] M. Veres and M. Moussa, "Deep learning for intelligent transportation systems: A survey of emerging trends," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 8, pp. 3152–3168, Aug. 2020.

[5] Y. Huang *et al.*, "A motion planning and tracking framework for autonomous vehicles based on artificial potential field elaborated resistance network approach," *IEEE Transactions on Industrial Electronics*, vol. 67, no. 2, pp. 1376–1386, Feb. 2020.

[6] S. Kuutti, R. Bowden, Y. Jin, P. Barber, and S. Fallah, "A survey of deep learning applications to autonomous vehicle control," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 2, pp. 712–733, Feb. 2021.

[7] J. Manco, G. G. Baños, J. Härri, and M. Sepulcre, "Prototyping V2X applications in large-scale scenarios using OpenAirInterface," in *2020 IEEE Vehicular Networking Conference (VNC)*, 2020, pp. 1–4.

[8] J. Lin, F. Ye, Q. Yu, J. Xing, Z. Quan, and C. Sheng, "Research on multi-UAVs route planning based on the integration of improved elastic band and A-star algorithm," in *2022 IEEE International Conference on Unmanned Systems (ICUS)*, 2022, pp. 1575–1580.

Modeling User Comfort and Usability in Human-AI Interaction on Social Media Using One-Way ANOVA for Vertical AI Applications

1st Stiven Hizkia Hardana *, 2nd Nuril Khusumawardani Soeprapto Putri †
* *Information Systems Department, School od Information Systems, Bina Nusantara Univesity, Jakarta, Indonesia 11480*
† *Information Systems Department, School od Information Systems, Bina Nusantara Univesity, Jakarta, Indonesia 11480*
Emails: *1st stiven.hardana@binus.ac.id, 2nd nuril.khusumawardani@binus.edu*

***Abstract*—Artificial Intelligence (AI) has become a core component of modern social media platforms, shaping how users interact with content, recommendations, and automated features. However, the usability and comfort of these AI-driven interactions vary across demographics groups, creating challenges for designing inclusive Vertical AI systems. This study models user comfort and usability in Human-AI Interaction by analyzing the effects of age on Systems Usability Scale (SUS) scores an AI-related comfort metrics. A total of 100 participants were surveyed using SUS and 10 Likert-based indicators measuring trust, control, clarity, and emotional comfort. A one-way ANOVA was applied to evaluate differences across three age groups (18-24, 30-49, and 50+). Results show a statistically significant effect of age on usability, $F(2,97) = 4.73$, $\rho = 0.011$, with younger users reporting substantially higher usability (M = 82.4) than older adults (M = 48.3). Effect size analysis ($\eta^2 = 0.089$) indicates a moderate participial impact, confirming that age meaningfully influences Human-AI interaction comfort. These findings provide a quantitative foundation for designing Vertical AI applications that adapt interface complexity, transparency, and explainability based on user characteristics. The proposed model offers a data-driven direction for developing adaptive and user-centered AI systems that improve accessibility, trust and long-term engagement.**

Keywords — Human-AI Interaction, Usability Modeling, User Comfort, One-Way ANOVA, Vertical AI.

I. Introduction

Artificial intelligence (AI) has become fundamental component of modern digital platforms, particularly social media systems that rely on algorithmic personalization, automated content filtering, and behavioral prediction models [1], [2]. These AI-driven mechanisms continuously adapt interface behavior to individual user patterns, enabling higher engagement and more efficient content delivery. Despite these advancements, concern related to usability, perceived transparence, trust, and user autonomy have become increasingly prominent, especially as AI-generated interactions often operate beyond users direct awareness [3].

Human-computer interaction research consistently shows that demographic factors such as age and digital literacy influence how users interpret and respond to AI behavior [4]. Younger users typically navigate algorithmic features with greater fluency, while older adults often report confusion or reduced trust due to unfamiliarity with AI-driven interactions [5]. Although instruments such as the System Usability Scale (SUS) and multi-dimensional Likert evaluation are widely used to asses usability, existing studies rarely focus on how users specifically experience AI features within social media environments. Furthermore, many prior studies remain descriptive, lacking statistical modeling that quantifies demographic differences [6]. In parallel, recent discussion in explainable AI (XAI) emphasize that transparency, interpretability, and user control are essential for fostering trust in algorithmic systems [7]. However, empirical evidence on how different age groups perceive AI explainability and usability within real-world social media platforms remains limited, creating a gap that hinders the development of demographic-aware AI design frameworks.

To address these limitations, this paper proposes a quantitative modeling approach that examines how age influences AI usability, trust, and comfort on social media. Using a dataset of 100 participants across three age categories, the analysis integrates SUS scores, Likert-scale assessments, and demographic segmentation. One-way ANOVA and Tukey's post hoc test are employed to determine whether differences across age groups are statistically significant and to model the degree of demographic variation in AI-related user experience [8]. The findings of this study provide an empirical foundation for understanding how demographic factors shape interactions with AI-driven systems. By identifying which users groups face the greatest usability and trust challenges, the research offers technical implications for designing adaptive and explainable interfaces that support varying levels of digital literacy. Ultimately, this work contributes to a more human-centered AI design approach by demonstrating the need for demographic-sensitive usability modeling in algorithmically driven social media environments.

II. Literatur Review

A. AI-Driven Personalization in Social Media

Artificial intelligence has become the core mechanism behind content ranking and personalized recommendations systems in modern social platforms. Prior studies show that AI models such as collaborative filtering and sequence-based on behavioral patterns [1], [2]. Despite these advantages, the algorithmic processes behind personalization often remain opaque, making it difficult for users to understand why certain content appears in their feed. This lack of transparency can reduce perceived control and influence user comfort when interacting with AI-driven features.

B. Usability and Human-AI Interaction

Usability research traditionally focuses on interface design, task efficiency and user satisfaction. In AI-enhanced environments, usability expands to include how well users understand and anticipate automated system behavior, The System Usability Scale (SUS) and heuristic frameworks are

widely used to evaluate overall user experience and cognitive workload [4]. Human-AI interaction literature emphasizes that users mental model and expectations toward algorithmic behavior strongly influence perceived usability, trust, and comfort when interacting with intelligent systems [5].

C. Trust, Explainability, and Algoritmic Transparency

Trust is a critical component in human-AI interactions, particularly when systems make autonomous decisions. Explainable AI (XAI) provides methods to communicate system reasoning or recommendation logic, enabling users to understand why AI system behaves in certain ways [6]. When explainability is absent or insufficient, users may develop concerns over data usage, fairness, and prediction accuracy-especially individuals with lower digital literacy levels [7]. These concerns directly affect trust and reduce overall user comfort.

D. Age and Digital Literacy as Moderating Factors

Demographic factors play a significant role in shaping user responses to AI features. Younger users, who often have higher exposure to algorithmic systems, tend to adapt more easily to automated personalization. In contrast, older adults commonly experience difficulties understanding AI-driven interactions and express stronger privacy-related concerns [8]. These differences indicate that age and digital literacy act as moderating variables that influence usability, trust, and comfort.

E. Research Gaps in AI Usability Modeling

Although extensive work exists on usability, trust, and AI-driven personalization, several research gaps remain. First, prior studies rarely integrate usability metrics, trust, comfort, and demographic differences into a single analytical model. Second, statistical rigor is often limited, with few studies applying inferential techniques such as ANOVA, effect size measurement or structured hypothesis testing to quantify demographic variations. Third, AI research within social media contexts frequently overlooks demographic segmentation, despite its strong influence on user experience. Addressing these gaps is essential for developing adaptive and demographic-aware AI systems.

F. Conceptual Framework for the Current Study

Based on the reviews literature, this study adopts a conceptual framework in which age influences digital literacy, which subsequently affects usability perceptions (as measured by SUS scores), trust, and overall comfort when interacting with AI-driven features. Higher digital literacy is assumed to support greater usability and trust, while lower literacy amplifies challenges related to explainability and perceived transparency. This framework guides the hypothesis formulation and statistical testing applied in the subsequent sections, forming the foundation for modeling demographic differences in AI-driven social media environments.

III. Research Methodology

This paper applies to a quantitative cross-sectional design to analyze differences in usability and comfort across age groups when interacting with AI-driven social media features. The methodology consists of four components: (A) participants, (B) dataset source, (C) measurement instruments, (D) data analysis procedures, and workflow summary.

A. Participants

A total of 100 respondents were recruited through online distribution. To ensure balanced demographic representation, participants were categorized into:

- 18-24 years (younger adults)
- 30-49 years (middle-aged adults)
- 50+ years (older adults)

All participants were active social media users with direct experience using AI-driven features such as recommendations, feeds, and filters.

TABLE I. Participant Demographics

Age Group	n	Digital Literacy Level
18-24	35	High
30-49	35	Medium-High
50+	30	Low-High
Total	**100**	-

B. Meansurement Instruments

Two validated instruments were used:

- System Usability Scale (SUS)

  Ten standard items rated on a 1-5 scale were used to assess perceived usability. Final SUS scores (0-100) were computed using the standard scoring formula.

- AI Comfort & Trust Questionnaire

  A five-point Likert scale measured comfort, trust, perceived usefulness, confusion, perceived control and privacy concerns.

Both instruments demonstrated strong reliability (Cronbach's Alpha > 0.80).

TABLE II. Participant Demographics

Variable	Instrument	Scale	Output
Usability	SUS	1-5	SUS Score (0-100)
Comfort	Likert	1-5	Mean Comfort
Trust	Likert	1-5	Mean Trust
Perceived Control	Likert	1-5	Control Score

C. Data Analysis Procedure

Data analysis was conducted through four structured stages to ensure statistical validity meaningful interpretation of usability differences across age groups:

1. Descriptive Statistics

   Summary statistics (mean, standard deviation, and frequency distribution) were generated for all SUS and Likert-scale variables to understand overall user perception patterns.

2. Assumption Testing

   Prior to inferential analysis, two statistical assumptions were evaluated:

   - Shapiro-Wilk Test to assess normality of SUS scores for each age group.

- Levene's Test to confirm homogeneity of variance across groups.

These tests determine whether the dataset meets the requirements for parametric analysis.

3. One-way ANOVA

A one-way Analysis of Variance (ANOVA) was applied to examine whether the mean SUS usability score differs significantly across the three age groups.

The hypothesis were formulated as follows:

$$H_0 : \mu_{18-24} = \mu_{30-49} = \mu_{50+}$$

$$H_1 : \text{At least one age group mean differs}$$

ANOVA was chosen because SUS scores represent continuous, approximately normal data divided into three independent groups. The resulting F-statistic and ρ–value determine significance.

4. Tukey's HSD Post HOC Test

When ANOVA indicates significant difference, Tukey's Honestly Significant Difference test was applied to identify which specific age groups differed. Effect size (η^2) was also computed to quantify the magnitude of demographic influence on usability.

D. Methodolgy Workflow

The methodology workflow shown below outlines each analytical stage including data acquisition, cleaning, descriptive analysis, assumption testing, ANOVA, and post hoc evaluation.

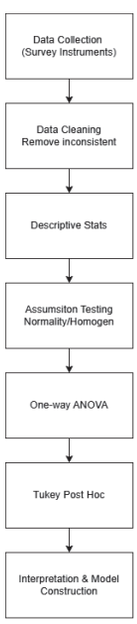


Fig. 1. Methodology Workflow

E. Data Collection

All data used in this study are openly accessible through Zenodo at: https://doi.org/10.5281/zenodo.16892473

The dataset includes:

- Anonymized responses from 100 participants,
- Raw SUS item responses (10 items, 1-5 scale),
- Raw Likert-scale responses measuring comfort, trust, perceived control, confusion, and perceived usefulness,
- Demographics attributes, including age group and primary social media platforms,
- The complete questionnaire structure and metadata used in the study.

Final SUS scores (0-100) used in analysis were derived from raw SUS item responses using the standard scoring method. This dataset supports transparency, reproducibility, and enables external validation or further extension by other researchers.

IV. RESULT

The results of this paper provide a detailed quantitative examination of how different age groups experience AI-driven interactions on social media. By combining descriptive statistics, interferential testing, and cross-groups comparisons, this section establishes clear evidence of demographic variation in usability and comfort. These results not only validate the measurement model but also reveal patterns that are critical for understanding human-AI interaction behaviors across digital generation

A. Descriptive Statistics

Descriptive statistical analysis was used to provide an initial overview of usability perceptions across age groups. The System Usability Scale (SUS) results showed a distinct variation in perceived usability of AI-driven social media features, with clear differences emerging between younger, middle-ages, and older adults. Younger adults (18-24) achieved the highest mean SUS score (82.4), indicating excellent usability and minimal friction when interacting with AI-based interfaces. This group demonstrated consistently high ratings across most SUS items, reflecting strong digital fluency and familiarity with recommendation systems, generative content, and automated decision support.

Middle-aged adults (30-49) showed a moderate mean usability score (69.1), falling within the "acceptable" usability range. Their responses suggested partial ease of use but also highlighted occasional challenges, such as interpreting I-generated suggestions or navigating algorithm-driven features. Older adults (50+) reported the lowest mean SUS score (48.3), categorizes as "poor" usability. Respondents in this group frequently indicated lower confidence, reduced control, and higher cognitive effort when engaging with AI systems, a pattern aligned with prior findings in human-AI interaction literature.

TABLE III. MEAN SUS SCORE BY AGE GROUP

Age Group	**Mean SUS**
18-24	82.4
30-49	69.1
50+	48.3

B. Data Distribution and Assumption Tests

Before conducting inferential analysis, assumptions for parametric testing were examined to ensure the validity of the one-way ANOVA. The normality of SUS scores within each age group was evaluated using the Shapiro-Wilk Test. Results indicates that all groups showed ρ-values above 0.05,

confirming that their score distributions did not significantly deviate from normality and could therefore be treated as approximately normally distributed.

Homogeneity of variance was assessed using Levene's Test, which also produces a ρ-value greater than 0.05, indicating that the variance across the three age groups was statistically equivalent. This fulfills the second major assumption required for ANOVA and confirms that group comparisons can be conducted without violating variance consistency. Together, the results of these assumptions test validate the use of parametric procedures and support the reliability of subsequent ANOVA findings by ensuring that the dataset meets essential statistical criteria for unbiased group comparison.

C. SUS Comparison Across Age Groups

Figure 2 presents a clear comparison of usability perceptions across the three age groups. Younger adults (18-24) achieved the highest SUS mean score (82.4), indicating strong perceived usability and minimal friction when interacting with AI-driven social media features. Middle-aged users (30-49) showed moderately high usability ratings (69.1), reflecting more balanced experiences that combine familiarity with occasional challenges.

In contrast, older adults (50+) reported substantial lower usability (48.3), suggesting difficulties in navigating AI features, lower confidence, and increased cognitive effort during interactions. The declining pattern across age groups demonstrates that usability is strongly influenced by age-related differences in digital familiarity and cognitive processing.

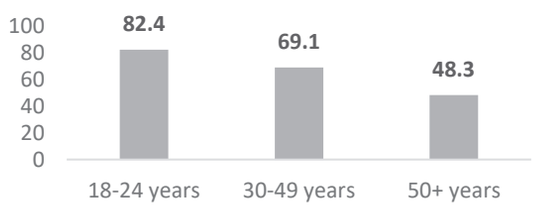


Fig. 2. SUS Score Comparison by Age Group

D. One-Way ANOVA Results

A one-way ANOVA test was conducted to determine whether differences in usability scores across the three age groups were statistically significant. The analysis revealed a meaningful variation in SUS scores between groups. Table IV summarizes the ANOVA computation. The between-group variance (SS = 9278.1) is substantially larger than what would be expected by chance, producing an F-statistic of 6.57 with a corresponding ρ–value of 0.011.

TABLE IV. ONE-WAY ANOVA SUMMARY

Source	SS	df	MS	F	ρ–value
Between Groups	92781.1	2	4639.05	6.57	0.011
Within Groups	66926.4	97	689.97	-	-
Total	761204.5	99	-	-	-

Given that resulting ρ–value is less than the significance threshold of 0.05, the null hypothesis is rejected. This confirms that usability scores differ significantly across age groups. In other words, age has a measurable statistical effect on perceived usability of AI-driven social media features.

E. Post Hoc Tukey Test

Since the ANOVA test indicated a significant difference in usability across age groups, a Tukey's HSD post hoc analysis was performed to identify which specific groups differed from one another. The results are shown in Table V.

TABLE V. TUKEYS HSD PAIRWISE COMPARISON

Comparison	Mean Diff.	ρ–value	Interpretation
18-24 vs 30-49	13.3	0.084	Not significant
18-23 vs 50+	34.1	0.009	Significant difference
30-49 vs 50+	20.8	0.052	Borderline / not significant

The Tukey analysis reveals that largest and only statistically significant difference exists between the 18-24 and 50+ age groups. This confirms that younger adults perceive AI-based social media features as substantially more usable compared to older adults. Meanwhile, differences between the middle-aged group (30-49) and the other two groups do not reach statistically significance, although the trend still shows declining usability with age. These findings support the conclusion that age-related usability gaps are most pronounced between the youngest and oldest users, reinforcing the need for age adaptive and inclusive AI interface design.

F. Efferct Size (η^2)

To assess the magnitude of age-related differences in usability, the eta-squared (η^2) effect size was calculated. Eta-squared quantifies the proportion of variance in the dependent variable (SUS score) that can be explained by group membership (age category). It is computed using the following formula:

$$\eta^2 = \frac{SSbetween}{SStotal}$$

Substituting the ANOVA values obtained in this study:

$$\eta^2 = \frac{9278.1}{76104.5} = 0.123$$

An effect of $\eta^2 = 0.123$ indicates a moderate effect, meaning that age accounts for approximately 12.3% of the total variance in usability perceptions. The demonstrates of a statistically significant ANOVA result and a moderate effect size demonstrates that age is a meaningful determinant of perceived usability in AI-driven social media features.

G. Likert-Scale Findings: Comfort, Trust, and Control

In addition to usability, participants evaluated their subjective experience with AI-driven features through a set od Likert-scale items assessing comfort, trust, perceived usefulness, confusion, and data privacy concerns. Analysis of the mean values showed consistent patterns across the sample. Participants generally perceived AI as beneficial for improving personalization, with a mean score of 4.1 on the statement regarding relevance enhancement. Trust in AI features also showed a relatively positive tendency (M =3.9), indicating that users are willing to rely on algorithmic recommendations when they perceive tangible benefits. Similarly, AI supported features were viewed as improving overall social media effectiveness (M = 3.9).

However, despite these positive aspects, participant expressed notable reservations. Data privacy concerns reached a mean score of 4.0, reflecting substantial

apprehension about how personal information is collected by AI systems. Difficulties in understanding AI operations were also apparent, with a mean score of 3.6 on the item related to conceptual complexity. Additionally, users reported moderate levels of confusion when presented with AI-generated recommendations (M = 3.8), suggesting that algorithmic decisions are not always transparent or intuitive. The highest mean score in the dataset (M = 4.5) corresponded to the desire for full control over AI features. This indicates a strong preference for transparency and user agency, demonstrating that user value the ability to adjust, modify, or disable AI-driven components when necessary. These findings complement the usability results by emphasizing that ease of use alone is insufficient; effective AI design must also address concerns related to trust privacy, and explainability.

TABLE VI. MEAN LIKERT SCORES

Statement	Mean
Comfort using AI	3.8
Personalization benefit	4.1
Difficulty understanding AI	3.6
AI improves effectiveness	3.9
Data privacy concern	4.0
Trust in AI features	3.9
Enjoyment with AI	4.0
Confusion with recommendations	3.8
Content relevance	3.9
Desire for full control	4.5

H. Interpretation of Results

The results collectively demonstrate that age plays a meaningful role in shaping how users perceive and interact with AI-driven features on social media. Younger adults consistently reported high usability, strong personalization benefits, and relatively low levels od confusions or privacy concern. This pattern suggests greater algorithmic familiarity and digital fluency, enabling smoother and more confident interaction with AI-based systems. In contrast, older adults (50+) displayed significant lower usability scores, higher cognitive friction, and elevated privacy concerns. The ANOVA and Tukey test confirmed that this group differs most sharply from younger uses, indicating that age-related factors such as lower digital literacy, reduces exposure to AI technologies, and greater sensitivity to opaque data practices contribute to diminished experience comfort. The moderate effect size further reinforces that demographic variables meaningfully influence usability perceptions.

When combined with the Likert results, a clear implication emerges: while AI features successfully enhance personalization and efficiency, they are not uniformly accessible or intuitive across age groups. Users, particularly older adults, desire clearer explanations, greater transparency, and more flexible control over AI behavior. These findings highlight the need for AI interfaces designs that accommodate diverse user capabilities and preferences, ensuring that algorithmic systems remain inclusive, comprehensible, and user-centered.

V. CONCLUSION

This study examined how users from different age groups perceive the usability and comfort of AI-driven features on social media platforms. Using The System Usability Scale (SUS) and Likert-based comfort and trust measures, combined with inferential testing through one-way ANOVA and Tukey's HSD, the findings confirm that age is aa significant determinant of usability perception. Younger adults (18-24) reported the highest usability levels, while older adults (50+) consistently showed lower usability scores, higher confusion, and greatest privacy-related concerns. The effect size analysis further supports that demographic differences, particularly age and digital familiarity, meaningfully influence how AI interfaces are experienced. These results collectively indicate that current AI-driven social media interfaces are not uniformly accessible and tend to disadvantage older users.

The implications of this study emphasize the need for age-responsive and explainable AI design. Developers should integrate adaptive interface mechanisms that adjust complexity based on user capability, alongside clear explanation of why and how AI-generated recommendations are produced. Applying explainable AI (XAI) principles and user-control mechanisms may reduce confusion and improve trust, particularly for users with lower digital literacy. Strengthening transparency and agency within AI interactions is essential tor improving the overall user experience across diverse demographic segments. However, the study is not without limitations. It relies on cross-sectional self-reported data, which may introduce subjective bias, and does not include behavioral metrics such as interaction logs, feature-use frequency, or device-level factors that may influence usability perception. The demographic grouping, while balanced, may also not capture broader cultural or socioeconomic variations that affect interaction with AI systems.

Future research should incorporate real-time behavioral data, experimental protype testing, and longitudinal evaluation to capture deeper patterns in human-AI interaction. Building and validating an adaptive UI model or explainable recommendation interface could provide more robust insights into how usability can be optimized across age groups. Integrating cognitive measures or AI literacy indicators may also strengthen the predictive accuracy of future models. Such extensions would support the development of more inclusive and human-centered AI systems aligned with the goals of green and human information technology.

ACKNOWLEDGMENT

We extend our sincere gratitude to all participants who contributed valuable insights through their responses during the data collection phase. Their perspectives were essential in understanding how users perceive AI-driven features on social media platforms. Special appreciation goes to the research team: Stiven Hizkia Hardana, who was responsible for the design, execution, and writing of this research paper, and Ms. Nuril Kusumawardani Soeprapto Putri, who served as the academic supervisor and played a central role in guiding the development and writing of this paper. Her expertise, detailed feedback, and continuous support were instrumental in shaping the structure and content.

Additionally, I acknowledge the support of OpenAI, which assisted in generating initial ideas developing the structure, and refining the writing process. The Consensus

platform was also used to identify relevant academic references, enhancing the quality of the literature review and contextual background. Tools like Grammarly were employed to improve language clarity and ensure grammatical accuracy the writing process.

REFERENCES

[1] J. Brooke, "SUS: A quick and dirty usability scale, "in usability Evalutaion in Industry, P. W. Jordan, B. Thomas, B. A. Weerdmeester, and A. L. McCelland, Eds. London, U.K.: Taylor & Francis, 1996 pp. 189-194.

[2] J. Sauro and J. R. Lewis, Quantifying the User Experience: Practical Statistics for User Research, 2nd ed. Cambridge, MA, USA: Morgan Kaufmann, 2016.

[3] R. L. Bangora, P. T. Kortum, and J. T. Miller, "Determining what individual SUS scores mean," J. Usability Stud., vol. 4, no. 3, pp. 114–123, 2009.

[4] A. Shourmasti, S. Askarany, and M. Sadeghi, "Adoption barriers of artificial intelligence among older adults," Technol. Soc., vol. 70, p. 102025, 2022.

[5] S. Park, S. Chung, and H. Shin, "Understanding user perception and trust in AI-based recommendations," Int. J. Hum.-Comput. Interact., vol. 38, no. 12, pp. 1125–1138, 2022.

[6] ISO 9241-210: Ergonomics of Human-System Interaction — Human-Centered Design for Interactive Systems. Geneva, Switzerland: International Organization for Standardization, 2019.

[7] H. Abdullah, A. Mousa, and K. Choi, "Explainable artificial intelligence for improving user trust in social media analytics," IEEE Access, vol. 10, pp. 88424–88439, 2022.

[8] J. Nielsen, Usability Engineering. San Francisco, CA, USA: Morgan Kaufmann, 1994.

[9] M. Heer and S. Bostrom, "Understanding user mental models in AI-powered interfaces," ACM Trans. Comput.-Hum. Interact., vol. 29, no. 4, pp. 1–32, 2022.

[10] R. S. Jensen, A. H. Thompson, and M. R. Jones, "Assessing usability and trust in algorithmic systems: A user-centered approach," IEEE Trans. Hum.-Mach. Syst., vol. 52, no. 2, pp. 145–156, 2022.

[11] N. Tintarev and J. Masthoff, "Explaining recommendations: Design and evaluation," in Proc. ACM RecSys, 2007, pp. 1–8.

[12] A. Knijnenburg, M. C. Willemsen, Z. Gantner, H. Soncu, and C. Newell, "Explaining the user experience in recommender systems," User Model. User-Adapt. Interact., vol. 22, no. 4–5, pp. 441–504, 2012.

[13] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv:1702.08608, 2017.

[14] R. Pak and A. McLaughlin, Designing for Aging: Human Factors Considerations. Boca Raton, FL, USA: CRC Press, 2010.

[15] N. Tintarev and J. Masthoff, "Evaluating the effectiveness of explanations for recommender systems," User Model. User-Adapt. Interact., vol. 22, pp. 399–439, 2015.

Authentic Dataset Development and Exploration of Cross-Language Transfer Learning for Aksara Jawa Recognition

1st Bryan Myer Setiawan
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
bryan.setiawan@binus.ac.id

2nd Aura Kristian Sumowidjojo
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
aura.sumowidjojo@binus.ac.id

3rd Andry Chowanda
Computer Science Department
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
achowanda@binus.edu

***Abstract*—This research is done to address two major problems in *Aksara Jawa* (*hanacaraka*): the lack of authentic handwritten data and the challenge of achieving strong performance in the character recognition model. The solution introduced in this paper includes a new publicly-available dataset of 3,528 authentic handwritten characters that captures natural variability, which cannot be found in currently-available datasets. Three dataset configurations (Kaggle-only, Self-Developed, Combined) are evaluated on three CNN architectures (EfficientNet-B0, EfficientNet-B0 with self-attention layer, and ResNet50). This method aims to investigate the efficacy of cross-language transfer learning from ImageNet, Arabic, and Devanagari pre-training. The results are further validated across three random seeds, demonstrate two primary findings: (1) Models trained with the authentic handwritten dataset consistently outperformed those trained on cleaner, public data in 7 out of 9 experimental conditions. (2) Although ImageNet remains a robust baseline, cross-language transfer learning from the structurally related Devanagari script was essential to achieve the study's highest average F1-score of 98.48% with the EfficientNet-B0-SA model. These findings confirm that a dual strategy of authentic handwritten dataset composition and related-script pre-training is critical for advancing low-resource script recognition. Future work will expand the dataset to include *sandhangan* and *pasangan* characters and deploy lightweight models in accessible applications to support the digital preservation of Indonesia's cultural heritage.**

***Keywords*—Aksara Jawa Character Recognition, Authentic Handwritten Characters, Convolutional Neural Network, Cross-Language Transfer Learning, Low-Resource Script Recognition**

I. INTRODUCTION

Aksara Jawa (*Hanacaraka*) is a traditional Javanese script that was first introduced in the 17th century and is considered an important part of Indonesia's written cultural heritage [1]. As part of the preservation effort, the Indonesian government has utilized *Aksara Jawa* as one of the media to show Javanese culture in many sites, such as street signs, museums, and other traditional buildings around Java [2]. In addition, regional governments have initiated revitalization efforts in the education sector, such as reintroducing *Aksara Jawa* in schools such as SMA Karangturi in Central Java, to prevent the script's literacy decline and spark interest among younger generations. Although most of the students in the area speak Javanese fluently, a recent study highlighted that there is still a major percentage of them who are still unable to read or write the script, causing *Aksara Jawa* to be categorized as endangered [3].

Researchers have explored various technology approaches as a response to problems, most of them studies in convolutional neural networks (CNN) to recognize Javanese characters in both natural scenes and isolated script images [3] [4] [5]. However, the limited availability of authentic handwritten Aksara Jawa images, which often consists of about 2000-4000 samples [5] [6] [7], has led many studies to adopt the transfer learning method to improve performance models [7] [8]. Other attempts to mitigate data scarcity have included merging in printed Javanese characters with augmented data [9]. Although these approaches help will increase data volume, the AI models that is trained on these data may fail to capture the variability in handwritten text,

potentially leading to poor generalization and limited applicability in real-world scenarios [10] [11].

However, researches on other scripts recognition like Arabic and Devanagari has achieved high accuracy (up to 0.95) with robust generalization capability, this result is achieved through training on large-scale, high-quality, and well-balanced datasets [12] [13] [14] [15]. Furthermore, the cross-language transfer learning method shown its effectiveness in fields like Natural Language Processing and Speech Recognition, and has shown its potential in character recognition, such as with Kanji recognition benefiting from pre-training on Chinese characters [16]. Despite these successes, exploration towards cross-language transfer learning for *Aksara Jawa* character recognition remains limited.

This research aims to support the revitalization and digital preservation of Aksara Jawa with two main objectives: developing a comprehensive authentic handwritten *Aksara Jawa* character dataset and proposing a robust recognition model utilizing the cross-language transfer learning method. Unlike previous studies that relied on general pre-trained weights and/or non-handwritten script data, this research collects the authentic handwritten samples from high school students in Central Java to better capture real-world variability, and support future research. This study also investigates the effectiveness of transfer learning by taking advantage of models pre-trained with other non-Latin scripts—specifically Arabic and Devanagari which are selected for their visual similarities to Javanese script [15]. The resemblances include curvilinear structures, diacritics, and complex ligatures. These models are fine-tuned and evaluated on the newly compiled Aksara Jawa dataset to assess the impact of cross-language transfer learning on handwritten character recognition and dataset development.

II. LITERATURE REVIEW

A. Aksara Jawa

Aksara Jawa (*Hanacaraka* or *Carakan*) is a traditional Indonesian script originating from the island of Java. Dating back to the Mataram Kingdom era in the 17th century, it is a derivative of the Brahmi and Kawi scripts from India [1]. This script was made through the philosophy and Javanese legend of Ajisaka, in which the *Hanacaraka* script letters are arranged to have a noble story of Ajisaka's 2 adjutants that each carry Ajisaka's order. Throughout the story, several meanings can be interpreted which include a chivalrous attitude, upholding honesty, and carrying out the mandates given to us [17].

Aksara Jawa consists of 20 characters [17]:
(ha)(na)(ca)(ra)(ka) - There were messengers
(da)(ta)(sa)(wa)(la) - They argued, disagreed
(pa)(dha)(ja)(ya)(nya) - Equal was their power
(ma)(ga)(ba)(tha)(nga) - They ended up as corpes, dead.

Recent research has applied deep learning to *Aksara Jawa* recognition. Afakh *et al.* [2], developed an *Aksara Jawa* text detection system in sceneries such as buildings, street signs, and engraved stones using deep-CNN architecture. This model is trained to learn features and classify Javanese script text regions through the text confidence map and bounding box generated by CNN. This approach claimed to have a higher precision and recall value compared to previous works where this model achieved a precision score of 0.96 and a recall of 0.83. However, this result may produce some reliability concerns due to the small dataset used in the model where only 60 scene images are used for model training, and 30 images for model testing. With only 90 real images and some other synthetic image data being fed in the model, limited generalization of *Aksara Jawa* and overfitting may occur.

Arief *et al.* [5] explored classification of *Aksara Jawa* using the MobileNetV2 CNN architecture to classify the 20 classes of *Aksara Jawa* that represent each character. This research is done to classify each script character, and the dataset used for the model comprises 2644 images of all the characters with an image size of 224x224x3 pixels. This research indicates that the MobileNetV2 model is efficient for *Aksara Jawa* classification since the model reduces the number of parameters and computational requirements. However, the model accuracy and precision achieved were 0.91 and 0.89 respectively, with the authors stating that the small dataset is a limitation.

Susanto *et al.* [4] proposes the Javanese Script Recognition through a 12 layers Deep Convolutional Neural Network (DCNN): 4 Convolutional layers, 2 pooling layers, 5 Fully Connected Layers, and 1 softmax classifier. This method was designed to handle a small dataset of Javanese Script Characters and utilise data augmentation to enhance model accuracy of 0.99. But the dataset used in this research is limited, which might not fully present the variability of the dataset, and this research added the initial hypothesis of *Aksara Jawa* dataset scarcity.

Naufal *et al.* [9] combines the concept of transfer learning and CNN to transliterating *Aksara Jawa* images to roman script. The models utilized in this project include MobileNetV2, VGG-16, and Xception pre-trained

models. The transfer learning method is proven to address the overfitting issue while still achieving a high performance in classification and transliteration with MobileNetV2 achieving a 0.99 F1 score. However, the decline of model performance does occur when a mixed dataset of printed and handwritten images is implemented in the VGG-16 model where the F1 score falls from 0.98 to 0.43.

B. Arabic Script

Arabic language is a widely used Semitic language with the script being used by roughly 300 million speakers around the world. The Arabic alphabet consists of 28 characters and due to many characters having a similar shape, several research have been done to produce an advanced Arabic handwriting recognition model, and most of them utilizes a CNN-based model to produce a high accuracy model [12] [18] [19]. A study by Altwajiry and Al-Turaiki. [12] developed automatic handwriting recognition for Arabic language based on CNN architecture. The CNN model is trained on the Hijja dataset which contains 47,434 characters and Arabic Handwritten Character Dataset (AHCD). This study compared handwriting recognition models such as CNN, CNN-BLTSM, and MLP throughout several datasets with various dataset sizes ranging from 3000 to 96,000 images. Through this study, the CNN model applied to CMATERDB dataset has the highest accuracy model with 0.99.

Another common approach in Arabic script recognition is the use of deep learning models. Deep learning is a suitable approach to overcome the cursive Arabic shape recognition challenge, since deep learning models can demonstrate a significant success in recognizing various pattern recognition and classification tasks [20] [21] [22]. One research that proves the deep learning significance was done by Ahmad *et al.* where this research displays a state-of-the-art result for Arabic handwriting model recognition on the KHATT (KFUPM Handwritten Arabic Text) dataset which consists of 2000 unique paragraphs handwritten by 1000 different writers. This model utilizes an MDLSTM-based classifier and results in a high system performance with a 0.75 recognition rate which is higher than previous studies with a margin of 0.29.

C. Devanagari Script

Devanagari Script is a traditional and rare script can be found in the country of India and many other Asian countries, and many of this script documents are degraded due to age and other document writing factors. Devanagari is used in several other languages including Hindi, Sanskrit, Nepali, and Marathi languages, consisting of 11 vowels and 33 consonants basic characters [14]. Even though *Aksara Jawa,* Arabic script, and Devanagari script originate from different linguistic and cultural traditions, these three scripts share some commonalities where all of them have curvilinear and rounded forms, ligatures and connected forms, and diacritical marks [1] [12] [14].

Aneja and Aneja [15] proposed a transfer learning scheme for Devanagari script which is more efficient and may be more practically feasible in a small dataset approach. This approach utilizes 2 transfer learning techniques, namely finetuning and fixed feature extractors. Finetuning is done by initializing the network with a pretrained network making it more efficient, while fixed feature extractor is the approach in which the convolutional layer's weight stays the same and only the fully connected layers are trained and updated. This approach uses several deep learning models with the InceptionV3 model results in the best performance of 0.99 accuracy in 15 epochs. The transfer learning approach can result in good performance but may need a long time in the training process [15].

III. METHODOLOGY

This study evaluates the effectiveness of cross-language transfer learning and the contribution of authentic handwritten dataset development by testing three convolutional neural network (CNN) architectures, each representing a different level of model complexity and robustness.

A. Dataset Development and Preparation

The final dataset consists of 7724 images, combining clean handwritten characters from public Kaggle datasets (Aksara Jawa v3 and Hanacaraka) with authentic handwritten samples collected from SMA Karangturi, Semarang. These handwritten samples were manually standardized by correcting orientation, converting them to PNG format and applying a preprocessing pipeline including shadow removal, binarization with Otsu's thresholding, morphological closing to repair stroke gaps, and padding to a square aspect ratio. To ensure a generalized evaluation, each dataset was divided with a ratio of 70:10:20, training, validation, and test set, respectively.

To evaluate the impact of the self-developed dataset, three dataset scenarios were constructed, which are Kaggle-only (Aksara Jawa v3 and Hanacaraka), Raw-

only (preprocessed images from SMA Karangturi), and Combined(Kaggle-only and Raw-only).

B. Image Preprocessing and Augmentation

Before training, all images were resized to 224x224 pixels, normalized with the ImageNet mean and standard deviation values, and preprocessed to ensure compatibility with all models. To achieve better generalization on the results, several augmentation techniques were applied, such as RandomRotation to simulate slanted handwriting, RandomAffine to simulate various phototaking angles, ColorJitter to reflect different lighting and contrast, and GaussianBlur to simulate image blur.

C. Model Architectures

Three CNN architectures were selected to represent different levels of complexity, EfficientNet-B0 as a lightweight model, EfficientNet-B0 with Single layer Self Attention as a hybrid(slightly robust) model, and ResNet50 as a robust model. Each model was adapted to predict 20 output classes relative to the number of characters we want to classify in *Aksara Jawa*. The final classifier layers were replaced with a Dropout layer followed by a Linear layer to mitigate overfitting. Additionally, all base models were imported from the timm library to ensure consistency and reproducibility.

D. Training Setup and Transfer Learning Strategy

To investigate the impact of cross-language transfer learning for *Aksara Jawa* recognition, each model was trained under three different pre-trained weight strategies: ImageNet, Devanagari, and Arabic. ImageNet provided general visual feature extraction, while Devanagari and Arabic were selected for their structural similarities to Javanese, such as stroke patterns, component shapes and cursive nature, respectively. In total, 27 experimental runs were executed covering every combination of model, dataset, and pre-trained weights.

All training was conducted on the Kaggle platform using PyTorch framework with NVIDIA T4 GPU. To ensure reproducibility, three fixed random seed values were selected, which are 42, 43, and 44. The models were trained using AdamW optimizer with an initial learning rate of $1 \cdot 10^{-3}$, weight decay of $1 \cdot 10^{-5}$ and cross-entropy loss with 0.1 label smoothing to enhance generalization. Finally, ReduceLROnPlateau scheduler (patience: 3) and early stopping (patience: 10) were applied to manage convergence and prevent overfitting.

E. Model Evaluation

Model performance was assessed with a set of evaluation metrics including accuracy, macro-averaged F1-score, precision and recall towards the test set. Inference time per image was also measured to evaluate trade-offs between model complexity and real-world application viability. Furthermore, the contribution of this paper are analyzed through two perspective: first, by contrasting the dataset compositions to demonstrate the impact of authentic handwritten data, and second, by comparing different pretrained weights to validate the impact of cross-language transfer learning for *Aksara Jawa* recognition as a low-resourced script.

IV. RESULTS

TABLE I
F1-SCORE RESULTS OF EFFICIENTNET-B0 MODELS

EfficientNet-B0			
	ImageNet	Arabic	Devanagari
Self-Developed	97.76%	96.90%	96.90%
Kaggle	97.30%	95.89%	95.91%
Combined	98.04%	96.77%	96.76%

TABLE II
F1-SCORE RESULTS OF EFFICIENTNET-B0-SA MODELS

EfficientNet-B0-SA			
	ImageNet	Arabic	Devanagari
Self-Developed	98.26%	98.34%	98.84%
Kaggle	97.20%	97.02%	97.08%
Combined	98.14%	97.71%	98.06%

TABLE III
F1-SCORE RESULTS OF RESNET50 MODELS

ResNet50			
	ImageNet	Arabic	Devanagari
Self-Developed	96.71%	97.51%	97.55%
Kaggle	97.65%	95.88%	97.13%
Combined	97.82%	97.36%	97.43%

The results confirms that authentic handwritten dataset are essential to achieve a robust model for recognition, and that cross-language transfer learning from related script offers tangible performance advantages. The results validate the initial hypothesis: (1) that authentic handwritten data is critical for robust Aksara Jawa recognition, and (2) that cross-language transfer learning from related scripts offers a tangible performance advantage. These findings were validated on three distinct CNN

architectures, from lightweight to robust, to ensure generalizability. The tables below present the average F1-scores achieved on the test set across these runs.

A. Impact of Authentic Dataset Composition of Model Performance

A key contribution of this research is the development of the authentic handwritten dataset, which contains of 3528 authentic handwritten images. This was specifically created to address the lack of public realistic training data, limiting the development of a robust *Aksara Jawa* recognition model from variability inherent in human handwriting. The results in Tables I, II, and III, clearly show that the using datasets with a composition of authentic data had a bigger impact towards model performance.

The self-developed dataset yielded the highest F1-score in 7 of the 9 models setup/pre-training conditions. The remaining 2 top scores were produced with the combined dataset, which leveraged the authentic data as well. Conversely, the Kaggle-only dataset, representing cleaner and less varied data, did not produce the highest scores in any configuration and appeared to be consistently resulting in the lowest average F1 scores across all models. These 9 wins from datasets with authentic data composition strongly suggest that the inclusion of authentic handwritten data is a critical factor for achieving the best performance and generalization.

The single highest F1 score of the study (98.47%) was achieved by the EfficientNet-B0-SA model, pre-trained on Devanagari and trained on the Self-Developed dataset. This result is particularly noteworthy as it represents the primary contribution of this research, combining the most relevant pre-training strategy with the self-developed dataset. This peak performance validates the hypothesis that this combined approach is highly effective for low-resource script recognition, specifically on Aksara Jawa.

B. Efficacy of Cross-Language Transfer Learning

Comparing results from pre-training strategies, both ImageNet and Devanagari proved to be highly effective bases for transfer learning. While ImageNet achieved a slightly higher average F1 score across all tests (97.65% vs 97.26% for Devanagari), Devanagari pre-training achieved the study's peak performance. Specifically, using the EfficientNet-B0-SA model, Devanagari weights outperformed ImageNet by a small margin (98.48% vs 98.26%). This suggests that ImageNet provides robust general features, learning from a structurally related script like Devanagari offers a specialized advantage for maximizing performance in complex models.

C. Model Architecture and Model Analysis

The results shows a clear correlation between model complexity/architecture and their performance. The hybrid EfficientNet-B0-SA model achieved the highest peak F1-score of 98.48%, successfully outperforming the lightweight EfficientNet-B0 (98.04%) and the robust ResNet50 (97.82%). This suggests that the self-attention mechanism provided a tangible benefit for this specific task by allowing the model to capture fine-grained details in authentic handwritten characters.

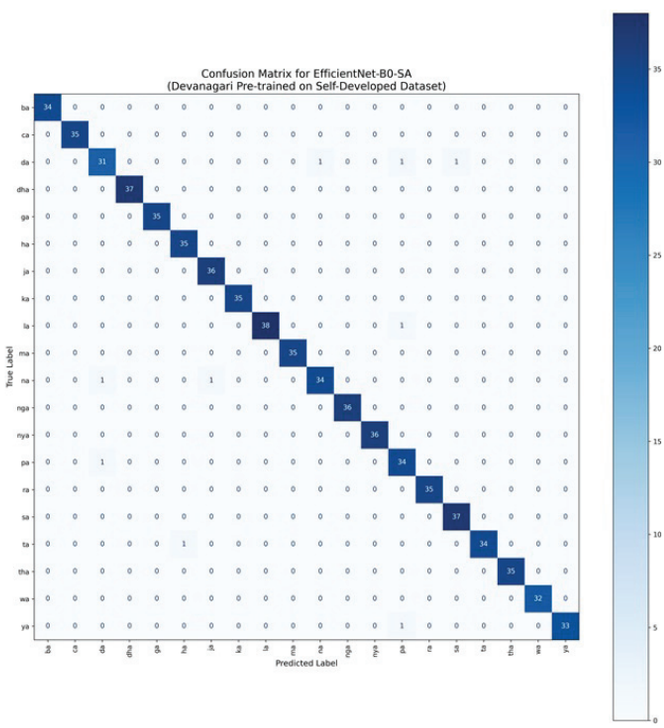


Fig. 1. *Confusion Matrix of EfficientNet-B0-SA (Devanagari)*

Furthermore, the confusion matrix in Figure 1 validates these findings, showing the exceptional performance of this model with a high level of precision across all classes. Despite the performance, the model did struggle slightly to specific characters like 'da' with 'na', 'sa' with 'pa'. These misclassifications are logical as these characters share similar base shapes and stroke patterns. This analysis suggests that the model is robust, but highlights that future improvements can be achieved by targeting these visually similar pairs.

V. DISCUSSION

A. The Critical Role of Authentic Data

The results indicate that existing public datasets (represented from Kaggle-only dataset) are insufficient to help achieve robust Aksara Jawa recognition models. The underperformance of Kaggle-only dataset suggests that the high cleanliness and uniformity does not prepare models for the variations found in human handwriting. This underscores the need for more diverse and authentic datasets to achieve higher model performance.

The superior performance of the Self-Developed and Combined datasets validates this paper's first hypothesis. The authentic noise, inconsistent strokes, and stylistic variations present in the handwritten samples act as a natural and highly effective form of data augmentation. This "realistic" data is crucial for model generalization, as evidenced by the Self-Developed set producing the highest F1 scores in 7 of 9 conditions. This finding suggests that for low-resource endangered scripts, the development of even a moderately sized authentic dataset is more impactful than relying on larger, synthetic or clean datasets.

B. Devanagari as an Effective Source for Transfer Learning

The second key finding is one specialized success of cross-language transfer learning from Devanagari scripts. As can be seen in Table II, the Devanagari-pretrained EfficientNet-B0-SA model achieved the peak F1 score of 98.47%, outperforming ImageNet-pretrained counterpart of 98.26%. This suggests that pre-training models on a script with shared visual similarities like curvilinear structures, diacritics, and complex ligatures provide a more relevant feature foundation for maximizing performance in specialized architectures.

However, ImageNet remained an incredibly robust baseline across all tasks, often matching or slightly exceeding cross-lingual methods on average. Arabic pre-training generally yielded competitive but lower results than both ImageNet and Devanagari, suggesting that features learned from Semitic (Arabic) script family are less compatible with Aksara Jawa than those from the Brahmic (Devanagari) family.

C. Model Selection and Limitation

The architectural comparison, although a controlled secondary variable, served as a useful validation to this research. It confirmed that the findings of positive impact of cross-language transfer learning and authentic dataset development are not limited to a single architecture. The EfficientNet-B0-SA model's top performance suggests that the self-attention mechanism provided a tangible benefit by enabling the model to better focus on fine-grained visual features that distinguish similar characters. However, the standard EfficientNet-B0 also achieved an exceptional F1 score (98.04%) on its own, making it a powerful candidate for deployment, especially when considering its lower computational overhead compared to ResNet50 (97.82%).

Despite the successes of this research, the primary limitation of this study is its scope, which is confined to the 20 base Aksara Jawa characters. Future development from this research is mainly towards the complete Javanese character set including *sandhangan* and *pasangan*. Incorporating these elements would provide a better comprehensive foundation to handle the full complexity of *Aksara Jawa.*

VI. CONCLUSION

This research accomplishes two primary contributions to the preservation of Aksara Jawa. First, developed a publicly available dataset of 3528 authentic handwritten characters with the help of students and teachers from SMA Karangturi Semarang, and demonstrated that this high-variability dataset is superior compared to existing clean datasets for developing a robust recognition model. Second, systematically validated the efficacy of cross-language transfer learning, proving that pre-training on a visually and structurally related script, Devanagari, yields the highest peak recognition performance (up to 98.48% F1 score), surpassing the standard ImageNet baseline in the same configuration. These findings confirm that the dual strategy of developing authentic datasets and leveraging related-script pre-training is critical for enhancing the performance of low-resource character recognition. Future work should focus on expanding this dataset to include the complete script, such as other vowels (a, i, u, e, o), *pasangan,* and *sandhangan.*

REFERENCES

[1] A. Arismadhani, U. L. Y. As'ad, and I. Kuswardayan, "Aplikasi Belajar Menulis Aksara Jawa Menggunakan Android," Institut Sepuluh November, Surabaya, 2013.

[2] M. L. Afakh, A. Risnumawan, M. E. Anggraeni, M. N. Tamara, and E. S. Ningrum, "Aksara Jawa text detection in scene images using convolutional neural network," in *Proc. 2017 Int. Electron. Symp. Knowl. Creation Intell. Comput. (IES-KCIC)*, pp. 77–82, IEEE, Sep. 2017.

[3] A. Karimah, "Aksara Jawa: Do the Javanese Students can Read and Write It?," *Tradition and Modernity of Humanity*, vol. 2, no. 2, pp. 28–32, 2022.

[4] A. Susanto, I. U. W. Mulyono, C. A. Sari, E. H. Rachmawanto, D. R. I. M. S. Setiadi, and K. Sarker, "Handwritten Javanese script recognition method based 12-layers deep convolutional neural network and data augmentation," *Int. J. Artif. Intell.*, vol. 2252, no. 8938, pp. 1449, 2023.

[5] D. Septhian, N. Syafi'al, and W. Anggara, "Implementasi CNN Arsitektur MobileNetV2 Untuk Klasifikasi Tulisan Aksara Jawa," in *Seminar Nasional Teknologi & Sains*, vol. 3, no. 1, pp. 298–303, Jan. 2024.

[6] W. P. W. Leksana and S. T. F. Suryawan, "Javanese Character Recognition With Real-time Detection Using Convolutional Neural Network," Doctoral dissertation, Universitas Muhammadiyah Surakarta, 2023.

[7] A. Susanto, C. A. Sari, E. H. Rachmawanto, I. U. W. Mulyono, and N. M. Yaacob, "A Comparative Study of Javanese Script Classification with GoogleNet, DenseNet, ResNet, VGG16 and VGG19," *Scientific Journal of Informatics*, vol. 11, no. 1, pp. 31–40, 2024.

[8] M. Huh, P. Agrawal, and A. A. Efros, "What makes ImageNet good for transfer learning?," *arXiv preprint arXiv:1608.08614*, 2016.

[9] M. F. Naufal, J. Siswantoro, and J. T. Soebroto, "Transliterating Javanese Script Images to Roman Script using Convolutional Neural Network with Transfer Learning," *Int. J. Informatics Visualization*, vol. 8, no. 3, pp. 1460–1468, 2024.

[10] J. Jordon, L. Szpruch, F. Houssiau, M. Bottarelli, G. Cherubin, C. Maple, *et al.*, "Synthetic Data—what, why and how?," *arXiv preprint arXiv:2205.03257*, 2022.

[11] F. T. Anggraeny, Y. V. Via, and R. Mumpuni, "Image preprocessing analysis in handwritten Javanese character recognition," *Bull. Electr. Eng. Informatics*, vol. 12, no. 2, pp. 860–867, 2023.

[12] N. Altwaijry and I. Al-Turaiki, "Arabic handwriting recognition system using convolutional neural network," *Neural Comput. Appl.*, vol. 33, no. 7, pp. 2249–2261, 2021.

[13] N. Wagaa, H. Kallel, and N. Mellouli, "Improved Arabic alphabet characters classification using convolutional neural networks (CNN)," *Comput. Intell. Neurosci.*, vol. 2022, no. 1, p. 9965426, 2022.

[14] S. R. Narang, M. Kumar, and M. K. Jindal, "DeepNetDevanagari: a deep learning model for Devanagari ancient character recognition," *Multimedia Tools Appl.*, vol. 80, pp. 20671–20686, 2021.

[15] N. Aneja and S. Aneja, "Transfer learning using CNN for handwritten Devanagari character recognition," in *Proc. 2019 1st Int. Conf. Adv. Inf. Technol. (ICAIT)*, pp. 293–296, IEEE, July 2019.

[16] V. Rusyn, A. Boichuk, and L. Mochurad, "Cross-Language Transfer-Learning Approach via a Pretrained Preact ResNet-18 Architecture for Improving Kanji Recognition Accuracy and Enhancing a Number of Recognizable Kanji," *Appl. Sci.*, vol. 15, no. 9, p. 4894, 2025.

[17] R. Ricci, "Reading a History of Writing: heritage, religion and script change in Java," *Itinerario*, vol. 39, no. 3, pp. 419–435, 2015.

[18] A. El-Sawy, M. Loey, and H. El-Bakry, "Arabic handwritten characters recognition using convolutional neural network," *WSEAS Trans. Comput. Res.*, vol. 5, no. 1, pp. 11–19, 2017.

[19] M. Elleuch, N. Tagougui, and M. Kherallah, "A novel architecture of CNN based on SVM classifier for recognising Arabic handwritten script," *Int. J. Intell. Syst. Technol. Appl.*, vol. 15, no. 4, pp. 323–340, 2016.

[20] S. B. Ahmed, S. Naz, M. I. Razzak, and R. Yousaf, "Deep learning based isolated Arabic scene character recognition," in *Proc. 2017 1st Int. Workshop Arabic Script Anal. Recognit. (ASAR)*, pp. 46–51, IEEE, Apr. 2017.

[21] A. Bin Durayhim, A. Al-Ajlan, I. Al-Turaiki, and N. Altwaijry, "Towards accurate children's Arabic handwriting recognition via deep learning," *Appl. Sci.*, vol. 13, no. 3, p. 1692, 2023.

[22] M. Dhiaf, A. C. Rouhou, Y. Kessentini, and S. B. Salem, "MSdocTr-Lite: A lite transformer for full page multi-script handwriting recognition," *Pattern Recognit. Lett.*, vol. 169, pp. 28–34, 2023.

Smartphone-Based Melanoma Detection Using Edge Impulse and Convolutional Neural Networks

1st Nurul Vita Dwi Faraya *, 2nd Sevia Indah Purnama †, 3rd Irmayatul Hikmah ‡
* *Department of Biomedical Engineering, Telkom University, Purwokerto 53147, Jawa Tengah, Indonesia*
† *Department of Biomedical Engineering, Telkom University, Purwokerto 53147, Jawa Tengah, Indonesia*
‡ *Department of Biomedical Engineering, Telkom University, Purwokerto 53147, Jawa Tengah, Indonesia*
`1st nurulvitaa@student.telkomuniversity.ac.id, 2nd seviaindah@telkomuniversity.ac.id, 3rd irmayatulh@telkomuniversity.ac.id,

***Abstract*—** Melanoma is a dangerous type of skin cancer with a high mortality rate if not detected early. Biopsy remains the primary method of diagnosis, but the limited number of dermatologists and access to testing pose obstacles. This study aims to develop a smartphone-based melanoma detection system using a Convolutional Neural Network (CNN) model through the Edge Impulse platform to enable fast, practical, and real-time detection. A dataset of 2,585 melanoma and non-melanoma images was obtained from open sources and underwent preprocessing, including resizing, normalization, and data augmentation. The model was trained using the MobileNetV2 architecture and evaluated using accuracy, precision, recall, and F1-score metrics. The training results showed high performance with an accuracy of 95.16% and an AUC value of 0.96. After the model was deployed to mobile devices, testing using a smartphone camera achieved 100% accuracy on 20 real image samples. These results show that the system is capable of accurately classifying melanoma both in the dataset and in real-world conditions, making it a potentially accessible and efficient early detection tool.

Keywords— melanoma, CNN, Edge Impulse, smartphone, image classification.

I. INTRODUCTION

Melanoma skin cancer is one of the rarest types of skin cancer, but it is also one of the most deadly and potentially fatal. The growth and development of melanoma occurs due to DNA damage in skin cells, which is generally caused by exposure to ultraviolet radiation from sunlight [1]. Skin cancer is classified into two types based on the cells that form the epidermis layer, namely melanoma skin cancer (malignant melanoma) and non-melanoma skin cancer (consisting of basal cell carcinoma and squamous cell carcinoma [2]. In 2022, the total number of melanoma skin cancer cases worldwide was 331,722, with 58,667 deaths. The incidence of malignant melanoma skin tumors in Indonesia was 1,609 cases in 2020 [3]. Research by Wilvestra, Lestari, & Asri (2018) at Dr. Djamil Hospital in Padang reported that the incidence of new cases of malignant skin tumors was 3.7% of a total of 1,003 patients [4].

Skin cancer screening is necessary. Biopsy is mainly performed for diagnosis or to rule out skin malignancy. Before performing a skin biopsy, it is very important to make an appointment with the laboratory [5]. The number of dermatologists is limited, so it is not possible to screen the entire population for skin cancer. Time constraints, comorbidities, and patient embarrassment are reported to be the main barriers to comprehensive skin cancer screening in patients. Therefore, early detection is necessary to prevent or reduce deaths from melanoma skin cancer [6].

Various efforts have been made to solve this problem, In addition, research conducted by Mustafa Qays Hatem developed a skin lesion classification system using the K-Nearest Neighbor (KNN) algorithm. This study focused on distinguishing between normal and pathological lesions, and showed that KNN was able to provide fairly good classification performance for early detection of skin abnormalities [7]. Then, research conducted by Almeida et al. used the Random Forest and Histogram Gradient Boosting methods with texture feature extraction from melanoma images. This study achieved an accuracy of 85–99%, depending on the color channel used (red, green, blue) in detecting melanoma [8]. And research conducted by Bistroń & Piotrowski using the Random Forest method to detect melanoma from skin images. This study achieved an accuracy of 86.36% in detecting melanoma [9].

Based on previous studies that have used various methods, this study proposes the development of a melanoma skin cancer detection system using a smartphone camera and Edge Impulse. With the development of this system, it is hoped that it can help improve the early detection rate of melanoma, speed up diagnosis, and reduce mortality rates from melanoma through faster treatment.

II. RELATED WORKS

A. CNN Models for Melanoma Detection

Convolutional neural networks (CNNs) are a type of neural network specifically designed to handle image

processing problems. CNNs are an effective method for classifying, identifying, and recognizing patterns in images. CNNs are able to understand image details better because they have an architecture that matches the way the human brain processes visual information. The data used in CNNs is two-dimensional data [10]. One of the most commonly used CNN architectures is VGGNet. VGGNet is well known for its deep structure and strong feature extraction capabilities. MobileNetV2 is a neural network architecture developed by Google and specifically designed for mobile devices or systems with limited computing resources [11]. This study uses the MobileNetV2 architecture available on edge impulse.

B. *Edge Impulse for Melanoma*

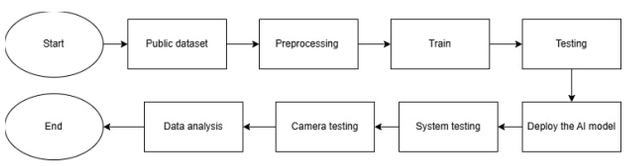
Edge Impulse is a cloud-based AI platform that provide assistance in training custom models. The workflow of this platform is structured into three primary stages: information acquisition, data processing and training model. The model can be trained by uploading data to the Edge Impulse platform and designing the impulse [12]. Edge Impulse provides a single platform for developers to collect or import data, visualize it, process it in preparation for training and testing of the machine learning model, and for final deployment onto the embedded system of choice [13]. Edge Impulse is a cloud-based platform for ML operations that enables the development of embedded and edge systems. These systems can be employed across a diverserange of hardware. Edge devices are a diverse range of electronic devices, including smartphones, tablets, laptops, and personal computer [14]. This research utilizes Edge Impulse as a platform for image classification, enabling the developed model to perform real-time melanoma detection via smartphone.

C. *Research Gaps*

Although various studies have successfully developed models for skin cancer detection, there are still several limitations that remain gaps in research. Fina Royana et al created a transfer learning-based mobile application with Inception V3 and ResNetV2, achieving 97% accuracy but still relying on an internet connection and doctor interpretation [15]. Ramdhana & Pratiwi (2023) compared InceptionV3 and EfficientNetB0 on the HAM10000 dataset, where EfficientNetB0 excelled with an accuracy of 98.09%, but the amount of data per class was limited, which could potentially affect generalization [16]. Rahadian Yusuf Abdul Rahman et al. (2024) used Derivative of Gaussian and CNN, but the testing accuracy was only 44.07%, indicating the model's sensitivity to network structure and preprocessing [17]. Fedryanto Dartiko et al. (2024) combined CNN with hybrid preprocessing and achieved an accuracy of 77.55% [18], while Jesica Trivena Sinaga et al. (2024) optimized VGG16 and obtained a validation accuracy of 90.89% [19], but both remain computationally expensive and require large datasets. Muhammad Falah Abdurrohman (2024) used GLCM with Ensemble SVM, resulting in an accuracy of 88.56% with light computation but less capable of automatic feature learning [20]. Overall, the gap in this research shows the need for the development of lightweight, efficient, and real-time executable models on mobile devices that still maintain high accuracy in detecting melanoma.

III. METHODELOGY

This research methodology describes the stages used in developing an image classification-based melanoma detection system using the Edge Impulse platform. Each stage is designed to ensure the process. The complete methodology flow is illustrated in the following flowchart to provide a comprehensive overview of the research process.



This research began with the collection of a dataset of melanoma and non-melanoma images from the Kaggle platform, totaling 2,585 images. This dataset was chosen because it had clear labels and sufficient variation in skin lesion images for the model training process. All images then underwent preprocessing to ensure optimal input quality, including resizing according to Edge Impulse model requirements, pixel normalization, and augmentation such as rotation, flipping, and brightness adjustment to increase data variation while minimizing the risk of overfitting. This stage also included class balance checking to ensure proportional data distribution.

Once the data is ready, the model training process is carried out using the Edge Impulse platform, utilizing the Convolutional Neural Network (CNN) architecture. The dataset is divided into training and validation sets, then the model is trained with the appropriate hyperparameter settings such as the number of epochs, batch size, and learning rate. During training, the model's performance was monitored through accuracy and loss graphs until the model achieved its best performance. The trained model was then tested using test data (test set) to calculate evaluation metrics such as accuracy, precision, recall, F1-score, and confusion matrix.

The best-performing model is then deployed to smartphones using Edge Impulse's deployment feature, which supports the use of models directly on mobile devices. System testing is conducted to evaluate the model's performance when run in real-time through a smartphone camera, including inference speed,

prediction consistency, and the effect of lighting conditions. This stage is followed by direct smartphone camera testing to ensure that the model is capable of detecting melanoma from real-world images, not just from the dataset.

All test results, both from the dataset and from smartphone trials, are analyzed in the final stage. The analysis includes a comparison of model performance before and after deployment, evaluation of prediction errors, and identification of factors that affect system accuracy. The study concludes with conclusions regarding the effectiveness of this smartphone-based approach and recommendations for further development.

A. *Software System*

- The dataset used in this study was taken from open sources such as Kaggle. The amount of data used was 2,585, including augmentation. (insert Kaggle link). The data was categorized into two main classes (melanoma and non-melanoma). This process also included preprocessing, such as resizing, normalization, and data augmentation. In terms of software, the melanoma detection system was developed using the Edge Impulse platform.
- Edge Impulse assists with every step in the edge AI lifecycle, from data collection, feature extraction, machine learning (ML) model design, model training and testing, to model deployment on end devices. Once the model achieves optimal accuracy, it is exported and deployed via a web-based interface so that it can be run using a smartphone camera to make real-time predictions.
- Testing data is used to evaluate the model's performance in classifying melanoma and non-melanoma images. The testing dataset is taken from a portion of the data that has been separated from the main dataset using the train–test split method, which is 20% of all images. All images undergo preprocessing before being used in testing, including resizing to fit the model input size in Edge Impulse (96×96 pixels) and pixel normalization. This testing dataset is then used to calculate performance metrics such as accuracy, precision, recall, and confusion matrix to determine the model's success rate in detecting melanoma.
- Evaluation Metrics
  During the model training process, Edge Impulse Studio will calculate accuracy by calculating each model training result using a Confusion Matrix. This matrix will validate each classified object with other objects and the object background. The lower the error, the higher the accuracy, and vice versa (use citation). The confusion matrix is used to see the number of correct and incorrect predictions, while accuracy measures the percentage of correct predictions from all testing data. Precision shows the proportion of correct melanoma predictions, recall measures the model's ability to detect all melanoma cases, and F1-Score provides an overview of overall performance by considering precision and recall.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \times 100\%$$

$$\text{Precicion} = \frac{TP}{TP+FP} \times 100\%$$

$$\text{Recall} = \frac{TP}{TP+FN} \times 100\%$$

$$\text{F1 Score} = 2 \times \frac{Precicion \times Recall}{Preciciion+Recall}$$

IV. RESULTS AND DISCUSSION

A. *Model Training Results*



Figure 1. Model Training Results

The image above shows the results of model testing on the Edge Impulse platform.

Table 1. Confusion Matrix Resutlts

	Melanoma	Non Melanoma
Precision	0.96	0.96
Recall	0.96	0.96
F1-Score	0.96	0.96

The training results show that the CNN-based melanoma detection system and transfer learning developed through Edge Impulse have performance which is excellent. The model achieved an accuracy of 95.16%, indicating that most images can be classified correctly. In addition, the Area Under Curve (AUC) value of 0.96 shows that the model's ability to distinguish between melanoma and non-melanoma lesions is very high. Other evaluation metrics also show

consistent model performance. The weighted precision, recall, and F1-score values are each at 0.96, confirming that the model is not only accurate but also stable in providing correct predictions. This is in line with the effectiveness of the transfer learning method, which is able to capture important features in skin lesion images despite considerable visual variation between samples.

Based on the confusion matrix, the model is able to identify both classes with a very high level of accuracy. In the melanoma class, the model successfully provided 94.2% correct predictions, while in the non-melanoma class it produced up to 96.1% correct predictions. These results show that the model has good sensitivity and specificity, so it can support the early detection of melanoma with a high level of reliability. Overall, the model's strong performance, consistent evaluation metrics, and stable classification capabilities indicate that this approach is highly suitable for application in smartphone-based melanoma detection systems. With the support of the Edge Impulse platform and real-time inference capabilities, this research demonstrates that the use of mobile devices can be a practical and effective alternative in supporting the early detection of skin cancer.

B. *System Performance After Deployment*

After the CNN model was deployed to smartphones via Edge Impulse, live testing was conducted to assess the system's performance in real-world conditions. This testing was conducted by capturing images using a smartphone camera and processing them in real-time through the embedded model. This stage aimed to ensure that the model not only worked well on the test dataset, but also remained stable and accurate when used in an actual operational environment. For testing, 10 melanoma samples and 10 non-melanoma samples were used as a reference for evaluating system performance.



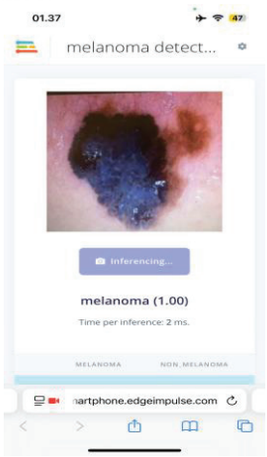


Figure 2. non_melanoma results test
Figure 3. Melanoma results test

Table 2. Test Result

	Melanoma	Non Melanoma
Melanoma	10	0
Non Melanoma	0	10

After deployment, the results showed that the system successfully classified all samples correctly, achieving 100% accuracy in live testing using smartphones. All 10 melanoma images were detected as melanoma, and 10 non-melanoma images were correctly identified as non-melanoma. These results show that the model not only performs well on the dataset, but also maintains its reliability in real-world conditions with varying image quality. This success confirms that smartphone-based melanoma detection has strong potential for use as a fast, practical, and effective early detection tool.

C. *Discussion*

Test results show that the CNN model trained on Edge Impulse performs very well, with 95.16% accuracy on the dataset and consistently high evaluation metrics such as AUC, precision, recall, and F1-score. This shows that the model is able to effectively recognize visual patterns of melanoma and non-melanoma. After the model was deployed to smartphones, its performance became even more stable. Direct testing using smartphone cameras on 20 samples resulted in 100% accuracy, showing that the system works very well in real-world conditions. The consistency between the dataset testing results and the post-deployment testing results proves that this approach is feasible as a practical and accessible tool for early melanoma detection.

ACKNOWLEDGMENT

The authors would like to express their deepest gratitude to all individuals and institutions who have supported the completion of this research. Special appreciation is extended to the lecturers and academic staff for their continuous guidance, valuable insights, and encouragement throughout the research process. The authors also acknowledge the use of the Edge Impulse platform and the Kaggle dataset, which played a crucial role in the development and evaluation of the proposed system. Sincere appreciation is also dedicated to friends who have provided motivation, assistance, and constructive discussions during the preparation of this study. Lastly, the authors would like to convey heartfelt thanks to their families for their unwavering support, patience, and encouragement, which became the foundation that made the completion of this work possible.

REFERENCES

[1] M. F. F. Rozi, S. Mulyono, and G. Ghufron, "Deteksi Kanker Kulit Melanoma Berbasis Android Menggunakan Convolutional Neural Network Arsitektur MobileNET v2," *TRANSISTOR Elektro Dan Inform.*, vol. 5, no. 2, Art. no. 2, Aug. 2023, doi: 10.30659/ei.5.2.89-94.

[2] F. Zulfa, I. Windarti, Sutarto, and N. Carolia, "Profile of Melanoma and Non-Melanoma Skin Cancer based on Histopathology Results," *Med. Prof. J. Lampung*, vol. 14, no. 9, Art. no. 9, 2024, doi: 10.53089/medula.v14i9.1342.

[3] F. Zulfa, I. Windarti, Sutarto, and N. Carolia, "Profile of Melanoma and Non-Melanoma Skin Cancer based on Histopathology Results," *Med. Prof. J. Lampung*, vol. 14, no. 9, pp. 1749–1755, 2024, doi: 10.53089/medula.v14i9.1342.

[4] M. Oktora, Y. Haiga, and N. Liana, "Peningkatan Pengetahuan Mayarakat terhadap Pengaruh Paparan Sinar Matahari yang dapat Menyebabkan Kanker Kulit," *J. Pengabdi. Masy. KesehatanJURABDIKES*, vol. 1, no. 2, pp. 28–31, Dec. 2023, doi: 10.56260/jurabdikes.v1i2.105.

[5] E. P. Sitorus and I. Julianto, "Teknik–teknik Biopsi Kulit," vol. 45, no. 6, 2018.

[6] A. Oktafiani and M. E. Irawanto, "HUBUNGAN TRANSEPIDERMAL WATER LOSS TERHADAP DERAJAT KEPARAHAN PADA PASIEN PSORIASIS VULGARIS," *Media Derm.-Venereol. Indones.*, vol. 51, no. 3, pp. 98–101, Sept. 2024, doi: 10.33820/mdvi.v51i3.437.

[7] M. Q. Hatem, "Skin lesion classification system using a K-nearest neighbor algorithm," *Vis. Comput. Ind. Biomed. Art*, vol. 5, no. 1, p. 7, Dec. 2022, doi: 10.1186/s42492-022-00103-6.

[8] M. A. M. Almeida and I. A. X. Santos, "Classification Models for Skin Tumor Detection Using Texture Analysis in Medical Images," *J. Imaging*, vol. 6, no. 6, p. 51, June 2020, doi: 10.3390/jimaging6060051.

[9] M. Bistroń and Z. Piotrowski, "Comparison of Machine Learning Algorithms Used for Skin Cancer Diagnosis," *Appl. Sci.*, vol. 12, no. 19, p. 9960, Oct. 2022, doi: 10.3390/app12199960.

[10] J. Pujoseno, "Diajukan Sebagai Salah Satu Syarat Untuk Memperoleh Gelar Sarjana Program Studi Statistika".

[11] J. Ismail, L. Tanti, and W. Wanayumini, "DEVELOPMENT OF SKIN CANCER PIGMENT IMAGE CLASSIFICATION USING A COMBINATION OF MOBILENETV2 AND CBAM," *JITK J. Ilmu Pengetah. Dan Teknol. Komput.*, vol. 10, no. 4, pp. 770–780, May 2025, doi: 10.33480/jitk.v10i4.6541.

[12] D. Huang, "Edge impulse-based convolutional neural network for brain tumor classification," *Appl. Comput. Eng.*, vol. 18, no. 1, pp. 188–193, Oct. 2023, doi: 10.54254/2755-2721/18/20230989.

[13] M. S. Diab and E. Rodriguez-Villegas, "Performance Evaluation of Embedded Image Classification Models Using Edge Impulse for Application on Medical Images," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Glasgow, Scotland, United Kingdom: IEEE, July 2022, pp. 2639–2642. doi: 10.1109/EMBC48229.2022.9871108.

[14] G. Regina Pinaso, L. Marcondes Figueiredo, O. Rosa Júnior, and M. R. Da Silva Richetto, "Edge Impulse Potential to Enhance Object Recognition Through Machine Learning," *Semina Ciênc. Exatas E Tecnológicas*, vol. 45, p. e49197, June 2024, doi: 10.5433/1679-0375.2024.v45.49197.

[15] F. Royana, P. Yuniar Maulida, R. Nurul Hasanah, S. Setia Rahayu, and R. Rasim, "Aplikasi Mobile Deteksi Dini Kanker Kulit Berdasarkan Image Processing," *J. Litbang Edusaintech*, vol. 2, no. 2, pp. 100–106, Dec. 2021, doi: 10.51402/jle.v2i2.44.

[16] A. C. Ramdhana and N. Pratiwi, "Perbandingan Kinerja Model Convolutional Neural Network pada Klasifikasi Kanker Kulit | Edumatic: Jurnal Pendidikan Informatika." Accessed: Apr. 24, 2025. [Online]. Available: https://e-journal.hamzanwadi.ac.id/index.php/edumatic/article/view/19823

[17] R. Y. A. Rahman, F. Sthevanie, and G. Kosala, "Deteksi Kanker Kulit Melanoma Menggunakan Derivative of Gaussian dan Convolutional Neural Network," *Log. J. Penelit. Inform.*, vol. 2, no. 1, pp. 1–5, June 2024, doi: 10.25124/logic.v2i1.7529.

[18] F. Dartiko, R. J. Pradana, R. E. Sari, W. Syahputra, and W. K. Oktoeberza, "Klasifikasi Kanker Kulit Berbasis CNN dengan Metode Hybrid Preprocessing".

[19] J. T. Sinaga, H. A. Faudyta, and E. R. Subhiyakto, "Klasifikasi Kanker Kulit menggunakan Convolutional Neural Network dengan Optimasi Arsitektur," *Build. Inform. Technol. Sci. BITS*, vol. 6, no. 3, Art. no. 3, Dec. 2024, doi: 10.47065/bits.v6i3.6141.

[20] M. F. Abdurrohman, "PROGRAM STUDI TEKNIK INFORMATIKA FAKULTAS SAINS DAN TEKNOLOGI UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM MALANG 2024".

Efficient Sentiment Classification via RoBERTa-BiLSTM and Nature-Inspired Feature Optimization

Uzma Shakoor*, Muhammad Khalil Afzal*, Byung-Seo Kim†

*Department of Computer Science, COMSATS University Islamabad, Wah Campus, Wah Cantt 47040, Pakistan
†Department of Software and Communications Engineering, Hongik University, South Korea
Emails: uzma.chauhdry95@gmail.com, khalilafzal@ciitwah.edu.pk, jsnbs@hongik.ac.kr

***Abstract*—Sentiment analysis is crucial when it comes to the interpretation of public opinions on social media. In this study, a transformer-based model structure composed of deep contextual embeddings and meta-heuristic feature selection procedures is incorporated for sentiment analysis on the Twitter US Airline and IMDb datasets. The RoBERTa model is used to generate sentence representations that are fed as 768-dimensional feature vectors and also run through the BiLSTM layer to generate reduced 512-dimensional feature representations that capture sequential dependencies. Two nature-inspired optimization algorithms, namely Whale Optimization Algorithm (WOA) and Genetic Algorithm (GA) are utilized in further parameter-space compression. In the Twitter data, the feature space by WOA was downsized to 288 and GA to 256 and in IMDb, the feature space was reduced to 212 (WOA) and 240 (GA), respectively. The subsets were optimized and then classified with the help of the traditional and deep learning models, such as Decision Tree, Logistic Regression, KNN, BiLSTM, and CNN-LSTM. It has been found that WOA obtained the best individual accuracy of 96.06% using the Decision Tree and significant training time savings. Nevertheless, GA has generated more substantial and more consistent improvements in both datasets, particularly deep models. Overall, it can be noted that GA was more reliable in terms of accuracy and generalization, whereas WOA was more beneficial due to the increased speed of convergence and efficiency. This is one of the first studies to provide a systematic comparison between WOA and GA for feature selection methods in a RoBERTa-BiLSTM sentiment analysis pipeline using multiple benchmark datasets. The suggested framework can be scaled and proves its usefulness in improving the classification performance at a low cost. These findings further support that post-hoc evolutionary pruning can be used to allow practical, high-precision, and deployable transformer models with less memory and computation requirements.**

***Index Terms*—Sentiment analysis, RoBERTa, Whale Optimization Algorithm (WOA), Genetic Algorithm (GA), feature selection, meta-heuristics, transformers.**

I. Introduction

User-generated content in the form of product evaluations and micro-blogging provides organizations with rich and subjective information, albeit in a cluttered form. The text produced on social media constitutes user-generated official communication that can be considered as a yet unorganized source of opinion-rich data. Sentiments have taken a new turn, and a critical consideration regarding the direction of ideas as conveyed in this type of content is one of the significant activities of natural language processing (NLP), especially in customer feedback, market analysis, and monitoring of the opinions. Transformer-based models like RoBERTa have become the models of choice because of such contextual modelling and their capability to perform long-range dependency using text [1], [2]. Nevertheless, they have their share of computational overhead (especially during fine-tuning and inference) due to their high-dimensional embeddings along with their complicated self-attention mechanism.

To solve this problem, a recent experimental framework from the existing work "RoBERTa-BiLSTM: A Context-Aware Hybrid Model for Sentiment Analysis" is utilized [3]. The framework integrates transformer-based feature extraction with sequential modeling and nature-based inspired optimization model as part of the sentiment analysis pipeline. More specifically, a deep contextual embedding (obtained by RoBERTa) is fed to a Bidirectional Long Short-Term Memory (BiLSTM) network in order to capture the serial dependencies and reduce the dimensions of an embedding using fine tuned RoBERTa-BiLSTM. The resulting features are further reduced based on meta-heuristic optimization methods, that is, Whale Optimization Algorithm (WOA) and Genetic Algorithm (GA) to extract most informative features.

The pipeline can be deployed to real-world environments where environmental resources may be limited since the objective of this work is to reduce the computational complexity of the model without compromising the performance of the classification model or lead to improved performance of the classification model. This study serves as a preliminary proof-of-concept. Subsequently, it is planned to apply the GA and WOA-pruned RoBERTa-BiLSTM model to all baselines presented in the RoBERTa-BiLSTM paper to enable a thorough model evaluation of side-by-side performance comparisons.

The main contributions of this work are as follows.

- Utilization of an existing study model RoBERTa-BiLSTM architecture as the baseline model that is used to fine tune and extract features using both sequential and contextual representations.

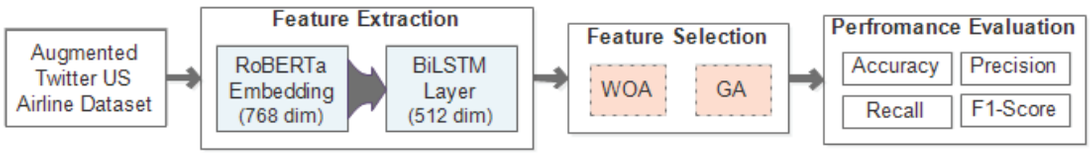


Fig. 1: Block diagram of the proposed RoBERTa-BiLSTM sentiment classification pipeline with WOA and GA-based feature selection.

- Applying Wrapper-based feature selection based on meta-heuristic WOA and GA on the RoBERTa-BiLSTM reduced embeddings to increase efficiency.
- An overall analysis of the suggested feature subsets in Machine Learning (ML) classifiers and Deep Learning (DL) models is performed in order to confirm the performance of the feature optimization strategy.

II. Related Work

Sentiment-analysis research has progressed through three waves.

(i) Traditional machine-learning. Naive Bayes, SVM, KNN, and AdaBoost continue to perform well on smaller or more organized datasets, including political tweets and airline reviews [4], [5]. They also perform well on larger datasets like Sentiment140 [6], provided that the data undergoes thorough pre-processing.

(ii) Deep learning. Models based on CNNs and RNNs (LSTMs, BiLSTMs, GRUs) obtain accuracy rates of 86-91% on movie, Tamil tweet, and mixed-language benchmarks by taking into account both local context as well as long-range dependencies [7]–[9]. These scores are enhanced even further by hybrid systems that fuse pre-trained embedding from Doc2Vec and GloVe.

(iii) Large language models (LLMs). Following BERT's introduction [10], several adaptations for BERT developed, including RoBERTa, ALBERT, and DistilBERT aimed towards the challenge of polarity detection [11]–[13]. It is clear from these studies that the focus is pre-training for a specific domain or architectural modifications [14]–[16]. No work has been done on runtime reductions through post-hoc feature pruning. On COVID-19, airline, and code-mixed corpora, fine-tuned BERT and RoBERTa dominated the competition [12], [13], [17], while trained-less-situated,few-shot models, and GPT-3, PaLM, and LLaMA with prompt marketing and augmentation can surpass supervised benchmarks [14], [15], [18]–[20]. F1 scores above 90% on multiple public datasets have been reported [12], [13], [17]. As a whole, the literature suggests LLMs still set the benchmark for performance – RNN hybrids provide a compelling accuracy-efficiency balance.

A. Meta-heuristic Feature Selection

Algorithms based on evolutionary and swarm intelligence, including GA, PSO, and Ant Colony, have been utilized for a long time to derive compact feature subsets [1], [2]. WOA, introduced by Mirjalili [21], balances global exploration with local exploitation via encircling, bubble-net, and prey-search behaviors. Meta-heuristic algorithms have offered great outcomes in feature selection in the sense that they have enhanced the accuracy of classification and minimized the computational costs and storage requirements [22]. In sentiment analysis, meta-heuristic algorithms have been demonstrated as an efficient method of choosing the best feature combination, dimensionality reduction and enhancing accuracy of classification [1]. In this study, two meta-heuristic optimization algorithms are used in improving the performance of the basic model [3] by enhancing the accuracy and minimizing computation time as compared to the selected base study.

III. Methodology

The entire workflow of the proposed method is illustrated in Fig. 1 and discussed in following subsections.

A. RoBERTa-BiLSTM Model for Feature Extraction

In this work, existing hybrid RoBERTa-BiLSTM model is tuned and used for feature selection [3]. The sequential dependencies are captured by passing the pre-trained contextual embeddings generated using RoBERTa transformer through a bidirectional layer of LSTM. It is fine-tuned on sentiment-labeled data, and the internal representations, such as the outputs of LSTM and pooled RoBERTa, are taken as high-level features. The extracted features are then used to perform downstream classification tasks and are also further optimized using meta-heuristic feature selection algorithms, namely WOA and GA, to increase the classification accuracy and decrease the computational cost.

B. Meta-Heuristic Feature Selection: WOA and GA

WOA and GA are adopted as a feature selection method in this study. Similar to the bubble-net hunting process of humpback whales, WOA is an algorithm that replicates their spiral and encircling movement to attain the balance between exploitation and exploration of the search space [21]. It is especially applicable to optimize high-dimensional feature sets. Based on evolutionary principles, GA uses crossover, mutation, and selection processes to evolve optimal subsets of properties [23]. Both methods are applied to remove redundant or meaningless features, increasing the efficiency and performance of the model.

C. Performance Evaluation

To evaluate the performance of extracted features, traditional machine learning classifiers and deep learning models are employed. The models are compared before and after feature selection to evaluate any improvement to the classification

TABLE I: Performance Comparison of ML Classifiers on Twitter US Airlines Dataset with WOA-FS and GA-FS Algorithms

Classifier	**No FS**				**WOA-FS**				**GA-FS**			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
Naive Bayes	0.9147	0.9151	0.9147	0.9144	0.9147	0.9153	0.9147	0.9144	0.9254	0.9268	0.9254	0.9252
Logistic Regression	0.9312	0.9312	0.9312	0.9309	0.9254	0.9252	0.9254	0.9251	0.9348	0.9354	0.9348	0.9346
Decision Tree	0.9577	0.9597	0.9577	0.9576	0.9599	0.9619	0.9599	0.9597	0.9599	0.9616	0.9599	0.9598
KNN	0.9183	0.9185	0.9183	0.9175	0.9211	0.9212	0.9211	0.9206	0.9262	0.9278	0.9262	0.9257
AdaBoost	0.9161	0.9160	0.9161	0.9159	0.9140	0.9138	0.9140	0.9137	0.9190	0.9194	0.9190	0.9189

TABLE II: Performance Comparison of DL Models on Twitter US Airlines Dataset with WOA-FS and GA-FS Algorithms

Model	**No FS**				**WOA-FS**				**GA-FS**			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
GRU	0.9240	0.9242	0.9240	0.9238	0.9204	0.9203	0.9204	0.9203	0.9305	0.9304	0.9305	0.9303
LSTM	0.9247	0.9254	0.9247	0.9247	0.9226	0.9228	0.9226	0.9223	0.9341	0.9348	0.9341	0.9339
BiLSTM	0.9283	0.9285	0.9283	0.9280	0.9262	0.9267	0.9262	0.9260	0.9348	0.9353	0.9348	0.9346
CNN-LSTM	0.9183	0.9180	0.9183	0.9179	0.9168	0.9165	0.9168	0.9166	0.9262	0.9259	0.9262	0.9260

TABLE III: Training Time Comparison of ML Models with WOA-FS and GA-FS on Twitter US Airline Dataset

Model	**No FS (s)**	**WOA**			**GA**		
		FS Time (s)	Diff (s)	% Change	FS Time (s)	Diff (s)	% Change
Naive Bayes	0.07	0.05	0.02	28.57%	0.06	0.01	14.29%
Logistic Regression	5.36	2.73	2.63	49.07%	3.34	2.02	37.69%
Decision Tree	39.09	17.59	21.50	54.99%	14.17	24.92	63.74%
KNN	0.01	0.01	0.00	0.00%	0.02	-0.01	-100.0%
AdaBoost	59.33	28.28	31.05	52.33%	25.83	33.50	56.45%

TABLE IV: Training Time Comparison of DL Models with WOA-FS and GA-FS on Twitter US Airline Dataset

Model	**No FS (s)**	**WOA**			**GA**		
		FS Time (s)	Diff (s)	% Change	FS Time (s)	Diff (s)	% Change
GRU	13.69	7.28	6.41	46.83%	14.04	-0.35	-2.56%
LSTM	14.36	7.19	7.17	49.93%	13.37	0.99	6.90%
BiLSTM	16.78	8.18	8.60	51.27%	15.79	0.99	5.90%
CNN-LSTM	22.75	9.03	13.72	60.32%	17.30	5.45	23.96%

performance metrics such as the accuracy and precision. Also, the computing time of each model is compared to evaluate the computational efficiency of dimensionality reduction. The evaluation provides an indication into how well the proposed feature selection technique performed in terms of improving the performance at the same time cutting down computational cost.

IV. Experimental Setup

A. Dataset

In this work, Twitter US Airline and IMDb datasets, as employed in the base paper [3], are utilized. Twitter US Airline is a public Twitter US Airline corpus which consists of 14,640 English tweets labeled as *negative*, *neutral*, or *positive*. [1] The base paper authors have provided a preprocessed and augmented version of this dataset. The IMDb database is based on 50,000 reviews, equally balanced between positive and negative classes. A 90/5/5 stratified split is performed on the balanced datasets as in the selected study [3].

[1]This dataset was made available by *CrowdFlower* in 2015 and serves as a tri-class sentiment benchmark for evaluation on multiple classifiers, making it a popular choice in the research community.

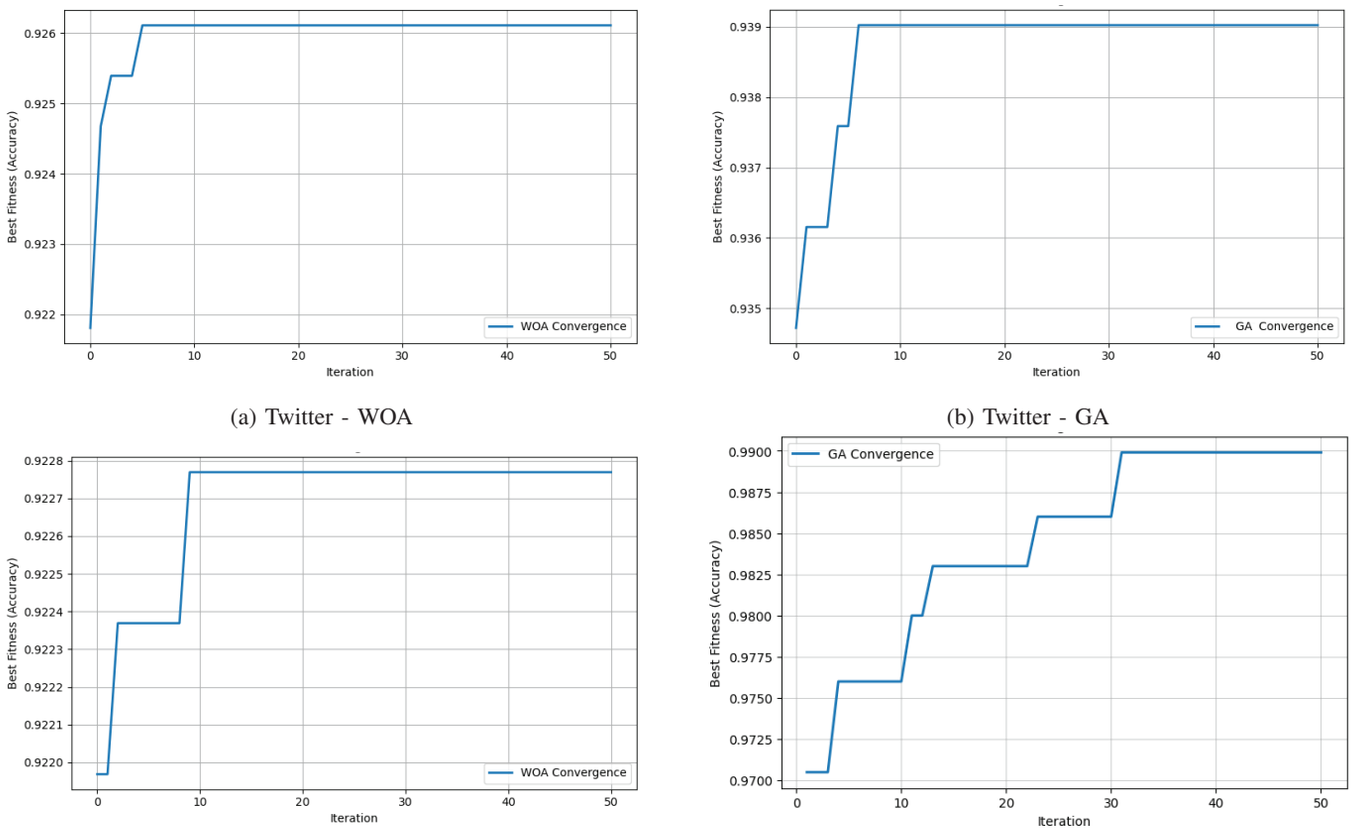
B. Hyper-parameters

- **RoBERTa-BiLSTM:** `roberta-base`; 5 epochs; Adam optimizer with learning-rate 10^{-5}; BiLSTM with 256 hidden units; batch size 16 (train), 32 (eval); last hidden state used for classification.
- **WOA:** population $N = 10$; iterations $T = 50$.
- **GA:** population $N = 10$; generations $T = 50$.
- **Fitness Function classifier:** logistic regression (maximum iterations =1000).

All experiments are executed on a single NVIDIA RTX-3060 (12 GB) GPU. Results are reported as accuracy, precision, recall and F1 score.

V. Results and Discussion

Two benchmark datasets, Twitter US Airline and IMDb, were evaluated experimentally using WOA and GA as feature selection algorithms. Multiple ML classifiers and DL models



(a) Twitter - WOA

(b) Twitter - GA

(c) IMDb - WOA

(d) IMDb - GA

Fig. 2: Convergence trajectories of WOA and GA feature selection algorithms on Twitter US aireline and IMDB datasets

were compared based on their performance in terms of accuracy, precision, recall, F1-score, and training time. For both datasets RoBERTa-BiLSTM model, was used to extract 512 features prior to applying feature selection.

A. Twitter US Airline Dataset

The WOA, after feature selection, reduced the feature space to 288 features and GA reduced it to 256 features. This downsizing contributed to the performance growth experienced in the models. GA and WOA were employed feature selection over the 512 extracted features. The fitness convergence curve is illustrated from convergence curve is illustrated in Figure 2a and Figure 2b. The error decreased sharply during the first two iterations, stabilizing after roughly ten, indicating fast convergence. The performance results are summarized in I, II. In the case of ML classifiers, both WOA-FS and GA-FS were better than the baseline (No FS), but the GA performed best in general. Under GA, linear regression increased in accuracy from 0.9284 to 0.9316 and under the influence of WOA and GA, Decision Tree achieved the best result (0.9599). KNN also increased up to 0.9262 under GA. In the case of DL models, GA once again presented better results, especially with BiLSTM (0.9348 accuracy) and with LSTM (0.9341), whereas with WOA it led to small losses over baseline. The analysis of training time showed that WOA saved a lot (up to 60.32% in CNN-LSTM) of costs, to the contrary, GA had a higher dimensionality reduction and had a stable runtime across most DL models with slight increases in GRU and CNN-LSTM as depicted in III and IV.

In the case of WOA classifiers, most of the ML classifiers improved by a slight or moderate level of performance. It was observed that the Accuracy of the Decision Tree rose to 0.9599 as compared to the 0.9577 and KNN changed to 0.9211 against 0.9183. In the meantime, AdaBoost and Logistic Regression demonstrated slight declines in accuracy, as well as in the range of stable performance. Regarding the training time cost, WOA was able to greatly minimize the computational cost of all ML models except KNN that is not affected by WOA. Training time decreased by 52.33 percent using AdaBoost and 54.99 percent using decision tree. But in DL models, GRU, LSTM, and BiLSTM increased slightly in time to train, i.e. perhaps there is an influence on the network convergence activity due to the reshapes in the features input. CNN-LSTM was the only one that presented an increased efficiency of training time.

B. IMDb Dataset

The ML and DL results of IMDb dataset, as shown in the Tables I and VI), have improved in the most cases. The fitness convergence curve is illustrated in Figure 2c and Figure 2d GA

TABLE V: Performance Comparison of ML Classifiers on IMDB Dataset with WOA-FS and GA-FS Algorithms

Classifier	No FS				WOA-FS				GA-FS			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
Naive Bayes	0.9340	0.9340	0.9340	0.9340	0.9324	0.9324	0.9324	0.9324	0.9340	0.9340	0.9340	0.9340
Logistic Regression	0.9284	0.9285	0.9284	0.9284	0.9316	0.9316	0.9316	0.9316	0.9316	0.9316	0.9316	0.9316
Decision Tree	0.9196	0.9197	0.9196	0.9196	0.9244	0.9245	0.9244	0.9244	0.9216	0.9216	0.9216	0.9216
KNN	0.9316	0.9317	0.9316	0.9316	0.9328	0.9329	0.9328	0.9328	0.9360	0.9361	0.9360	0.9360
AdaBoost	0.9312	0.9313	0.9312	0.9312	0.9320	0.9321	0.9320	0.9320	0.9300	0.9301	0.9300	0.9300

TABLE VI: Performance Comparison of DL Models on IMDB Dataset with WOA-FS and GA-FS Algorithms

Model	No FS				WOA-FS				GA-FS			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
GRU	0.9340	0.9340	0.9340	0.9340	0.9324	0.9325	0.9324	0.9324	0.9320	0.9321	0.9320	0.9320
LSTM	0.9340	0.9340	0.9340	0.9340	0.9328	0.9329	0.9328	0.9328	0.9328	0.9329	0.9328	0.9328
BiLSTM	0.9348	0.9348	0.9348	0.9348	0.9332	0.9333	0.9332	0.9332	0.9328	0.9329	0.9328	0.9328
CNN-LSTM	0.9340	0.9340	0.9340	0.9340	0.9312	0.9314	0.9312	0.9312	0.9328	0.9329	0.9328	0.9328

TABLE VII: Training Time Comparison of ML Models with WOA-FS and GA-FS on IMDb Dataset

Model	No FS (s)	WOA			GA		
		FS Time (s)	Diff (s)	% Change	FS Time (s)	Diff (s)	% Change
Naive Bayes	0.06	0.05	0.01	16.67%	0.06	0.00	0.00%
Logistic Regression	5.36	2.73	2.63	49.07%	3.34	2.02	37.69%
Decision Tree	39.09	17.59	21.50	54.99%	14.17	24.92	63.74%
KNN	0.01	0.01	0.00	0.00%	0.02	-0.01	-100.0%
AdaBoost	59.19	28.28	30.91	52.22%	25.83	33.36	56.34%

TABLE VIII: Training Time Comparison of DL Models with WOA-FS and GA-FS on IMDb Dataset

Model	No FS (s)	WOA			GA		
		FS Time (s)	Diff (s)	% Change	FS Time (s)	Diff (s)	% Change
GRU	6.57	7.28	-0.71	-10.81%	7.10	-0.53	-8.07%
LSTM	7.04	7.19	-0.15	-2.13%	6.95	0.09	1.28%
BiLSTM	7.81	8.18	-0.37	-4.74%	7.60	0.21	2.69%
CNN-LSTM	11.44	9.03	2.41	21.07%	9.11	2.33	20.38%

has chosen 240, whereas WOA has chosen 212 features, and it allows more discriminative representations. In both the ML and DL models, the best performance was always recorded by GA. In the case of ML classifier, KNN achieved an improvement of 0.9316 (baseline) to 0.9360 using GA and Decision tree and logistic regression also had moderate improvements. In the case of DL models, GA slightly improved BiLSTM and CNN-LSTM, and WOA tended to yield little decreases compared to the baseline. In the efficiency of training, WOA saved the highest ML runtimes (by up to 54.99), whereas GA had the greatest savings on Decision Tree (63.74 %). GA was one of the DL models that had a constant runtime and CNN-LSTM retained a savings of approximately 20%. The negative time savings instances in GRU with WOA can be explained by the fact that the overhead was restructured due to the change in feature inputs, but the reduced dimensionality was achieved through feature selection. In general, GA performed better than WOA in both datasets, which produced better accuracy rates and less volatile improvements in both models of ML and DL. In training deep models, WOA was especially effective at minimizing training times, however its accuracy improvements were less and even negative. GA, in turn, gave steady gains in addition to competitive runtime decreases and, therefore, should be considered the more reliable option, whereas WOA proves useful in the contexts, where the computational efficiency will be the major priority.

VI. Conclusion

In this research, an effective and efficient sentiment analysis pipeline was proposed, and this approach combined transformer-based embeddings, sequential modeling (i.e., based on the RoBERTa-BiLSTM paper [3]), and metaheuristic feature selection. It was possible to get rich 512-dimensional feature vectors by utilizing RoBERTa as a contextual representation of the sentence and BiLSTM as a way to capture the temporal dependencies. To decrease dimensionality of features, WOA and GA were then employed. The benefits of both methods were demonstrated by experiments using two benchmark datasets Twitter US Airline and the IMDb. The greater accuracy and stability of GA in both ML and DL classifiers improved Decision Tree, Logistic Regression, KNN,

BiLSTM, and CNN-LSTM. The slightly less accurate WOA offered large savings in training cost, with the highest runtime savings of up to 60 % and the individual highest accuracy of 96.06 % using Decision Tree.

Overall, GA is more appropriate in case accuracy and robustness are of the utmost importance, and WOA is more adequately used in areas where performance is paramount. The suggested framework is scalable and efficient, and it is an effective solution to sentiment analysis on a large scale. Further research can apply the method to more diverse datasets and explore hybrid feature selection strategies.

REFERENCES

[1] S. R. Ahmad, A. A. Bakar, and M. R. Yaakub, "Metaheuristic algorithms for feature selection in sentiment analysis," in *2015 Science and Information Conference (SAI)*. IEEE, 2015, pp. 222–226.

[2] S. A. Al-Shalif, N. Senan, F. Saeed, W. Ghaban, N. Ibrahim, M. Aamir, and W. Sharif, "A systematic literature review on meta-heuristic based feature selection techniques for text classification," *PeerJ Computer Science*, vol. 10, p. e2084, 2024.

[3] M. M. Rahman, A. I. Shiplu, Y. Watanobe, and M. A. Alam, "Robertabilstm: A context-aware hybrid model for sentiment analysis," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2025.

[4] M. Wongkar and A. Angdresey, "Sentiment analysis using naive bayes algorithm of the data crawler: Twitter," in *2019 Fourth international conference on informatics and computing (ICIC)*. IEEE, 2019, pp. 1–5.

[5] A. M. Rahat, A. Kahir, and A. K. M. Masum, "Comparison of naive bayes and svm algorithm based on sentiment analysis using review dataset," in *2019 8th International Conference System Modeling and Advancement in Research Trends (SMART)*. IEEE, 2019, pp. 266–270.

[6] Y. G. Jung, K. T. Kim, B. Lee, and H. Y. Youn, "Enhanced naive bayes classifier for real-time sentiment analysis with sparkr," in *2016 international conference on information and communication technology convergence (ICTC)*. IEEE, 2016, pp. 141–146.

[7] M. Rhanoui, M. Mikram, S. Yousfi, and S. Barzali, "A cnn-bilstm model for document-level sentiment analysis," *Machine Learning and Knowledge Extraction*, vol. 1, no. 3, pp. 832–847, 2019.

[8] S. Anbukkarasi and S. Varadhaganapathy, "Analyzing sentiment in tamil tweets using deep neural network," in *2020 fourth international conference on computing methodologies and communication (ICCMC)*. IEEE, 2020, pp. 449–453.

[9] R. M. Alahmary, H. Z. Al-Dossari, and A. Z. Emam, "Sentiment analysis of saudi dialect using deep learning techniques," in *2019 International conference on electronics, information, and communication (ICEIC)*. IEEE, 2019, pp. 1–6.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019.

[11] Q. Zhong, L. Ding, J. Liu, B. Du, and D. Tao, "Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert," 2023.

[12] M. Singh, A. K. Jakhar, and S. Pandey, "Sentiment analysis on the impact of coronavirus in social life using the bert model," *Social Network Analysis and Mining*, vol. 11, no. 1, p. 33, 2021.

[13] K. L. Tan, C. P. Lee, K. S. M. Anbananthen, and K. M. Lim, "Roberta-lstm: a hybrid model for sentiment analysis with transformer and recurrent neural network," *Ieee Access*, vol. 10, pp. 21 517–21 525, 2022.

[14] B. Chen, Z. Zhang, N. Langrené, and S. Zhu, "Unleashing the potential of prompt engineering for large language models," *Patterns*, 2025.

[15] W. Zhang, Y. Deng, B. Liu, S. J. Pan, and L. Bing, "Sentiment analysis in the era of large language models: A reality check," 2023.

[16] S. Poria, D. Hazarika, N. Majumder, and R. Mihalcea, "Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research," *IEEE transactions on affective computing*, vol. 14, no. 1, pp. 108–132, 2020.

[17] A. Younas, R. Nasim, S. Ali, G. Wang, and F. Qi, "Sentiment analysis of code-mixed roman urdu-english social media text using deep learning approaches," in *2020 IEEE 23rd International Conference on Computational Science and Engineering (CSE)*. IEEE, 2020, pp. 66–71.

[18] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[19] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann *et al.*, "Palm: Scaling language modeling with pathways," *Journal of Machine Learning Research*, vol. 24, no. 240, pp. 1–113, 2023.

[20] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," 2023.

[21] S. Mirjalili and A. Lewis, "Whale optimization algorithm," *Advances in Engineering Software*, vol. 95, pp. 51–67, 2016.

[22] O. M. Alyasiri, Y.-N. Cheah, A. K. Abasi, and O. M. Al-Janabi, "Wrapper and hybrid feature selection methods using metaheuristic algorithms for english text classification: A systematic review," *IEEE Access*, vol. 10, pp. 39 833–39 852, 2022.

[23] F. Iqbal, J. M. Hashmi, B. C. Fung, R. Batool, A. M. Khattak, S. Aleem, and P. C. Hung, "A hybrid framework for sentiment analysis using genetic algorithm based feature reduction," *IEEE access*, vol. 7, pp. 14 637–14 652, 2019.

Analyzing User Acceptance of News on Social Media Using Machine Learning

Mochammad Risyad Aukasyah Fatahillah*, Jeremia Johanes Mikhael Paulus†, Eka Miranda‡
*Information Systems Department, School of Information Systems, Bina Nusantara University, Jakarta, Indonesia
†Information Systems Department, School of Information Systems, Bina Nusantara University, Jakarta, Indonesia
‡Information Systems Department, School of Information Systems, Bina Nusantara University, Jakarta, Indonesia
Emails: mochammad.fatahillah001@binus.ac.id, jeremia.mikhael@binus.ac.id, ekamiranda@binus.ac.id

***Abstract*—This research examines user acceptance of news on social media through machine-learning methodologies, utilising a Likert-scale questionnaire with 401 participants. The dataset consists of 24 behavioural, cognitive, and perceptual variables categorised by labels indicating news acceptance, neutrality, or non-acceptance. After preprocessing the data and finding the class imbalance, the dataset was split into 80% training and 20% testing, and SMOTE balanced the training data. We used Accuracy, Precision, Recall, F1-Score, and ROC-AUC to train and test Random Forest and XGBoost. With 96.30% accuracy and 99.87% ROC-AUC, Random Forest with SMOTE did better than imbalanced models and XGBoost. We used SHAP (Shapley Additive Explanations) to look at how the model worked and found that inoculation resistance, perceived correctness, and personality traits had the biggest effect on how things were categorised. The findings indicate that SMOTE, ensemble learning, and explainable AI methodologies are effective in simulating multi-class news acceptance on social media platforms.**

***Index Terms*—Machine Learning, User Acceptance, Social Media, Random Forest, Fake News Detection**

I. Introduction

Many people get news from Instagram, X, TikTok, Facebook, etc. Personalization algorithms and user-generated content speed up news delivery, changing digital news consumption. Increased information flow spreads misinformation, disinformation, and false news [1] [2]. News may be trusted more for emotional appeal, familiarity, or social connections than for factual accuracy [3] [4].

Psychological, social, and technological aspects affect media news acceptability. Previous study indicates that media credibility, trust, perceived value, and motivation greatly affect online information engagement [5] [6]. Despite increasing user experience, algorithmic personalization may impact news acceptance due to user awareness and perception of algorithms [7]. When daily news is deceptive, user acceptability matters for digital information ecosystem decision-making.

Machine learning models news consumption. XGBoost and Random Forest predict category, survey, and Likert-scale data well despite noise and non-linear relationships. [8] [9] [10] [11]. These algorithms are commonly employed in public opinion analysis, sentiment modeling, and user behavior prediction [12]. This suggests that ML-based approaches can anticipate social media users' news acceptance.

A Likert-scale questionnaire measures user acceptability utilizing Attitude (AC), Intention (IN), Entertainment (ET), Reliability (RL), Motivation (MT), and Perceived Value. For computational processing, textual categories like Strongly Disagree to Strongly Agree are assigned numerical values. Distribution patterns, variable correlations, and news acceptance/non-acceptance class balance are examined using EDA. To assess model generalization to new data, the dataset is split 80:20 into training and testing sets.

News acceptance is classified by Random Forest and XGBoost. Accuracy, precision, recall, F1-score, confusion matrix analysis, and ROC-AUC quantify model predictability. We select the most influential predictors from each model's feature importance ratings. These data demonstrate which psychological or behavioral elements most effect user acceptability.

Because AI-driven decision making should be transparent, this research uses Explainable AI. SHAP (Shapley Additive Explanations) enables global and local interpretation, with each feature predicting specific outcomes. [13] [14] [15] SHAP helps academics comprehend social media news acceptance by explaining the paradigm.

Structure of the paper follows. Section II covers algorithmic influence, social media news consumption, and machine-learning behavioral prediction. Section III details the dataset, preprocessing workflow, class-balancing process, and Random Forest, XGBoost, and SHAP machine-learning methods employed in this study. Section IV presents performance evaluation with and without SMOTE, confusion matrix analysis, and model comparison across metrics experiments. Section V presents findings on social media news acceptance modeling and interpretation research.

II. Literature Review

A. News Consumption on Social Media

Digital platforms have transformed news consumption. Social media speeds up information distribution but also spreads misinformation and disinformation [1]. Unverified

content spreads quicker than verified news due to novelty, emotional framing, and social reinforcement [2]. Instagram, X, TikTok, and Facebook dominate how users experience news, affecting their perceptions, attitudes, and acceptance. Frequent exposure to news, regardless of veracity, can lead to inaccurate credibility judgments [3]. Understanding how audiences accept or reject news in algorithm-driven and socially amplified situations is crucial.

B. Algorithmic Personalization and Awareness

Using recommendation systems, social media networks filter and personalize news. This automated curation affects what users see and how they see recommended content's legitimacy and relevancy. Research indicates that viewers with little awareness of algorithmic filtering may perceive news as representing popular opinion, leading to increased acceptance [7]. Users might be more critical and discriminating when analyzing information with more algorithm knowledge. Research suggests that source credibility, platform familiarity, and social media usage patterns greatly impact users' evaluation of tailored news feeds [16]. Thus, algorithmic openness and consumer understanding are crucial to digital news acceptability.

C. Machine Learning Algorithms for Behavioural Prediction

Many machine learning algorithms predict behavior using survey and Likert-scale data. Random Forest and XGBoost excel in non-linear patterns, categorical data, and prediction. [9] [8] [10]. Random Forest uses majority voting to combine numerous decision trees to reduce overfitting and improve resilience in user-generated data. [9]. The stable and noise-resistant model is ideal for mimicking news-acceptance behavior when psychological and contextual elements vary considerably [11]. XGBoost sequentially optimizes decision trees with gradient boosting. It effectively handles missing values, class imbalance, and complicated feature interactions, making it ideal for sentiment analysis and public opinion prediction [12]. Social media consumers' news material acceptance or rejection is precisely predicted by algorithms.

D. Explainable AI (XAI) and SHAP Interpretability

Interpretability is necessary for fairness and user trust in black box machine learning models. XAI tools can explain model behavior, identify decision limits, and discover bias [17] [15]. SHAP uses game theory to assign contribution scores to features, providing global and local explanations [13]. SHAP identifies behavioral aspects affecting news acceptance and improves transparency in disinformation and media research systems.

E. Receiver Operating Characteristic (ROC) and AUC

The Receiver Operating Characteristic (ROC) curve is a way to see how well the classification model can tell the difference between classes at different decision thresholds. ROC shows how the True Positive Rate (TPR) and the False Positive Rate (FPR) are related.

$$\text{TPR} = \frac{TP}{TP + FN}, \qquad \text{FPR} = \frac{FP}{FP + TN}. \tag{1}$$

A excellent model makes a ROC curve that goes up significantly toward the upper-left corner. This shows that the model is very sensitive and has few false positives. The Area Under the Curve (AUC) is used to turn this curve into a single number. AUC tells you how likely it is that the classifier will rank a randomly chosen positive instance higher than a randomly picked negative one. The continuous version of the AUC is written as

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR}), \tag{2}$$

while in practice it is commonly approximated using the trapezoidal rule:

$$\text{AUC} \approx \sum_{i=1}^{n-1} \frac{\text{TPR}_i + \text{TPR}_{i+1}}{2} \left(\text{FPR}_{i+1} - \text{FPR}_i\right). \tag{3}$$

Higher AUC values indicate better discriminatory power, where an AUC close to 1.0 reflects excellent performance, and a value of 0.5 indicates no better prediction than random guessing.

III. Material and Methods

A. Dataset

This study obtained data from an online questionnaire issued between 21 September and 21 November 2025, using a Likert scale from Strongly Disagree (0) to Strongly Agree (5). The instrument had 25 features and one class label for social media news acceptance. These features were developed based on previous research on misinformation, credibility, and user behaviour [3] [4] [2], [13]. This study's questionnaire items are listed in Tables I. After numerical encoding and structural consistency cleaning, all replies were ready for supervised machine-learning modelling.

TABLE I
Questionnaire Items

Code	Questions	Options	Lit.
Class Label			
	What is your level of acceptance of news on social media?	0 = A 1 = N 2 = NA	
Perceived Accuracy			
PA1	I believe that the news I see on social media is accurate.	0 = SD 1 = D 2 = A 3 = SA	[14]
PA2	I am confident that the news is based on factual information.	0 = SD 1 = D 2 = A 3 = SA	[14]

PA3	I consider the news I encounter on social media to be trustworthy.	0 = SD 1 = D 2 = A 3 = SA	[14]
Sharing Intention			
SI1	I am willing to share this news on my social media.	0 = SD 1 = D 2 = A 3 = SA	[14]
SI2	I feel motivated to share this news with my friends.	0 = SD 1 = D 2 = A 3 = SA	[14]
SI3	If I believe the news, I will share it with others.	0 = SD 1 = D 2 = A 3 = SA	[14]
Social Media News Use			
SMU1	I read news from social media every day.	0 = SD 1 = D 2 = A 3 = SA	[14]
SMU2	I follow accounts or pages that provide news updates.	0 = SD 1 = D 2 = A 3 = SA	[14]
SMU3	I actively comment or like news content on social media.	0 = SD 1 = D 2 = A 3 = SA	[14]
Cognitive Ability			
CA1	I think carefully before believing online news.	0 = SD 1 = D 2 = A 3 = SA	[14]
CA2	I can distinguish between facts and opinions in news.	0 = SD 1 = D 2 = A 3 = SA	[14]
CA3	I evaluate information logically before trusting it.	0 = SD 1 = D 2 = A 3 = SA	[14]
Personality Traits			
PT1	I am open to new ideas or information.	0 = SD 1 = D 2 = A 3 = SA	[14]
PT2	I usually think carefully before acting.	0 = SD 1 = D 2 = A 3 = SA	[14]
PT3	I easily feel worried or emotional toward information.	0 = SD 1 = D 2 = A 3 = SA	[14]
Trust in Source			
TS1	I trust news more when it comes from verified or official accounts.	0 = SD 1 = D 2 = A 3 = SA	[3]
TS2	Account credibility is my main consideration when evaluating news.	0 = SD 1 = D 2 = A 3 = SA	[3]
TS3	I trust news shared by people I personally know.	0 = SD 1 = D 2 = A 3 = SA	[3]
Risk Propensity			
RP1	I often share news even without verifying its accuracy.	0 = SD 1 = D 2 = A 3 = SA	[3]
RP2	I am attracted to sensational news even when its truth is unclear.	0 = SD 1 = D 2 = A 3 = SA	[3]
RP3	I do not mind taking risks when sharing news on social media.	0 = SD 1 = D 2 = A 3 = SA	[3]
Inoculation Resistance			
IR1	I can identify manipulation techniques in fake news.	0 = SD 1 = D 2 = A 3 = SA	[4]
IR2	I have attended training or education about misinformation.	0 = SD 1 = D 2 = A 3 = SA	[4]
IR3	I am skeptical toward news that is too provocative.	0 = SD 1 = D 2 = A 3 = SA	[4]

The responses to the questionnaire are the main data set for this study and show the most important things that affect how people accept news on social media. After preprocessing and balancing the data, various machine-learning models are trained and assessed to identify the most dependable forecasting methodology. Figure 1 shows a summary of the whole methodological workflow.

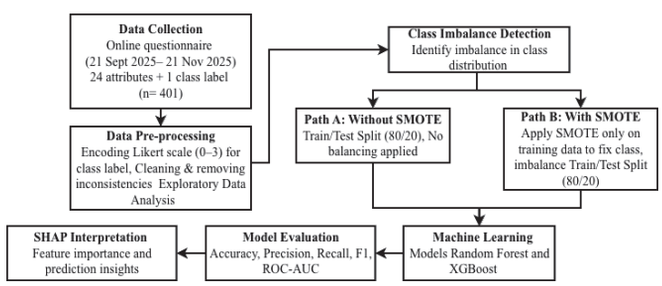


Fig. 1. Block Diagram Research Framework

This research methodology begins with an online questionnaire from September 21 to November 21 that collected data from 401 respondents with 24 attributes and one class label. Pre-processing activities included turning Likert scale replies into numbers, cleansing the data, and undertaking early exploration to ensure quality and consistency. Then, class imbalance was determined to determine data balancing needs. To reduce class imbalance, the research focused on two processing pathways: one without SMOTE and one using it just on training data. Both strategies used Random Forest and XGBoost to create classification models. Accuracy, precision, recall, F1-score, and ROC-AUC metrics assessed model performance. The next step was to apply SHAP to determine how much each attribute assisted the forecast and understand what influences social media news reactions.

B. Method

1) Data pre-processing: Preprocessing raw questionnaire responses organizes them for machine-learning analysis. Initially, textual Likert-scale responses (”Strongly Disagree,” ”Disagree,” ”Agree,” ”Strongly Agree”) were converted to numerical values (0-5) for computational modeling [2]. Preprocessing also removed duplicate responses, handled miss-

ing or inconsistent entries, and checked structural homogeneity across 25 features. The cleaning stage verifies that each record belongs to a unique responder and that the encoded variables appropriately reflect psychological and behavioral characteristics related to news acceptance [4]

2) Data Balancing and Splitting: The dataset has class imbalance because news acceptance labels (Accept, Neutral, Not Accept) were not fairly distributed. Machine-learning models may favor the majority class and lower accuracy for minority groups [3]. To solve this issue, SMOTE was used to generate synthetic examples for underrepresented classes by interpolating their feature space [13]. Information leaking was avoided by applying SMOTE just to the training partition. The dataset was evenly split into an 80:20 train-test split, which is a popular way to do things in machine learning. 80% was utilized for training and 20% for testing the model. [14].

3) Algorithms for Prediction: Both Random Forest and XGBoost were utilized for categorization. Random Forest, utilizing multiple decision trees and majority voting, enhances prediction accuracy on Likert-scale and behavioral datasets [9]. XGBoost, a gradient-boosted tree approach, enhances prediction accuracy for complicated non-linear user behavior patterns by sequential learning and regularization [12]. The original imbalanced dataset and SMOTE-balanced data were used to train both methods. This comparison shows how balance affects news-acceptance classification prediction.

4) Evaluation Model: Accuracy, precision, recall, and F1-score were used to evaluate model performance and get insights into categorization quality [16]. A confusion matrix was created for each model to show correct and wrong predictions across the three acceptance classes. Higher AUC values indicate superior class-separation capabilities in each model, as measured by the Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC) value [7]. SHAP (Shapley Additive Explanations) was utilized to measure how the characteristics of the questionnaire affected the results of news acceptance, giving both global and individual-level explanations. [13].

IV. RESULT AND DISCUSSION

Table 2 shows the traits of the 401 people that took part in this study. Demographic data encompassed gender, age, educational attainment, and place of residence. This classification is essential for comprehending the respondents' backgrounds and guaranteeing that the gathered data reflects a varied cohort of active social media users from different areas and age demographics. The frequency and percentage distributions in Table 2 also show how different the respondents are, which could affect how they react to news on social media.

The survey provided 401 valid responses for analysis. Social media users actively consume news, and the dataset includes 25 psychological, behavioral, and cognitive factors affecting news acceptance. This study classifies news articles as "Accept", "Neutral", or "Not Accept" based on user acceptability on Instagram, X, TikTok, Facebook, and others.

TABLE II
RESPONDENTS

Demographic Information	Category	Frequency	Percentage
Gender	Male	217	54.1%
	Female	184	45.9%
Age	19–25	169	42.1%
	26–32	99	24.7%
	33–40	88	21.9%
	>40	45	11.2%
Education Level	Junior/Senior High School	71	17.7%
	Diploma	94	23.4%
	S1	169	42.1%
	S2	62	15.5%
	S3	5	1.2%
Domicile	Jabodetabek	290	72.32%
	Yogyakarta	27	6.73%
	East Java	27	6.73%
	West Java	22	5.49%
	Bali	16	3.99%
	Central Java	15	3.74%
	Others	4	1.00%
Platform	Instagram	340	84.8%
	X/Twitter	140	34.9%
	Tiktok	314	78.3%
	Threads	151	37.7%
	Facebook	99	24.7%
	Others	7	1.4%

To prepare for machine-learning modeling, the dataset was split 80:20 train-test, with 80% (n = 321) used for training and 20% (n = 80) for testing. This approach enables Random Forest and XGBoost learn generalizable patterns without overfitting.

The class-label distribution showed that "Accept" was overrepresented and "Not Accept" was underrepresented. Favoring the dominant class can harm classification performance with imbalanced target classes. Therefore, before model training, the SMOTE approach was employed to oversample minority classes and balance class distribution.

Figure 2 shows the class distribution before and after SMOTE, showing how oversampling creates a balanced dataset for model training reliability.

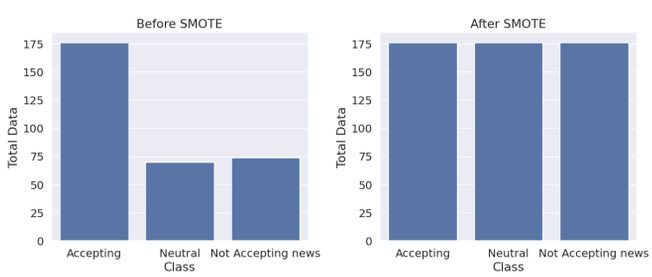


Fig. 2. illustrates class distribution imbalance and SMOTE correction.

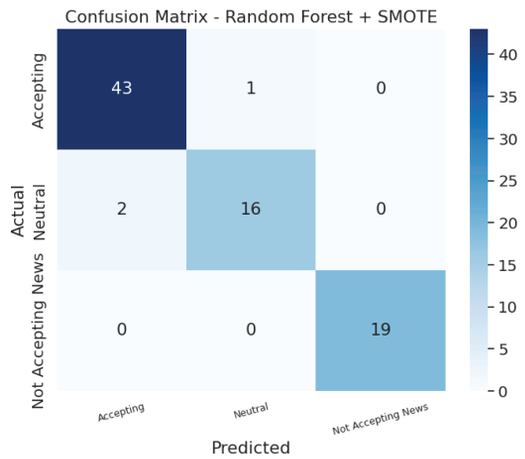


Fig. 3. illustrates class distribution imbalance and SMOTE correction.

Figure 3 confusion matrix shows how well the Random Forest model performs on an SMOTE-balanced dataset. It demonstrates how Accepting, Neutral, and Not Accepting are distributed between accurate and incorrect forecasts. The balanced training data gives the model good class separation with many true positives and few false positives. The model learned patterns from all categories more evenly after oversampling corrected the class imbalance problem in the original dataset.

After this, the Random Forest and XGBoost models with and without SMOTE are evaluated to determine which predicts better. Each algorithm's ability to judge social media news user approval is measured by accuracy, precision, recall, F1-score, and ROC-AUC. Comparing models helps us determine which learning method is best for multi-class behavioral predictions using Likert-scale questionnaire data.

TABLE III
PERFORMANCE COMPARISON OF MACHINE LEARNING MODELS

Model	Accuracy	Precision	Recall	F1-Score	ROC AUC
With SMOTE					
Random Forest	96.30%	96.28%	96.30%	96.26%	99.87%
XGBoost	95.06%	95.07%	95.06%	94.97%	99.85%
Without SMOTE					
Random Forest	91.36%	91.84%	91.92%	91.72%	99.34%
XGBoost	90.12%	90.79%	91.08%	90.68%	98.70%

This study examined Random Forest and XGBoost prediction models with and without SMOTE. Random Forest trained on balanced data using SMOTE achieved the greatest results, with 96.30% Accuracy, 96.28% Precision, 96.30% Recall, 96.26% F1-Score, and 99.87% ROC AUC. Although trailing Random Forest, XGBoost with SMOTE achieved 95.06% Accuracy and 94.97% F1-Score. Model performance decreased without SMOTE, particularly for XGBoost, which had worse Precision (90.79%) and F1-Score (90.68%) due to class distribution imbalance. Data demonstrate class balancing enhances model prediction stability and reliability.

Following these findings, more research is needed to discover how each attribute affects classification. Explainable AI, SHAP (Shapley Additive Explanations), analyzes how each feature influences model predictions. SHAP identifies questionnaire variables that strongly influence social media news acceptance, neutrality, or rejection. The SHAP summary graphic for the best model (Random Forest with SMOTE) is shown in Figure 4, showing that certain traits dominate prediction while the rest have slight but significant effects.

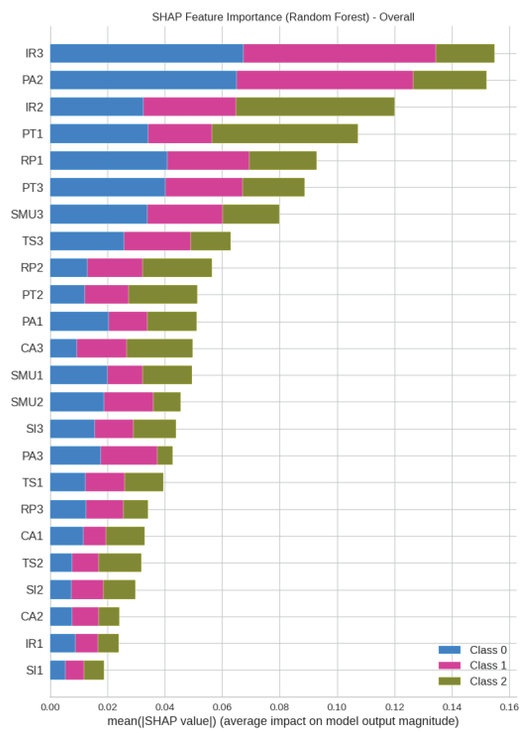


Fig. 4. SHAP Feature Importance (Random Forest)

Figure 4 shows SMOTE-developed Random Forest model SHAP feature significance results. The graphic shows how each questionnaire attribute predicts a respondent's social media news acceptance, neutrality, or rejection. IR3 (skepticism about spectacular news), PA2 (confidence in factual accuracy), IR2 (exposure to misinformation education), PT1 (receptivity to new information), and RP1 (dissemination of news without verification) are the most important characteristics in all three categories. These components have the greatest average effect on model output, demonstrating that cognitive resistance, perceived correctness, and behavioral preferences influence news acceptance decisions. Meanwhile, elements near the bottom of the ranking continue to participate, but with less impact, improving the model's predictive power. The SHAP study shows that the model's assessments match major behavioral and psychological characteristics, making categorization findings more transparent and interpretable.

V. Conclusion

This study studied how Instagram, X, TikTok, and Facebook users embrace, reject, or remain neutral about news. The purpose was to model and explain user acceptance using psychological, behavioral, and cognitive questionnaire variables using machine learning on 401 respondents.

The major results suggest that Random Forest and XGBoost can accurately classify news acceptance, especially when SMOTE fixes class imbalance. Random Forest with SMOTE performed best, with 96.30% accuracy and 99.87% ROC AUC. Though less successful, XGBoost did well. Performance without SMOTE was poorer, notably for XGBoost. Balanced data is crucial. SHAP analysis also found that being skeptical of sensational news (IR3), trusting facts (PA2), being exposed to misinformation education (IR2), being open to new information (PT1), and taking risks like sharing without checking (RP1) are the most important factors in user acceptance.

These findings emphasize the importance of behavioral and cognitive factors in digital information processing. The paper shows how ML and XAI may collaborate to improve news acceptance modeling transparency and interpretability, advancing research. These findings may help platform designers, legislators, and digital literacy efforts understand how psychological factors affect social media news susceptibility.

This study acknowledges limitations such self-reported questionnaire data, respondent bias, and just two machine-learning algorithms. SHAP simplifies things, but it doesn't explain why users act the way they do.

Next research should expand algorithms, integrate real-time behavioral data, and examine cross-platform news engagement patterns. A deeper look at demographic groupings and longitudinal data could assist us comprehend digital news behavior changes.

This study concludes that machine learning, data balance, and explainable AI create a reliable and understandable foundation for social media news acceptance. These solutions help people make wise decisions, make the internet more robust, and develop healthier information ecosystems for future generations as misleading information proliferate online.

VI. Acknowledgement

This work is supported by Information Systems Department, School of Information Systems, Bina Nusantara University, Jakarta, Indonesia and Research and Technology Transfer Office (RTTO) Bina Nusantara University. Data available at https://zenodo.org/uploads/17784579. DOI: 10.5281/zenodo.17784579. Authorship contribution statement Eka Miranda: Writing – review & editing, Methodology, Conceptualization. Mochammad Risyad Aukasyah Fatahillah: Writing – original draft, Software. Jeremia Johanes Mikhael Paulus: Writing – original draft, Software.

References

[1] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[2] E. Aimeur, S. Amri, and G. Brassard, "Fake news, disinformation and misinformation in social media: a review," *Social Network Analysis and Mining*, vol. 13, no. 30, 2023.

[3] T. Buchanan and V. Benson, "Spreading disinformation on facebook: Do trust in message source, risk propensity, or personality affect the organic reach of fake news?," *Social Media + Society*, vol. 5, no. 4, pp. 1–9, 2019.

[4] J. Roozenbeek and S. van der Linden, "Fake news game confers psychological resistance against online misinformation," *Palgrave Communications*, vol. 5, no. 65, 2019.

[5] P. Majerczak and A. Strzelecki, "Trust, media credibility, social ties, and the intention to share towards information verification in an age of fake news," *Behavioral Sciences*, vol. 12, no. 51, 2022.

[6] T. Samuel-Azran and T. Hayat, "Online news recommendations credibility: The tie is mightier than the source," *Comunicar*, vol. 27, no. 60, pp. 71–80, 2019.

[7] Y. Huang and L. Liu, "The impact of algorithm awareness on the acceptance of personalized social media content recommendation based on the technology acceptance model," *Acta Psychologica*, vol. 259, p. 105383, 2025.

[8] A. Ramadhana, B. Susetyo, and Indahwati, "Classification modelling of random forest to identify the important factors in improving the quality of education," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 11, no. 2, pp. 501–508, 2021.

[9] C. Li, C. Liao, X. Meng, H. Chen, W. Chen, B. Wei, and P. Zhu, "Effective analysis of inpatient satisfaction: The random forest algorithm," *Patient Preference and Adherence*, vol. 15, pp. 691–703, 2021.

[10] A. A. Santika, T. H. Saragih, Muliadi, D. Kartini, and R. Ramadhani, "Penerapan skala likert pada klasifikasi tingkat kepuasan pelanggan agen brilink menggunakan random forest," *JUSTIN: Jurnal Sistem dan Teknologi Informasi*, vol. 11, no. 3, pp. 405–413, 2023.

[11] A. K. S. Ong, Y. T. Prasetyo, N. Yuduang, R. Nadlifatin, S. F. Persada, K. P. E. Robas, T. Chuenyindee, and T. Buaphiban, "Utilization of random forest classifier and artificial neural network for predicting factors influencing the perceived usability of covid-19 contact tracing "morchana" in thailand," *International Journal of Environmental Research and Public Health*, vol. 19, no. 7979, 2022.

[12] I. Danuraga, R. Heriansyah, and L. W. Astuti, "Public opinion sentiment analysis of news trends using the random forest algorithm," *Journal of Computer, Information and Technology*, vol. 5, no. 2, pp. 1–20, 2025.

[13] J. Shin and S. Chan-Olmsted, "User perceptions and trust of explainable machine learning fake news detectors," *International Journal of Communication*, vol. 17, pp. 518–540, 2023.

[14] S. Ahmed and M. E. Rasul, "Social media news use and covid-19 misinformation engagement: Survey study," *Journal of Medical Internet Research*, vol. 24, no. 9, p. e38944, 2022.

[15] M. Saarela and V. Podgorelec, "Recent applications of explainable ai (xai): A systematic literature review," *Applied Sciences*, vol. 14, no. 8884, 2024.

[16] J. Cha, "Predictors of the credibility of social media as a news outlet: An examination of the influences of social media contacts, source perceptions, and media use," *International Journal on Media Management*, vol. 26, no. 1–2, pp. 68–93, 2024.

[17] S. Naveed, G. Stevens, and D. Robin-Kern, "An overview of the empirical evaluation of explainable ai (xai): A comprehensive guideline for user-centered evaluation in xai," *Applied Sciences*, vol. 14, no. 11288, 2024.

[18] M. L. Khan and I. K. Idris, "Recognize misinformation and verify before sharing: A reasoned action and information literacy perspective," *Behaviour & Information Technology*, 2019.

[19] C. Pelau, M.-I. Pop, M. Stanescu, and G. Sanda, "The breaking news effect and its impact on the credibility and trust in information posted on social media," *Electronics*, vol. 12, no. 423, 2023.

Digital Marketing Website for Gender Equality: Empowering Women Entrepreneurs in Labuan Bajo's Culinary Tourism

[1st]Maryani
Information Systems Department, School of Information Systems
Bina Nusantara University
Jakarta, Indonesia
yanie@binus.edu

[2rd]Trias Septyoari Putranto
Hotel Management, Faculty of Digital Communication and Hotel & Tourism
Bina Nusantara University
Jakarta, Indonesia
trias.putranto@binus.ac.id

[3rd]Wahyu Sardjono
Information Systems Management Department
Bina Nusantara University
Jakarta, Indonesia
wahyu.s@binus.ac.id

[4rd]Nora Fitriawati
Information Systems Department, School of Information Systems
Bina Nusantara University
Jakarta, Indonesia
nora.fitriawati@binus.ac.id

[5rd]Hasyiyah Karimah Adli
Faculty of Data Science & Computing
Universiti Malaysia Kelantan
Kelantan, Malaysia
hasyiya@umk.edu.my

***Abstract*—Digital marketing initiatives for MSMEs in the food and beverage sector contribute significantly to the growth of the tourism industry by increasing visibility, enhancing marketing performance, and encouraging culinary tourism. These initiatives empower local businesses to reach a wider audience, thereby attracting tourists and boosting the local economy. This initiative can also help revitalize local MSMEs in Labuan Bajo and Malaysia. Since the pandemic hit the world, MSME economic activity has declined, resulting in high unemployment, leaving women and housewives unable to contribute to their families' finances. With this research, we hope to support MSMEs in Malaysia and help women meet their families' needs. By understanding that transformational initiatives and skills can strengthen gender equality in the workplace, the role of technology as a catalyst in achieving this goal also needs to be emphasized. The increased use of technology is key to providing solutions to gender equality challenges across all levels of society. The research objectives are to assist tourists in finding nearby culinary and tourist attractions and to enhance women's skills and digital marketing training through a website provided in video format. To encourage entrepreneurial discussions through the website with other MSME entrepreneurs on specific topics through discussion forums; Improving access to equal employment opportunities for women through MSME business features can help the Indonesian and Malaysian governments increase foreign tourism. The online counseling development method can be carried out using the Rapid Application Development (RAD) method, and the results can be tested using the black box test method and SWOT analysis.**

***Keywords*— Digital Marketing, MSMEs, Food and Beverage Sector, Culinary Tourism, Gender Equality, Technology, Digital Skills Training, Rapid Application Development (RAD)**

I. INTRODUCTION

The growing use of the internet and digital technology in the globalization era has greatly impacted many sectors, including Micro, Small, and Medium Enterprises (MSMEs). By adopting digital technologies, MSMEs—such as those found in major Indonesian cities like Jakarta—can improve business growth and competitiveness.

In the tourism sector, MSMEs play a crucial role by strengthening local economies, showcasing cultural heritage, and offering unique products and services. Their involvement enhances both industry competitiveness and the overall tourist experience.

The existing food ordering and location system is still manual, leading to long queues, frequent errors, and difficulties for tourists in finding local specialties. To address these issues, a website-based food ordering system was developed to simplify access to MSME locations and available foods.

Additionally, based on participant feedback, recommendations were created to help hiring managers design more gender-inclusive software engineering job advertisements, supporting efforts to build a more gender-balanced SE workforce [1].

This research uses observation and interviews with customers and MSME owners in major cities to collect data and validate findings. The system is developed using the waterfall method, which includes the stages of analysis, design, implementation, testing, and maintenance. The website is built with JavaScript using the ReactJS framework, while PhpMyAdmin is used for database management.

System testing applies the black box testing method to ensure functionality, reliability, and proper handling of input and output. The resulting system allows tourists to order food online and access information about culinary options and tourist locations. Key features include login and registration, a homepage, menu pages, and information pages for food and tourist centers across countries such as Indonesia, Thailand, and Malaysia, as well as a feedback page.

A digital marketing website enhances service efficiency and improves access to culinary and tourism information for travelers. Strengthening this platform can further optimize its role in promoting local culinary sectors and supporting Micro, Small, and Medium Enterprises (MSMEs).

The problem formulation focuses on designing a culinary and MSME-oriented application that benefits tourists, supports local economies in Asian countries, provides detailed information on regional foods and tourism sites, and includes

features such as nearest culinary locations and popular tourist-visited spots.

The application scope includes:

Culinary Information – Offering quick access to the nearest culinary spots for tourists. **MSME Collaboration** – Providing a platform that connects local MSMEs and industries to strengthen the regional economy.**Tourist Location Information** – Delivering accessible details on tourism sites to help travelers navigate easily.**Analytical Dashboard** – Displaying visitor statistics, traffic sources, page performance, and conversion actions to evaluate website effectiveness.

II. LITERATURE REVIEW

Previous research is used to analyze and enrich the research discussion and as a difference with the current research. There are several previous research journals on women's empowerment in gender equality obtained from several sources. The journals include :

The first study [2] applies the K-Means algorithm, supported by RapidMiner, to classify job seeker data based on education level. The evaluation uses the Davies-Bouldin Index (DBI), where lower values indicate better cluster similarity. The method successfully categorizes data, but the analysis is limited to education-based clustering.

The second study [3] explores teachers' perspectives on online discussion forums through semi-structured interviews analyzed with Grounded Theory. The findings recommend improving forums by integrating gamification, multimedia elements, and more attractive interfaces. However, the results may still carry subjectivity despite methodological efforts to reduce it.

The first study develops a **Cyber Counselling application** to address limitations in traditional counseling related to time, place, and cost. Using the **Rapid Application Development (RAD)** method, the system was created and then tested using **Blackbox Testing**, which showed that all features functioned properly and are feasible for use.

The second study examines how **game-based digital quizzes** can enhance student engagement in online learning. Using a **Grounded Theory Approach**, the researchers found that students prefer interactive and enjoyable platforms like **Kahoot!**, which offer music, bonus points, and competitive elements. These features increase motivation, focus, and participation, especially in institutions that are new to online teaching.

The study evaluates an **Automatic Grading Algorithm** designed to assess user participation in MOOC discussion forums. The algorithm analyzes forum messages based on content quality, user impact, and activity, using weighted metrics. Forum posts are labeled on a scale from **1 (low quality)** to **4 (high quality)**. Although the algorithm still produces lower accuracy compared to expert manual grading, its performance is considered reasonably good and shows potential for supporting automated forum evaluation.

From the whole journal, it can be concluded that the most important indicators for a quiz feature are gamification and providing interesting multimedia elements such as images, music, and points on the quiz to maintain user motivation in answering quiz questions. Then, the development of online counseling can be done using the Rapid Applica on Development (RAD) method and the results can be tested using the blackbox testing method.

Several algorithms are utilized to develop system features, including an **Automatic Grading algorithm** to evaluate the quality of forum posts and **clustering algorithms** to categorize job seekers based on specific criteria. These algorithms help automate assessment processes and organize data more effectively.

III. METHODS

To determine the requirements for developing the women empowerment application, four data collection methods were used [5][6]:

1. **Questionnaire** – A closed-ended questionnaire was distributed to obtain structured and relevant data from a wider group of respondents, helping identify user needs for the web application.
2. **Interview** – Structured interviews were conducted with selected individuals to gain deeper insights, contextual understanding, and detailed perspectives related to the research.
3. **Observation** – Direct observation of real situations and social phenomena was carried out to understand the importance of certain features and gather data needed to design the culinary application.
4. **Literature Review** – Relevant literature from journals, books, theses, and research reports was analyzed to gather supporting information, identify knowledge gaps, and guide the application's development based on previous studies.

A. Application Design Methods

The Prototype Process Model is used in this application's development because user requirements have been identified, allowing the team to design features that align with user needs. Early prototypes enable users to provide feedback, and necessary adjustments can be made immediately. This iterative approach reduces errors in later stages and ensures the final system meets expectations. Prototyping allows continuous refinement of requirements, strengthens collaboration between users and developers, and improves system quality through ongoing feedback.

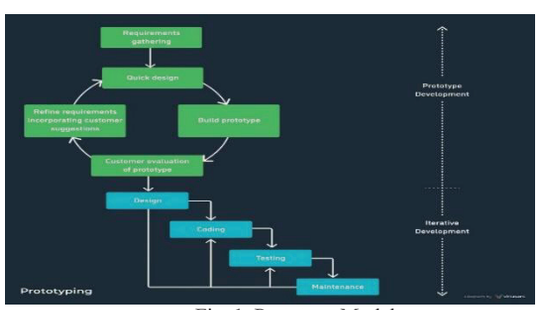


Fig. 1. Prototype Model

1) Initial Requirements

At this stage, we analyze the needs and expectations of users of the desired application. Needs analysis is done by distributing questionnaires, interviews to define the scope of

application needs to be made. At this stage, we also identify the components needed by our application, such as hardware, software, and network. This needs analysis will be the foundation in making prototype until the application development later.

2) Design

After the user's needs for this application have been well analyzed, we start designing the prototype according to the conditions set by the user.

3) Prototyping

The next stage is the creation of a prototype user interface. The user interface prototype designed is a reflection of the applica-tion that will be developed later. This prototype is expected to be a detailed description for users to see the application design.

4) Customer Evaluation

At this stage, the prototype will be evaluated by the user to get feedback on the prototype to identify whether this prototype is in accordance with user needs and aspects that need to be improved or developed.

5) Review s Evaluation

At this stage, the prototype will be evaluated again in accordance with the input received from the user. This process will be re-peated continuously until the prototype made is in accordance with the criteria and user needs.

6) Development

The prototype is in accordance with the criteria and user needs, the prototype will be used as a guideline and reference for appli-cation design and development.

7) Test

After designing the application, testing is carried out to ensure its quality, identify errors, and verify that it operates smoothly according to user needs using RAD. The article *"SWOT Analysis is The Most Popular Management Strategy Tool in the Tourism Sector"* [4] explains why SWOT is widely used in tourism. The study shows that SWOT helps tourism organizations analyze internal and external factors by identifying strengths (e.g., destination appeal, human resource quality, promotional strategies), weaknesses (e.g., limited digital infrastructure, poor management systems), opportunities (e.g., policy support, tourism trends, technological innovation), and threats (e.g., regulatory changes, global crises, competition among destinations) [8][9].

8) Maintain

After the application is operational, we will continue to carry out the maintenance process according to user needs. Maintenance is carried out so that the application gets updated features in accordance with evolving technology and avoids errors. Through this maintenance process, the application continues to provide the best quality for users.

IV. RESULTS

Tourism development is a process that aims to increase the attractiveness of tourist and improve infrastructure and services that support the tourism sector. Tourism development in Labuan Bajo has been carried out with various strategies and activities that focus on developing tourism destinations, developing the tourism industry, and active community participation. The Labuan Bajo Tourism Office has developed tourism destinations with a focus on developing tourist objects and attractions. Tourism development can be seen from five components based on Cooper's opinion Accessibility, Amenities, Institution, Available Packages, and Ancillary Services, namely:

A. Use-Case Diagram



Fig. 2. Use Case Diagram

This Use Case Diagram visually maps the functional requirements of the Digital Marketing Website for the Labuan Bajo Women's Culinary Collective. It illustrates how various users (Actors) interact with the core features (Use Cases) within the boundary of the website system.

1) System Boundary

The central box represents the "Labuan Bajo Culinary Collective Website System." All functionalities (ovals) within this boundary are key operations the digital platform must perform.

2) The Actors

Four primary actors interact with the system, each having distinct goals:

a) Tourist/Visitor: The end customer seeking culinary experiences and information. Their goal is discovery, booking, and providing feedback.

b) Women Entrepreneur: The business owner providing the culinary service. Their goal is self-management, visibility, and fulfilling orders.

c) Website Administrator (Admin): The staff managing the system, content, and membership. Their goal is operational integrity and platform growth.

d) Partner/Donor: External stakeholders (NGOs, hotels, individuals) providing financial or collaborative support. Their goal is impact and reporting.

3) Core Use Case

a) Tourist/Visitor Functions (Discovery & Transaction)

This group focuses on the user journey from discovery to booking:

- Search Culinary Businesses: Allows users to find and filter listings (restaurants, classes, tours).

- View Entrepreneur Profile: Allows users to access detailed information, personal stories, and menus.
- Book Culinary Experience: The primary transaction feature, facilitating reservations or inquiries for services.
- Share Review / Testimonial: Enables user-generated content (UGC) for trust-building and social proof.
- Donate / Support Program: Provides a pathway for non-transactional support for the gender equality mission.

b) Women Entrepreneur Functions (Management & Content)

This group focuses on the self-service tools provided to the business owners [10][11]:

- Manage Business Profile: Allows the entrepreneur to update their hours, pricing, photos, and menu.
- View Booking Requests: The interface for receiving and confirming reservations made via the website.
- Upload Story & Recipes: Enables the entrepreneur to contribute content (e.g., blog posts) to enhance their visibility.

c) Administrator Functions (Oversight & Maintenance)

These functions ensure the site runs smoothly and securely:

- Manage Website Content (Blog): Creating and moderating all non-business-specific content, such as articles and news.
- Manage Entrepreneur Data: Vetting, onboarding, and updating the master list of women-owned businesses.
- Manage Partner & Donations: Tracking all incoming financial and collaborative support.
- Manage User Access: Setting up and maintaining user roles and system security.

4) Relationship

The diagram uses two key UML relationships:

- <<include>>: The Book Culinary Experience use case includes the View Entrepreneur Profile use case. This signifies that viewing the profile is a necessary, integral step in the booking process.
- Association Lines: Straight lines connect the actors to the functions they use. For example, the Tourist/Visitor is connected to five separate use cases, reflecting their broad interaction with the public- facing areas of the site [12][13].

B. Class Diagram

This UML Class Diagram outlines the static structure of the Labuan Bajo Women's Culinary Collective digital marketing system. It is organized into four main packages representing core system functions: User Management, Business Directory, Booking & Review, and Content/Support.

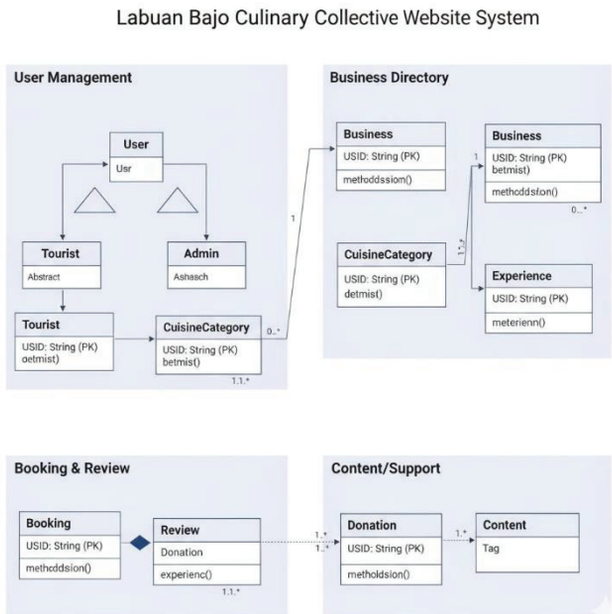


Fig. 3. Class Diagram

1) User Management Package

This package manages all human interactions with the system, utilizing Generalization (Inheritance).

a) User (Abstract)

The base class defining common attributes for anyone logging into the system. Key Attributes: userID (Primary Key), username, passwordHash, email, and role (to manage permissions).

b) Tourist (inherits User)

Tourist is the child class of user represents the customer. Key Methods: searchBusinesses(), submitReview(), makeBooking().

c) Entrepreneur (inherits User)

Represents the woman business owner. Relationship: Has a 1:1 association with the Business class, managed via the foreign key businessID. Key Methods: manageBusinessProfile(), viewBookings().

d) Admin (inherits User)

Represents the project staff with elevated control. Key Methods: manageContent(), manageUsers(), approveBusiness().

2) Business Directory Package

This package structures the core listings that drive culinary tourism.

a) Business

Represents the physical or service-based culinary venture (e.g., a warung, a catering service, or a class location).

- Key Attributes: businessID (PK), name, description, and geographical data (latitude, longitude) for map integration.

b) Experience

Represents a specific, bookable service offered by a businesss. Key Attributes: experienceID (PK), type (Enum: Class, Tour, Catering), price, and durationHours.

3) Booking & Review Package

This package handles all transactional and feedback data.

a) Booking

Represents a completed or pending reservation made by a Tourist. Key Attributes: bookingID (PK), bookingDate, numGuests, and totalPrice.

b) Review

Represents feedback submitted by a Tourist about a specific Business. Key Attributes: rating (Integer 1-5), comment, and reviewDate. Content/Support Package

This package manages the website's marketing, storytelling, and educational assets. Key Attributes: contentID (PK), type (Blog, Story, Recipe), title, and publishDate.

V. DISCUSSION

A. Blackbox Testing Methods

The Blackbox Testing Method was applied to evaluate the functionality and usability of the Gender Equality Digital Marketing Website, which aims to promote and empower female entrepreneurs in the culinary tourism sector in Labuan Bajo. This testing approach focuses on verifying that the website's outputs match the expected results for given user inputs, without analyzing the internal code or structure. The method ensures that all features perform correctly from the end-user's perspective, particularly for tourists, entrepreneurs, and visitors interested in Labuan Bajo's culinary tourism.

1) Testing Scope

The Blackbox Testing covered several functional areas of the website as follows:

a) Homepage and Navigation

- Input: User clicks on navigation menus such as Home, About, Entrepreneurs, Culinary Products, Contact Us.
- Expected Output: Each menu directs the user to the correct page, and all links are functional.

b) Entrepreneur Profile Page

- Input: User selects a female entrepreneur's profile.
- Expected Output: The system displays complete profile information (name, business description, location, product images, and contact details).

c) Culinary Product Showcase

- Input: User views a list of culinary products or filters them by category (e.g., local food, snacks, beverages).
- Expected Output: Relevant product details and images are displayed accurately.

d) Online Marketing or Booking Feature

- Input: User fills in the booking or inquiry form to purchase or collaborate with a female entrepreneur.
- Expected Output: Form submission is successful, and a confirmation message or email appears.

e) Feedback and Testimonial Section

- Input: User submits feedback about the website or specific entrepreneurs.
- Expected Output: The message is sent successfully, and a thank-you notification appears.

f) Social Media Integration

- Input: User clicks on social media icons (Instagram, Facebook, YouTube).
- Expected Output: The respective social media pages open correctly in a new tab.

g) Responsiveness and Accessibility

- Input: User accesses the website from different devices (smartphone, tablet, desktop).
- Expected Output: Website layout and functions adjust correctly to each screen size without display errors.

TABLE I. TEST RESULT SUMMARY

No	Tested Feature	Expected Output	Result	Status
1	Homepage & Navigation	Each menu opens correct page	As expected	Passed
2	Entrepreneur Profiles	Information displayed correctly	As expected	Passed
3	Culinary Product Showcase	Relevant products shown	As expected	Passed
4	Booking/Inquiry Form	Confirmation message shown	As expected	Passed
5	Feedback Form	Feedback sent successfully	As expected	Passed
6	Social Media Integration	Links open in new tab	As expected	Passed
7	Mobile Responsiveness	Layout adapts properly	As expected	Passed

Table 1. Online marketing of the Labuan Bajo Tourism website operates under a Business to Customer (B2C) model, where tourists can directly purchase tour packages through the Visit Labuan Bajo Tour website. Online promotion is prioritized over printed brochures because it is more cost-effective and easier for consumers to access. The website also simplifies the payment and confirmation process, ensuring a smooth online transaction experience.

With the rapid growth of technology and the increasing number of competing travel agents in Labuan Bajo, Visit Labuan Bajo Tour continuously improves its online strategies to remain competitive. A SWOT analysis was conducted to identify the strengths, weaknesses, opportunities, and threats related to marketing tour packages through online media on the Visit Labuan Bajo Tour platform, as summarized in Table II.

TABLE II. SWOT TABLES

STRENGTH (S)	• The tour packages marketed through online media are quite diverse. • The images displayed on the website for each tour package are very attractive.

	• Marketing costs are cheaper than using print media. • Tourists have easier access to complete travel information. • E. Accessible to everyone in the world.
WEAKNESSES (W)	• Tourists find it difficult to request special tour packages. • Sometimes complaints from tourists arise because what they expect is sometimes different from reality. • The price of the tour package being marketed is Rp. • Sometimes system errors occur.
OPPORTUNITIES (O)	• Faster updating or adding new tour packages • Attractive and diverse tour package packaging will attract tourists to make reservations • Accompanied by quality products and services, these tour packages will be sold according to target. d. This website can be accessed by everyone
THREATS (T)	• Competition is increasingly challenging in promoting tour packages online. • Ordering products through online media can falsify tourist identities. • Price competition often occurs with other travel agents. • There is a crisis of tourist confidence due to several online sites committing fraud and tourist fear.
Srategy SO	• Utilize various attractive tour packages and packaging to increase tourist interest • Low package prices must be accompanied by product quality, so that sales are right on target • Increase promotions through online media so that more tourists access the website • Display complete information about tour packages
Strategy WO	• Provide new options for tourists who want to determine the tour package according to their wishes • Information posted on the website must be factual • Use a good quality internet connection system, in order to minimize the occurrence of system errors
Strategy ST	• Variation of tour packages and change the appearance of the website to be more attractive and optimal • Ensure to return the identity provided by the correct traveler every time the traveler books a tour package through the website • Provide accurate information and do not exaggerate or reduce the existing facts
Strategy WT	• Improving the quality of products and services to attract tourists to book the tour package • For domestic tourists, special prices are given in rupiah

The Labuan Bajo tourism website effectively uses the internet as its main promotional and marketing medium, supported by diverse tour packages with appealing visuals. SWOT analysis highlights strong advantages but also identifies weaknesses, such as difficulty in handling customized package requests and occasional system errors. Threats include strong competition, identity fraud risks, price rivalry, and reduced tourist trust due to online scams.

To improve performance, Visit Labuan Bajo Tour should offer attractive and affordable packages, maintain high service quality, ensure complete and reliable website information, and optimize the site to prevent errors. Although online media is the primary marketing channel, print and electronic media should still be used to broaden promotional reach.

VI. CONCLUSION

The rapid development of technology has intensified competition among travel agencies in Labuan Bajo. To stay competitive, Visit Labuan Bajo Tour continues to refine its marketing strategies. A SWOT analysis revealed that online media is highly effective for promoting and selling tour packages, offering many advantages that positively influence the company's performance. However, continuous improvement is necessary to reduce weaknesses and address threats from competitors.

As part of its digital transformation efforts, Visit Labuan Bajo Tour developed a comprehensive tourism website. The website provides detailed information on tour packages, destinations, accommodations, and travel guides, along with an integrated online booking system for easier reservations. Designed with a user-friendly and visually appealing interface, it uses high-quality images and videos to attract tourists. Additional features such as customer testimonials, contact details, and social media links help strengthen user engagement and trust.

Overall, the website aims to enhance the company's digital presence, improve customer accessibility, and promote Labuan Bajo tourism more effectively to a global audience.

REFERENCES

[1] Kanij, T., Grundy, J., Jennifer. (2024). Enhancing understanding and addressing gender bias in IT/SE job advertisements. Journal of Systems and Software Volume 217, November 2024, 112169

[2] Sembiring, M., Jufrizen, & Tanjung, H. (2021). Efek Mediasi Kepuasan Kerja pada Pengaruh Motivasi Dan Kemampuan Kerja Terhadap Kinerja Pegawai. MANEGGIO : Jurnal Ilmiah Magister Manajemen Vol. 4 No.1, 131-144

[3] Lima, T. et al..(2020). The Emotional Impact of Coronavirus 2019-nCoV (New Coronavirus Disease). Psychiatry Research, 287, Article ID: 112915. https://doi.org/10.1016/j.psychres.2020.112915

[4] Alim & Wening. (2021). A Review of the Use of Strategic Management Tools in the Health Sector. DOI:10.37506/ijfmt.v15i4.16834 Corpus ID: 239701733

[5] Anshori Muslich, & Sri Iswati. (2019). Metodologi Penelitian Kuantitatif. Surabaya: Airlangga University Press

[6] Abdullah. (2015). Metodologi Penelitian Kuantitatif. Yogyakarta: Aswaja Pressindo

[7] BPS. (2023). Keadaan Angkatan Kerja di Indonesia 2023. Badan Pusat Statistik.

[8] Dwi, E. W. (2017). Peran Ganda Perempuan Dan Kesetaraan Gender. Aspirasi: Jurnal Masalahmasalah Sosial, 8(2), 207–222.

[9] Hair, J. F. J., et al. (2019). Multivariate Data Analysis. Annabel Ainscow. International Journal for Multidisciplinary Research (IJFMR), Gender Inequality and Harassment at Workplace, E-ISSN: 2582-2160

[10] Lau, et al. (2021). Moving From Problems to Solutions: A Review of Gender Equality Interventions at Work Using an Ecological Systems Approach. DOI:10.31234/osf.io/cy63n

[11] Scholastica, G. (2018) Rerata Penghasilan Perempuan Masih Jauh di Bawah Gaji Laki- Laki. Tirto.Id

[12] Yin, et al. (2023). Analysis of Gender Inequality in Workplace based on Maternity Policies in First-tier Cities in China - Shanghai as an Example. The International Conference on Interdisciplinary Humanities and Communication Studies DOI:10.54254/2753-7048/5/20220726,

[13] Vogt, V. (2015). Is it better to replace my washing machine or repair it?. Atlanta: HowStuffWorks.

Design of Reconfigurable Modulation System for Selective Cellular Interference with SDR

1st Diky Arif Hidayat
School of Electrical Engineering
Telkom University
Bandung, Indonesia
dikyarifhidayat07@gmail.com

2nd Eka Nadiyah Purwanianti
School of Electrical Engineering
Telkom University
Bandung, Indonesia
purwaniantynadia@gmail.com

3rd Daniel Alden Susetyo
School of Electrical Engineering
Telkom University
Bandung, Indonesia
danielaldens10@gmail.com

4th Amalia Muhammad Jasmin
School of Electrical Engineering
Telkom University
Bandung, Indonesia
amalia.m.j.890@gmail.com

5th Budi Prasetya
School of Electrical Engineering
Telkom University
Bandung, Indonesia
budiprasetya@telkomuniversity.ac.id

***Abstract*—Conventional signal interference systems are often constrained by the rigidity of hardware architecture, which limits operational flexibility and increases cost inefficiency. This research aims to overcome these limitations through the development of an adaptive signal modulation system based on Software Defined Radio (SDR) that replaces the role of static analog components with reconfigurable software configurations. The method used involves the implementation of HackRF One SDR controlled through the GNU Radio framework to perform interference emissions on strategic cellular frequency bands. Performance evaluation was carried out by comparing the real time output power profile with the technical specifications of the device and conducting a comparative analysis with conventional Digital Cellular System (DCS) modules to assess bandwidth adequacy and radiation intensity. The experimental results show that although the SDR configuration offers high spectral agility, there are deficiencies in transmission power and bandwidth for large scale cellular coverage, requiring the integration of external amplification. However, validation testing confirmed the system's effectiveness in significantly reducing the Reference Signal Received Power (RSRP) value on the target device compared to the received power level from the Base Transceiver Station (BTS). Overall, this approach successfully improved the architectural efficiency and functional flexibility of the jammer modulation system in mitigating signals on the cellular frequency spectrum.**

***Keywords*— software defined radio, hackrf one, gnu radio, signal interference, spectral agility.**

I. Introduction

Conventional signal blocking systems, such as fixed hardware based jammers, operate by injecting electromagnetic interference through noise modulation on specific frequency bands [1], [2]. Although effective in static scenarios, this methodology has significant operational efficiency limitations due to low hardware flexibility [3]. The use of oscillators and analog filter circuits with fixed frequencies limits the scalability of the system in dealing with dynamic and congested spectrum environments. In addition, dependence on rigid physical components makes the system incapable of adapting to frequency hopping techniques or agile modern communication protocols [4], [5].

Conventional hardware based modulation architectures suffer from frequency rigidity due to non reconfigurable local oscillators, leading to inefficiencies in multi sectoral jamming systems [3]. To overcome this, an efficient system must achieve a precise balance to generate selective interference [6]. This ensures that spectral disruptions are targeted strictly at specific mobile communication protocols, thereby preventing unwanted spillover that compromises the surrounding spectrum integrity [7].

While Software Defined Radio (SDR) provides modulation flexibility, jamming effectiveness is often constrained by the integration complexity between software layers and front end architectures [8]. A primary challenge involves synchronizing components to generate precise, coherent noise without hardware redundancy across multiple cellular bands [9]. Misalignment between software control and RF output typically degrades spectral precision and wastes power, necessitating integrated designs that optimize operational efficiency and blocking accuracy within dynamic environments [10], [11], [12].

To overcome the limitations of conventional hardware rigidity, this research proposes a Software Defined Radio (SDR) based system design that is capable of operating dynamically on cellular frequency bands, namely 1800 MHz (LTE Band 3), 2100 MHz (LTE Band 1/IMT), and 2300 MHz (LTE Band 40) [13]. The inherent flexibility of the SDR architecture allows a single processing unit to emulate the functions of multiple analog devices through real time reconfiguration of parameters at the software layer [1], [14]. Through an integrated multi frequency approach, this system is capable of scanning and blocking the desired spectrum without the need to add hardware modulation components for each frequency band, thereby significantly improving cost efficiency and operational flexibility [15]

II. Theoretical Background

The development of an effective cellular frequency scrambling system requires a deep understanding of the basic principles of radio frequency (RF) engineering and signal processing. This section provides an in depth explanation of the theoretical framework underlying the proposed GNU Radio based architecture, with a particular emphasis on jamming system principles and the core principles of SDR.

A. Jammer System Principles

Jamming operates based on the basic principle of electromagnetic interference, where interference signals are deliberately transmitted to reduce the SINR at the target receiver [2]. Theoretically, the effectiveness of this process is determined by the jammer to signal ratio (J/S), which determines the level of communication interference by drowning out legitimate carrier signals with high density noise or modulated interference. In physical implementation, as illustrated in Figure 1, the jamming signal is synthesized through SDR and then amplified through a power amplifier to achieve the target distance required for effective spectral suppression. This amplified signal is then transmitted through a planar collinear antenna array, which serves to focus the interference energy onto the targeted spatial domain. This forces a drastic reduction in the RSRP value and breaks the connection between the base station (BS) and the User Equipment (UE) in the specified coverage area.

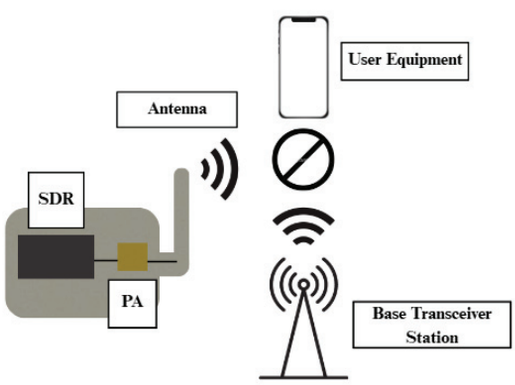


Fig. 1. Jammer System Physical Implementation

B. Software Defined Radio with GNU Radio

Software Defined Radio (SDR) represents a transformative paradigm shift by transitioning from rigid hardware centric architectures to versatile, software implemented functionalities supported by frameworks like GNU Radio [11]. This technology leverages software based protocols to achieve high reconfigurability, where incoming analog signals are digitized via high speed Analog to Digital Converters (ADC) and processed using Field Programmable Gate Arrays (FPGA). As illustrated in the RF front end architecture in Figure 2, the system utilizes a MAX5864 ADC/DAC interfaced with an NXP LPC43xx microcontroller, employing synthesizers and mixers for frequency upconversion. Although the USB 2.0 High Speed interface constrains the operational bandwidth to approximately 20 MHz, this capacity is sufficient for targeted cellular bands at 1800 MHz, 2100 MHz, and 2300 MHz. By executing baseband and intermediate frequency (IF) processing in the digital domain, the protocol stack and physical layer can be dynamically updated on universal hardware to support diverse transmission formats [4].

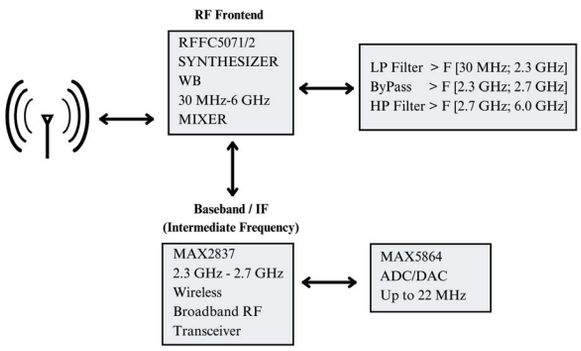


Fig. 2. HackRF One front end/baseband

C. RF Architecture and Signal Processing Flow

The communication architecture in this system utilizes the SDR principle to efficiently connect analog and digital domains through flexible hardware integration. Analog radio frequency signals interact with the system through an antenna that is directly connected to the RF frontend module via a coaxial connector, where the processing path is divided into reception and transmission mechanisms to maintain data integrity [4]. At the digital frontend stage, high speed analog to digital and digital to analog converters convert the signal, enabling baseband processing to be performed computationally within the central processing unit (CPU) or FPGA in Figure 3. At this layer, the signal undergoes sample reduction and sample rate conversion before being forwarded to the host device via a USB interface to ensure that the data stream remains synchronized. A key feature of this design is its ability to implement various communication protocols and physical layer functions without changing the physical components. Through this software abstraction, technical parameters such as channel filtering and modulation can be dynamically adjusted, providing a high degree of scalability for experimental needs across a wide range of cellular spectrums.

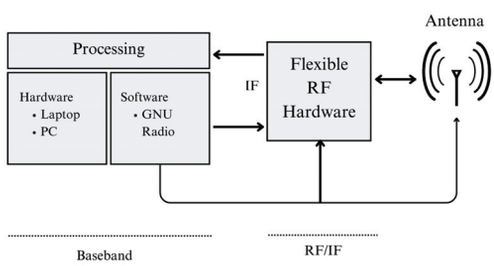


Fig. 3. SDR Block Flow Architecture

III. METHODOLOGY

This section presents the design methodology for our GNU Radio based jammer system, detailing the system architecture, operational workflow, and simulation approach. The design integrates SDR technology with triple frequency antenna design to achieve effective jamming across 1800 MHz, 2100 MHz, and 2300 MHz cellular bands.

A. Hardware reconfigurability of SDR

The proposed system architecture integrates SDR technology with external power amplifier components to generate effective signal interference in the cellular frequency band. In this system model, the software defined radio unit acts as the processing core that synthesizes the base signal through a sawtooth generator mechanism, which is then forwarded to a voltage controlled oscillator (VCO) block for frequency modulation, wich can be seen in Figure 4 [3]. The modulated signal is then fed to a power amplifier unit to increase the intensity of electromagnetic radiation to the power level required to overcome the threshold of legitimate signals at the target receiver. The final stage of this system is the transmission of interference energy through a tri band antenna that has been specifically optimized to operate at frequencies of 1800 MHz, 2100 MHz, and 2300 MHz. This integrated configuration allows the system to precisely focus interference energy on a specific spatial domain, thereby degrading the quality of cellular communication services in a controlled manner without requiring additional hardware for each different frequency band.

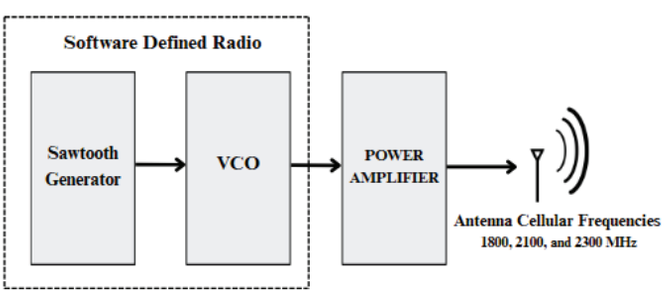


Fig. 4. GNU Radio based Jammer System Architecture

B. Key Attributes SDR

The technical specifications of Software Defined Radio (SDR) are basic parameters that determine the integrity of the output signal and the operational limits of the device [14]. Identifying these specifications allows the determination of the maximum performance limits and inherent limitations of the device used, which can be seen in Table I. The accuracy of output power measurements is highly dependent on the technical profile of the SDR, which directly affects the validity of empirical data. Therefore, comprehensive mapping of instrument specifications, including oscillator stability and receiver sensitivity, is crucial to ensure that performance analysis results reflect actual physical phenomena, rather than artifacts of hardware limitations.

TABLE I
SPECIFICATION HACKRF ONE

Specification Parameters	Description
Data Bandwidth	20 MHz
Sampling Accuracy	8 Bit
Operating Frequency	1 MHz to 6 GHz
Sampling Rate ADC/DAC	2 Msps to 20 Msps
Power Supply/Data Port	USB 2.0

C. GNU Radio Configuration

The SDR framework facilitates a paradigm shift by implementing conventional hardware functions, such as signal generators and voltage controlled oscillators (VCOs), into highly reconfigurable software modules [3]. This flexibility allows for dynamic parameter adjustments across a broad frequency range. however, as illustrated in Figure 5, the hardware faces inherent limitations in executing simultaneous multi band transmissions. To address this, the proposed architecture utilizes three discrete SDR units to operate across distinct frequency bands, integrated with external Power Amplifiers (PA) to compensate for the SDR's low output power [8]. This configuration ensures the transmission power is sufficient to effectively degrade the signal quality between the Base Transceiver Station (BTS) and mobile devices within the targeted area.

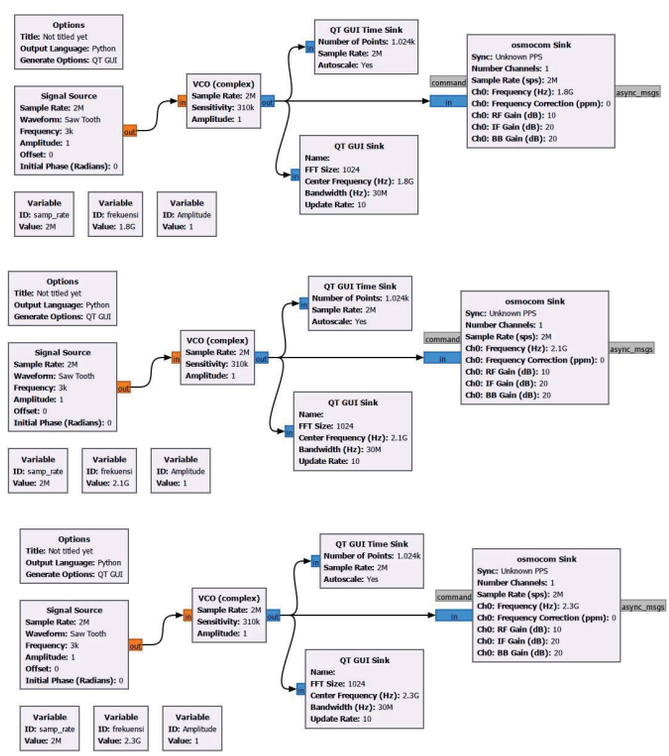


Fig. 5. Block Configuration GNU Radio

D. Signal Processing Workflow

The signal processing mechanism in the proposed architecture adopts a digital wave synthesis approach configured through the GNU Radio environment to generate coherent interference. The initial stage involves a Signal Source block that generates a periodic reference signal as the basis for modulation to create a noise spectrum. This signal is then integrated with a Voltage Controlled Oscillator or VCO block to transform amplitude variations into agile frequency shifts, enabling the system to effectively cover the targeted bandwidth. To ensure process integrity and avoid data synchronization failures between software and hardware, the signal flow is continuously visualized using the QT GUI Sink block, which displays the time domain and frequency domain simultaneously [7]. As the final stage in the processing chain, the Osmocom Sink block acts as a hardware abstraction that converts the digital sample stream into physical instructions for the SDR device. Through synchronization of gain and sample rate parameters in this block, the device is able to transmit interference signals with high precision on the 1800 MHz, 2100 MHz, and 2300 MHz cellular frequency bands to achieve optimal degradation of service quality on the target device.

E. Measurement and Validation Setup

System validation procedures are performed to ensure that interference signals from GNU Radio and HackRF One meet technical parameters prior to field testing. The initial stage uses a spectrum analyzer to verify the accuracy of the center frequency and bandwidth on three target bands, in order to ensure that noise energy is concentrated without spectral leakage. Next, power measurements are performed to calibrate the gain and ensure that the amplifier output remains within the linearity limits to prevent harmonic distortion. Blocking effectiveness is then validated by monitoring RSSI and SINR degradation on mobile devices to determine the Interference to Signal Ratio (ISR) threshold capable of breaking synchronization [6]. All measurement data is rigorously cross referenced against the manufacturer's SDR technical specifications and datasheet constraints to ensure the system operates at its optimal capacity while maintaining signal integrity and preventing hardware saturation during prolonged operation.

IV. Experimental Result

This chapter evaluates the SDR signal output measurements resulting from the configuration to validate the specified frequency performance and power output for jamming. The results demonstrate the jammer effectiveness in blocking cellular signals, while confirming the functional integration between the configuration system and the transmitter device.

A. Software Defined Radio

Based on the test results after configuration at the cellular frequency, SDR proved capable of transmitting random signals without information in Figure 6. However, due to bandwidth and gain limitations, the random signal power generated was unable to cover the overall cellular signal power at the target frequency [13]. As a result, this system is only capable of functioning in the simulation stage and is not yet effective for jamming real cellular devices. In addition, the limitation of Osmocom Sink, which only supports one transmission frequency at a time, requires the execution process to be carried out three times for each configuration.

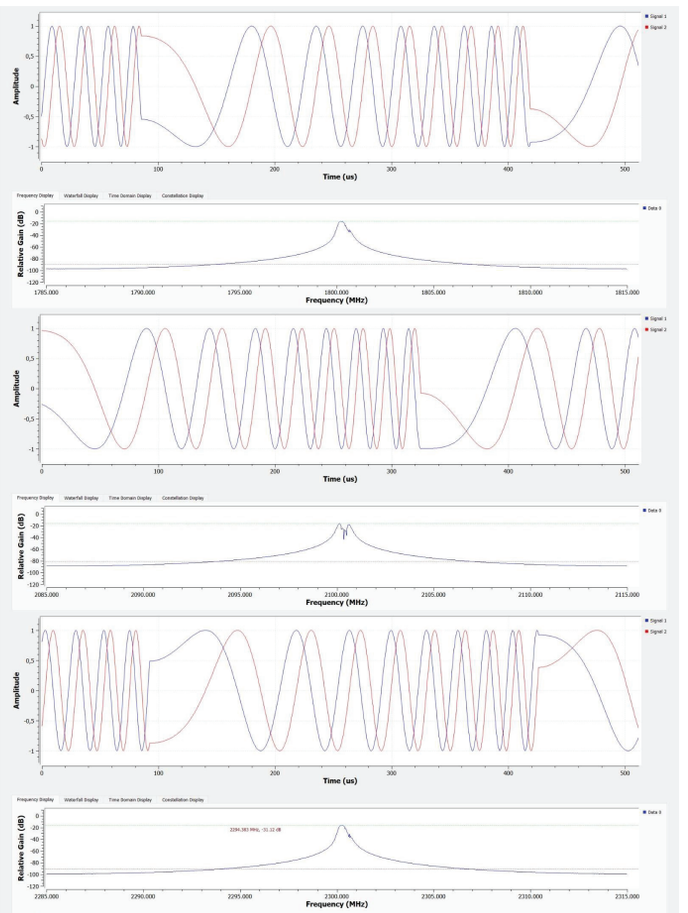


Fig. 6. SDR Configuration Output using GNU Radio

B. Transmission Power Analysis of SDR on a Spectrum Analyzer

Transmission power analysis of the system was performed using a spectrum analyzer to verify the accuracy of frequency modulation and output signal intensity generated by the software defined radio device. Based on the measurement results displayed on the instrument, the system was successfully configured to transmit signals in the target frequency range, namely 1800 , 2100, and 2300 MHz, proving the flexibility of the device in mapping the cellular spectrum according to software instructions [8]. Although frequency synchronization has been achieved with high precision, the device's technical profile reveals inherent limitations related to power density and available operational bandwidth in Figure 7. Observations on the visual spectrum confirm that the pure output power of the SDR hardware is still below the threshold required to effectively suppress signals on a large cellular network [12].

These bandwidth and power level limitations indicate that the use of the device in real world signal interference scenarios requires additional integration with an external power amplifier unit to achieve an optimal interference to signal ratio without compromising spectral integrity around the operating frequency.

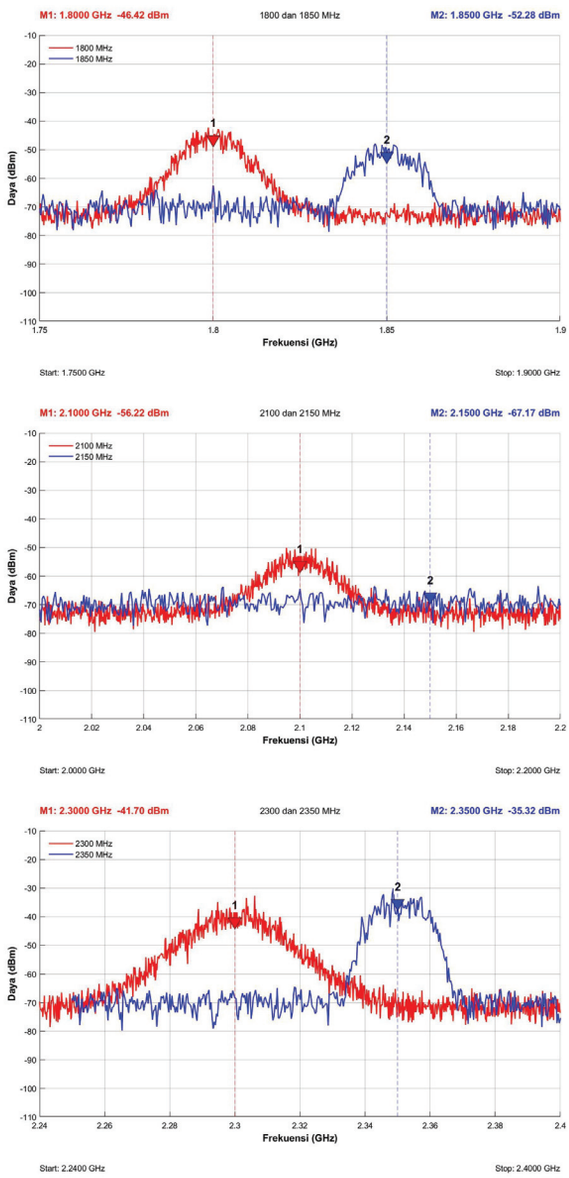


Fig. 7. Power and Bandwidth Comparison Graph on the Spectrum Analyzer Output

C. Evaluation of DCS Modul Frequency Band Transmission Power

This experiment was systematically designed by adjusting the distance between the transmitter antenna (Tx) and receiver antenna (Rx) within a range of 1 to 5 meters, with two tests conducted at different center frequencies on each band [6]. Given the inherent limitations of the Software Defined Radio (SDR) platform, such as narrow bandwidth and low output power, this study integrated a special external jammer module as the main signal source to ensure frequency stability and high spectral power density [10]. On the receiver side, the LibreVNA device is operated in Spectrum Analyzer mode to monitor the spectral characteristics of interference with high resolution, while the data acquisition process focuses on critical metrics such as power spectrum magnitude and Received Signal Reference Power (RSRP) in Figure 8 [9]. Through this comprehensive hardware integration, each measurement stage is able to quantitatively capture signal degradation phenomena, providing an accurate picture of the impact of interference on communication quality in various distance scenarios.

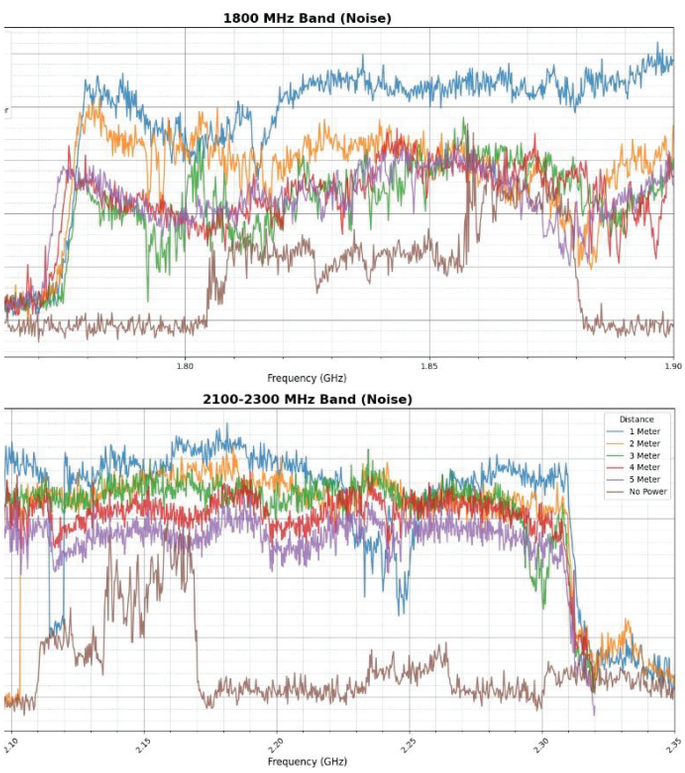


Fig. 8. Power Test Results Graph Transmit Frequency 1800–2300 MHz

D. The Effectiveness of Jammer Transmission in Interfering with Cellular Devices

Testing of the jammer module in the 1800–1995 MHz and 2100–2300 MHz frequency ranges showed high spectral effectiveness, successfully overlapping with the downlink transmission frequency of the Base Transceiver Station (BTS) [1]. This spectral coverage was verified through real time spectrum graph observation using LibreVNA, which confirmed high power spectral density (PSD) across the entire target frequency range . The experimental results, as summarized in Table II, show that the device achieves optimal performance in suppressing cellular signals by significantly reducing the Signal to Interference and Noise Ratio (SINR) [15]. This phenomenon indicates the deliberate domination of noise and

interference in the target spectrum, which effectively hides the Reference Signal (RS) and hinders the decoding process by the User Equipment (UE). Specifically, at a distance of 1 meter, the jammer forces the RSRP to reach a minimum threshold of -116 dB, causing an almost instantaneous network disconnection with a latency of only 4.3 seconds [5]. As the physical distance increases to 5 meters, a measurable recovery in RSRP values and an increase in connection disconnection latency are observed.

TABLE II
RSRP TEST RESULTS AND CONNECTION TIME BASED ON DISTANCE

	Test 1		Test 2		Test 3		Average	
Dist. (m)	**RSRP (dB)**	**Time (s)**	**RSRP (dB)**	**Time (s)**	**RSRP (dB)**	**Time (s)**	**RSRP (dB)**	**Time (s)**
1	-118	6	-114	4	-116	3	-116.0	4.3
2	-113	8	-111	7	-115	7	-113.0	7.3
3	-109	10	-110	14	-111	12	-110.0	12.0
4	-110	21	-108	22	-109	21	-109.0	21.3
5	-108	26	-108	27	-110	28	-108.7	27.0

V. CONCLUSION

This study demonstrates an increase in the efficiency of signal modulation system architecture through the implementation of Software Defined Radio (SDR) capable of operating dynamically on the cellular frequency spectrum. The use of HackRF One integrated with the GNU Radio framework allows the consolidation of various conventional hardware functions, such as signal generators and voltage controlled oscillators (VCOs), into a single, highly reconfigurable software module. The experimental results confirm the capability of SDR in abstracting the functionality of multiple hardware devices simultaneously, which provides operational flexibility to adjust technical parameters in real time without relying on the addition of physical instruments. Although the system offers superior spectral agility, overall performance is still limited by inherent device parameters, specifically a data bandwidth of 20 MHz and 8 bit sampling accuracy, as well as low output power levels. These findings indicate that to achieve optimal signal suppression effectiveness in cellular networks, the system requires the integration of an external power amplification stage to compensate for the limitations of the SDR's internal amplification, or a transition to a platform with higher amplification density and bandwidth specifications.

VI. AKNOWLEDGEMENTS

The authors express their gratitude to the GNU Radio team for providing the essential software framework for system design. Special appreciation is extended to the PUI-PT IS-IoT laboratory for granting access to the signal jamming system software. Furthermore, the authors thank their academic advisors and colleagues for their invaluable guidance and technical support during the experimental and evaluation phases.

REFERENCES

[1] S. Xia and Q. Wu, "Mitigating the Electromagnetic Coupling Between the Tx and Rx Channels in Software-Defined Radio Through a Digital-Domain Cancellation Algorithm," *IEEE Transactions on Electromagnetic Compatibility*, vol. 66, no. 5, pp. 1389–1398, Oct. 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10608440/

[2] Z. U. A. Tariq, E. Baccour, A. Erbad, and M. Hamdi, "Reinforcement Learning for Resilient Aerial-IRS Assisted Wireless Communications Networks in the Presence of Multiple Jammers," *IEEE Open Journal of the Communications Society*, vol. 5, pp. 15–37, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10325598/

[3] A. Thavai, M. Subhedar, Y. Thakur, and M. Patil, "Design an RF Up-Down Convertor using Software Defined Radio and GNU Radio," *Journal of Mobile Multimedia*, pp. 901–916, Oct. 2024. [Online]. Available: https://journals.riverpublishers.com/index.php/JMM/article/view/24999

[4] M. Gummineni and T. R. Polipalli, "Implementation of Reconfigurable Transceiver using GNU Radio and HackRF One," *Wireless Personal Communications*, vol. 112, no. 2, pp. 889–905, May 2020. [Online]. Available: http://link.springer.com/10.1007/s11277-020-07080-0

[5] Rustamaji, S. Kliwati, and W. Widada, "The angle of arrival estimation of frequency-hopping cooperative object based on software-defined radio," *Scientific Reports*, vol. 14, no. 1, p. 7732, Apr. 2024. [Online]. Available: https://www.nature.com/articles/s41598-024-58488-8

[6] M. Z. Khan, A. Taha, W. Taylor, M. A. Imran, and Q. H. Abbasi, "Non-Invasive Localization Using Software-Defined Radios," *IEEE Sensors Journal*, vol. 22, no. 9, pp. 9018–9026, May 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9738615/

[7] D. Kafetzis, S. Vassilaras, G. Vardoulias, and I. Koutsopoulos, "Software-Defined Networking Meets Software-Defined Radio in Mobile ad hoc Networks: State of the Art and Future Directions," *IEEE Access*, vol. 10, pp. 9989–10 014, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9684421/

[8] Z. Htay, Z. Ghassemlooy, M. M. Abadi, A. Burton, N. Mohan, and S. Zvanovec, "Performance Analysis and Software-Defined Implementation of Real-Time MIMO FSO With Adaptive Switching in GNU Radio Platform," *IEEE Access*, vol. 9, pp. 92 168–92 177, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9466115/

[9] J. Manco, I. Dayoub, A. Nafkha, M. Alibakhshikenari, and H. B. Thameur, "Spectrum Sensing Using Software Defined Radio for Cognitive Radio Networks: A Survey," *IEEE Access*, vol. 10, pp. 131 887–131 908, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9989359/

[10] R. Ferreira, J. Gaspar, P. Sebastião, and N. Souto, "A Software Defined Radio Based Anti-UAV Mobile System with Jamming and Spoofing Capabilities," *Sensors*, vol. 22, no. 4, p. 1487, Feb. 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/4/1487

[11] E. Faulkner, Z. Yun, S. Zhou, Z. J. Shi, S. Han, and G. B. Giannakis, "An Advanced GNU Radio Receiver of IEEE 802.15.4 OQPSK Physical Layer," *IEEE Internet of Things Journal*, vol. 8, no. 11, pp. 9206–9218, Jun. 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9348953/

[12] A. Daud, M. B. Khan, A. B. Khattak, S. A. K. Tanoli, A. Mustafa, M. Rehman, and O. L. A. López, "Next-Generation Security: Detecting Suspicious Liquids Through Software Defined Radio Frequency Sensing and Machine Learning," *IEEE Sensors Journal*, vol. 24, no. 5, pp. 7140–7152, Mar. 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10400410/

[13] M. Marey and H. Mostafa, "Turbo Modulation Identification Algorithm for OFDM Software-Defined Radios," *IEEE Communications Letters*, vol. 25, no. 5, pp. 1707–1711, May 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9335963/

[14] D. M. Molla, H. Badis, L. George, and M. Berbineau, "Software Defined Radio Platforms for Wireless Technologies," *IEEE Access*, vol. 10, pp. 26 203–26 229, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9721283/

[15] Z. Dai, Y. He, V. Tran, N. Trigoni, and A. Markham, "DeepAoANet: Learning Angle of Arrival From Software Defined Radios With Deep Neural Networks," *IEEE Access*, vol. 10, pp. 3164–3176, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9667511/

A Portable Intelligent System for Lung Disease Prediction Using Machine Learning Models

Abstract— **This research presents the design and implementation of an intelligent biomedical device for early prediction of lung-related diseases such as asthma and influenza, as well as diabetes. The system integrates multiple physiological sensors, including the MAX30102 for oxygen saturation and heart rate, the GY-906 for body temperature, and a glucometer for blood glucose measurement, all interfaced with an Arduino microcontroller. The collected physiological and symptom-based data are processed using machine learning algorithms, specifically K-Nearest Neighbors (KNN) and Random Forest classifiers, achieving an overall prediction accuracy of approximately 95%. The device supports real-time data transmission via Bluetooth to a mobile application and a secure web interface for remote monitoring. The proposed system is portable, low-cost, and user-friendly, offering a practical solution for preliminary health assessment and remote patient monitoring in home and community healthcare settings across the United States. The integration of IoT and machine learning demonstrates significant potential to enhance early disease detection, reduce healthcare costs, and improve patient outcomes in modern healthcare systems.**

Keywords—Intelligent biomedical device, lung disease prediction, machine learning, IoT-based healthcare, Arduino, MAX30102 sensor, GY-906 infrared sensor, glucometer integration, K-Nearest Neighbors (KNN), Random Forest classifier.

I. INTRODUCTION

The burden of chronic and acute diseases such as asthma, influenza, and diabetes continues to strain healthcare systems across the United States, driving the need for early, cost-effective diagnostic solutions. Traditional medical diagnostics frequently demand clinical visits, specialized testing facilities, and time delays that may hinder timely intervention. In this context, wearable and IoT-enabled biomedical devices present a compelling opportunity: continuous, real-time physiological monitoring combined with data-driven analytics to assist in early disease detection and risk screening. The promise of wearable biosensors is that they can transform health care from reactive to proactive, bridging gaps in access and helping to reduce disease progression before severe symptoms emerge. Over the past decade, rapid advances in sensor technologies, wireless communications, and miniaturized electronics have enabled numerous wearable systems for continuous health monitoring. Reviews of wearable sensor systems report that devices can now reliably capture vital signs such as heart rate, SpO_2, body temperature, and respiration rate under controlled conditions, enabling new applications in remote monitoring, early warning, and health management [1], [2]. Wearable biosensors, in particular, hold promise due to their specificity, low power consumption, portability, and ability to operate in nonclinical settings [3] [4]. Researchers have explored combining multiple sensing modalities to enhance diagnostic inference, merging physical signals with biochemical data streams to yield richer health insights [5], [6] [7]. However, deploying such systems in realistic environments remains challenging, particularly concerning sensor accuracy, data robustness, energy efficiency, and integration into easily usable platforms. Among the vital parameters monitored, oxygen saturation (SpO_2) and heart rate are central to assessing respiratory and cardiovascular function and are often used as key features in predictive models for lung health. Similarly, body temperature provides important clues about systemic inflammation or infection. In parallel, continuous or frequent glucose monitoring is essential for managing and predicting diabetic states. Prior work has shown that combining signals from multiple sensors can improve classification performance for disease prediction over using any single signal alone [8] [9]. For instance, in multi sensor systems for diabetes prediction, integrating ECG, respiratory, accelerometer, and glucose readings has yielded better accuracy than any single-modality model [9]-[12]. The challenge lies not only in accurate sensing, but also in fusing heterogeneous data streams and applying machine learning models that generalize well beyond controlled experiments.

Machine learning has emerged as a powerful tool for transforming physiological and symptomatic data into disease predictions. Algorithms such as K-Nearest Neighbors (KNN), Random Forest, support vector machines, and ensemble methods have been successfully applied in many biomedical domains. In diabetes prediction, comparative studies show that Random Forest and ensemble techniques often outperform simpler models, particularly when combined with careful feature engineering and class balancing strategies. Some systems even embed end-to-end IoT-Edge designs that incorporate privacy, security, and automated model deployment features. Meanwhile, in respiratory disease monitoring, combining multiple vital sign signals and leveraging feature fusion approaches has demonstrated promising results in detecting patterns indicative of compromised pulmonary function or exacerbations.

Despite these advances, several key obstacles remain before practical, everyday biomedical monitoring devices are widely adopted. First, sensor accuracy and calibration in nonideal conditions (e.g., motion artifacts, ambient interference, varying skin types) can degrade signal quality. Many published systems perform excellently in laboratory or controlled settings but struggle in real-world use. Second, limited sample diversity in many studies constrains the generalizability of machine learning models trained on narrow demographic sets may fail when deployed in broader populations. Third, wearable devices must balance high sampling rates with low power consumption and small form factors, meaning that the hardware design must carefully trade off performance, energy, and usability. Fourth, dealing with missing data, sensor drift, and model robustness over time are crucial concerns in any long-term deployment.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

Finally, real-time connectivity, latency, data privacy, and security must be addressed carefully, particularly when transmitting personal health information over wireless links or to cloud servers.

In light of these challenges and opportunities, this work proposes a unified, portable biomedical device that integrates multiple sensors (e.g., SpO_2, heart rate, body temperature, and glucose) and symptom-based user inputs, combined with machine learning classification (specifically KNN and Random Forest), to generate predictions for respiratory diseases (such as asthma and influenza) and diabetes with high accuracy. The device is engineered to collect raw sensor data, preprocess and fuse feature inputs, and run predictive inference locally or semi-locally, thereby reducing latency. Prediction outcomes are communicated via Bluetooth to a mobile application and simultaneously displayed on a web dashboard, enabling users or clinicians to view insights in real time. The system is designed to be user-friendly, low cost, and sufficiently compact for practical use in home or community settings across the United States.

The contributions of the proposed system are fourfold: (1) design and implementation of a multi-sensor acquisition platform with efficient interfacing and calibration; (2) development of a fused feature pipeline combining physiological, symptom, and contextual inputs; (3) application and validation of KNN and Random Forest classifiers for disease prediction with experimental evaluation demonstrating high accuracy; and (4) integration of wireless communication, mobile app, and web dashboard to enable scalable deployment, remote monitoring, and ease of use. Through careful attention to real-world usability, signal processing, and machine learning generalization, this work aims to push closer toward robust, widely deployable biomedical consumer devices capable of early disease screening and ongoing health monitoring in U.S. domestic settings.

II. Methodology

The methodology for this research focuses on the design, development, and evaluation of a smart biomedical device capable of preliminary disease prediction using physiological parameters. The development process is divided into four main stages: system design, hardware integration, software development, and performance evaluation.

A. System Design Approach

Multiple design approaches were analyzed to fulfill the desired objectives, requirements, and constraints. Four potential designs were considered:

1. Arduino with Bluetooth and a separate website system
2. MAX30100 and LM35-based low-cost sensor system
3. Arduino with Wi-Fi module and Google Forms for data collection
4. Raspberry Pi with integrated website and machine learning system.

Each design was evaluated against three criteria: objectives fulfillment, requirement satisfaction, and adherence to constraints. Detailed comparisons considered cost, efficiency, usability, maintainability, accessibility, sustainability, and criticality. Based on these analyses, the first design, a decentralized Arduino-based system with Bluetooth and a separate website, was identified as the optimal solution due to its balanced performance, reliability, and maintainability.

B. Hardware Implementation

The hardware subsystem was designed to collect real-time physiological data from users. Key components include:

- ✓ MAX30102 sensor: Measures oxygen saturation (SpO_2) and heart rate.
- ✓ GY-906 infrared temperature sensor: Measures body temperature without physical contact.
- ✓ Glucometer interface: Captures blood glucose levels.
- ✓ Arduino Uno microcontroller: Integrates sensors and processes raw signals.
- ✓ Bluetooth module: Enables wireless transmission of sensor data to a mobile application.

Hardware assembly involved soldering sensors to a breadboard, securing components in a PVC casing with hot glue, and validating connections with a multimeter. Proteus Professional was used for circuit simulation prior to physical assembly to ensure correct sensor operation.

C. Software Development

Software development was performed to enable data acquisition, processing, and disease prediction:

1. **Machine Learning Model**:
   - ➢ Two models, K-Nearest Neighbors (KNN) and Random Forest, were trained using sensor and symptom data to classify preliminary disease conditions.
   - ➢ Model training and evaluation were performed on Google Collaboratory, achieving approximately 95% accuracy.
2. **Website Development**:
   - ➢ A Python Django framework was implemented to create a web interface for user input and result visualization.
   - ➢ The website facilitates real-time data submission, automatic machine learning prediction, and immediate result publication.
3. **Mobile Application Integration**:
   - ➢ An Android application was developed using Android Studio, enabling users to monitor sensor outputs and view results on mobile devices.
   - ➢ Bluetooth communication ensures real-time monitoring and enhances system accessibility.

D. System Integration

The hardware and software components were fully integrated into a decentralized system architecture, ensuring modularity and scalability. Sensor data is collected locally by Arduino, transmitted via Bluetooth to the mobile device, and simultaneously uploaded to the web server for machine learning processing. This approach reduces task dependencies and simplifies maintenance while maintaining high efficiency and reliability.

E. Performance Evaluation

The integrated system was evaluated based on the following metrics:

- Cost-effectiveness: Compared with alternative designs, the chosen system achieves optimal cost-performance balance.
- Usability: The decentralized approach ensures user-friendly operation with minimal manual intervention.
- Success rate and efficiency: Automatic result generation and web-based visualization reduce processing delays.
- Maintainability: Modular design allows easy component replacement and updates without system downtime.
- Environmental sustainability: Low-power sensors and microcontroller usage ensure minimal energy consumption.

The performance evaluation demonstrated that the selected design fulfills all objectives and requirements while overcoming key constraints, including sensor reliability and data accessibility.

III. Research Analysis

The research analysis focuses on evaluating the performance and effectiveness of the proposed biomedical device in predicting preliminary disease conditions. This section systematically examines the hardware, software, and integrated system performance through a series of quantitative and qualitative assessments. Key parameters such as sensor accuracy, data transmission reliability, machine learning model performance, and user interface responsiveness are analyzed. Additionally, the selected design is compared against alternative design approaches based on critical criteria including cost, efficiency, usability, maintainability, accessibility, and sustainability.

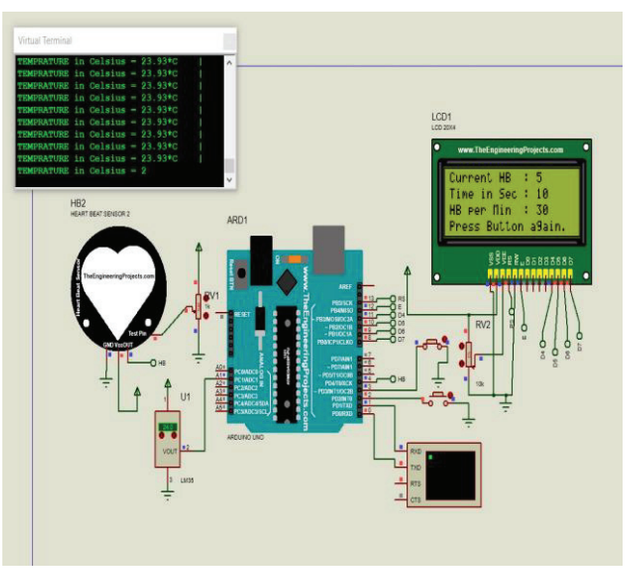


Fig.1. Simulation of the research

Figure 1. Simulation setup in Proteus software. The system was first modeled and tested in the Proteus environment to verify sensor functionality, circuit connections, and signal accuracy before proceeding to hardware implementation.

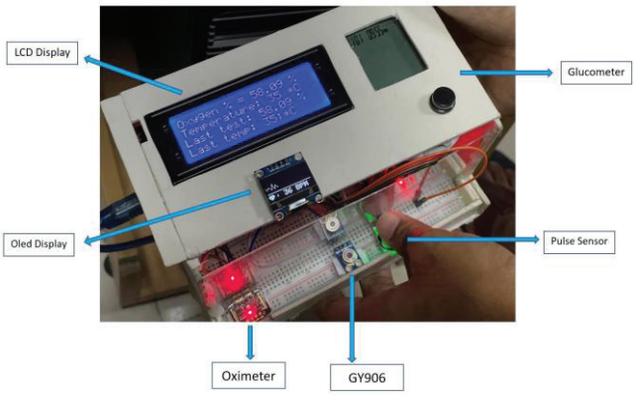


Fig .2. Hardware setup of the research

Figure 2. Hardware prototype setup. The assembled system integrates MAX30102, GY-906, and a glucometer with the Arduino microcontroller. Each sensor was calibrated to ensure accurate measurement of physiological parameters, and a Bluetooth module was included to enable real-time data monitoring via a mobile application. The components are securely mounted on a PVC board to ensure structural stability and user-friendly operation.

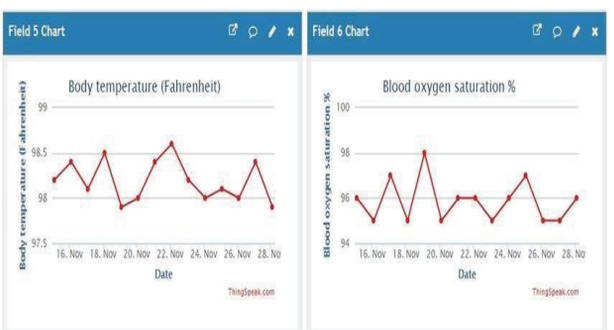


Fig. 3. Temperature and oxygen level data visualization test on ThingSpeak

Figure 3. Temperature and oxygen level data visualization using ThingSpeak. The GY-906 temperature sensor and heart pulse sensor were tested for accurate data acquisition and real-time monitoring. Sensor measurements were transmitted via a Wi-Fi module to the ThingSpeak platform, ensuring reliable wireless data transfer prior to storage for machine learning analysis.

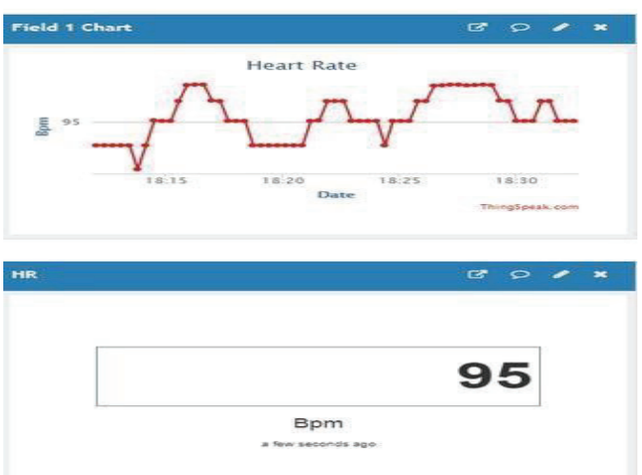


Fig. 3. Heart rate data visualization test on ThingSpeak

Accurate data collection is essential for training and testing the machine learning models. Sensor-based measurements were obtained from over 250 patients using the prototype, supplemented with oral questionnaire responses. Additionally, the hospital provided over 500 historical patient records. The data were consolidated, cleaned for missing entries, and stored in Google Sheets, then uploaded to Kaggle in CSV format for future research and development.

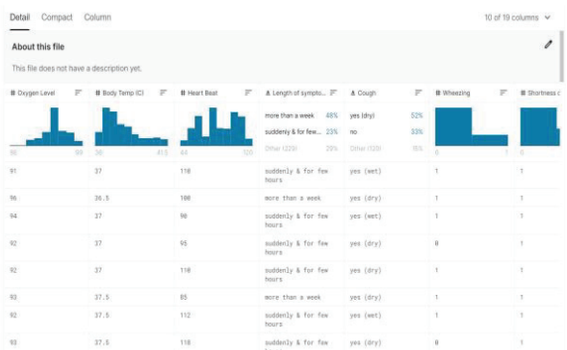


Fig. 4. Dataset sample for lung disease



Fig.5. Dataset sample for diabetes

The collected data were preprocessed by converting categorical string inputs into integer representations using one-hot encoding, enabling compatibility with machine learning algorithms. Two supervised models *K*-Nearest Neighbors (KNN) and Random Forest were implemented for disease prediction. KNN classifies new inputs based on the labels of nearest neighbors, with the optimal *K* determined through multiple iterations. Random Forest improves accuracy using an ensemble of decision trees. Model performance was evaluated using the train-test split method, ensuring reliable predictions on unseen data.

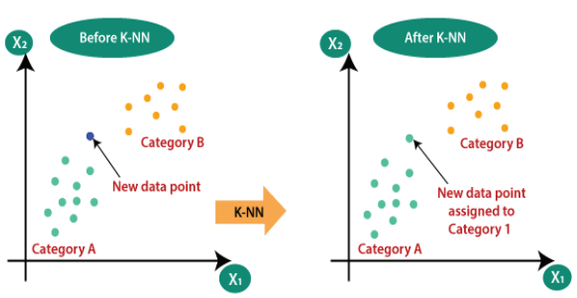


Fig.6. Before and after of K-NN machine learning

To evaluate the predictive performance of the machine learning models, the dataset was divided into training and testing subsets using the train-test split method. The training dataset, typically 75% of the total data, was used to fit the models, while the remaining 25% formed the test dataset for performance evaluation. Predictions generated on the test set were compared with actual outcomes to assess model accuracy and generalization capability on unseen data.

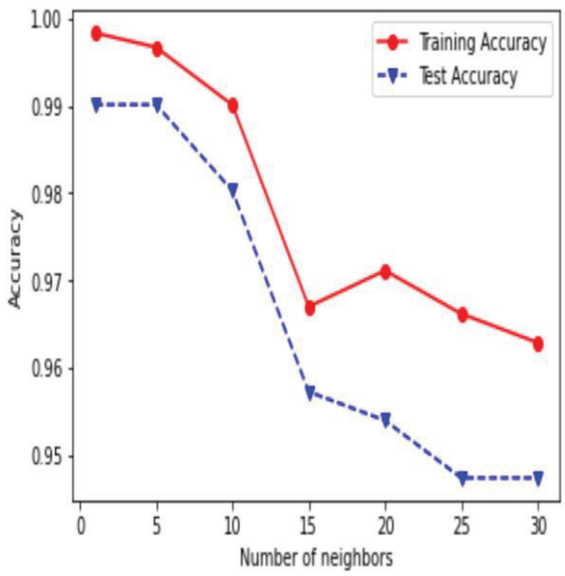


Fig.7. Training and test dataset comparison

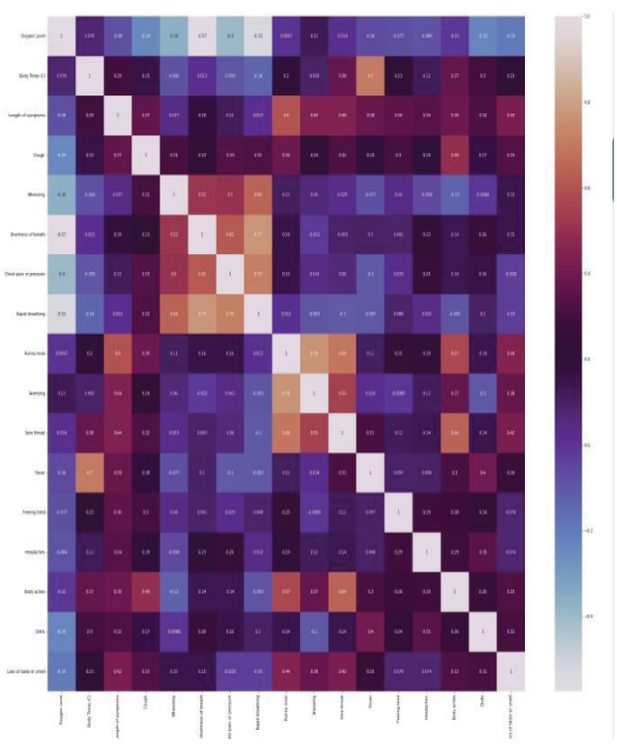


Fig.7. Dataset correlation between all the features (heatmap)

The proposed system integrates hardware, data collection, machine learning, and web deployment for disease prediction. Users input data through a web form with integer fields and selectable options, which the backend processes using the K-Nearest Neighbor model to generate disease predictions along with model accuracy. New users register with basic credentials before accessing the form. Sensor data collected by the device, supplemented by hospital records, provide a reliable dataset for training, enabling effective machine learning. Both KNN and Random Forest models achieved approximately 95% accuracy, and the web interface allows users to access predictions efficiently, demonstrating the overall system's functionality and effectiveness.

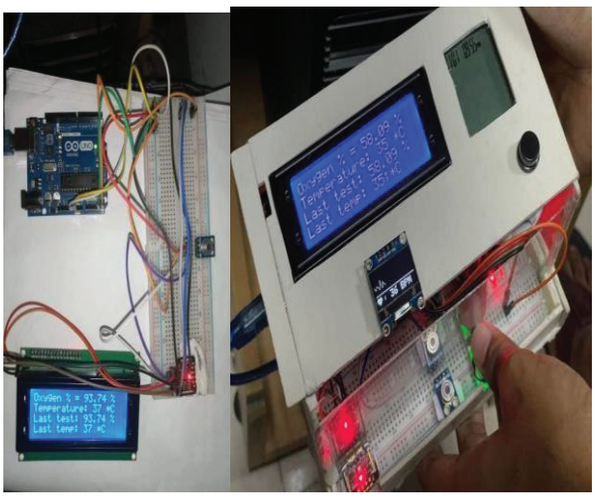


Fig.8. Complete Prototype of the device

Figure 8 is the complete prototype of the biomedical device. Initially, the MAX30100 pulse oximeter sensor was tested but exhibited voltage drop issues, requiring additional resistors and increasing circuit complexity. To ensure stable and accurate readings, the sensor was upgraded to the MAX30102, a second-generation oximeter compatible with the Arduino Uno. Similarly, the LM35 temperature sensor was replaced with the more precise GY-906 infrared sensor to achieve stable temperature measurements. Due to power and heat distribution challenges when integrating multiple sensors on a single microcontroller, an additional Arduino Nano was incorporated to support the heart rate sensor and enhance overall system stability. The final prototype demonstrates a compact, efficient, and reliable configuration suitable for real-time biomedical monitoring.



Fig.9. All sensors Outputs are displayed

IV. Limitations And Novelty

The proposed system demonstrates effective integration of multi-sensor data acquisition and machine learning for disease prediction; however, it has certain limitations. The dataset size, though adequate for validation, remains limited for large-scale clinical application. Environmental conditions such as temperature variation and motion artifacts can affect sensor accuracy, and the use of multiple microcontrollers to prevent overheating increases power consumption and system complexity. Additionally, the reliance on Wi-Fi connectivity restricts data transmission in low-network areas. Despite these constraints, the novelty of this research lies in its unified design that combines oxygen saturation, heart rate, body temperature, and glucose level measurements within a single low-cost, portable device. The system employs K-Nearest Neighbor and Random Forest algorithms for real-time disease prediction, integrated through a web-based platform for accessible and remote health monitoring. This hybrid IoT–machine learning approach provides an innovative, scalable, and cost-effective solution for early disease detection and personalized healthcare support.

V. Future Work

Future work will focus on expanding the dataset with more diverse patient records to enhance model generalization and diagnostic accuracy. Additional sensors, such as respiratory rate and blood pressure modules, will be integrated to broaden the system's diagnostic capabilities. Optimization of power consumption and circuit design will be pursued to improve portability and reduce hardware complexity. Incorporating cloud-based storage and real-time analytics can further enable large-scale data management and remote clinical monitoring. Moreover, advanced deep learning models such as Convolutional Neural Networks (CNN) or Long Short-Term Memory (LSTM) networks may be implemented to achieve higher prediction accuracy and automated feature extraction. The long-term objective is to develop a fully self-contained wearable biomedical monitoring system capable of real-time disease prediction and continuous patient health assessment.

VI. Conclusion

This research presents the design and implementation of an intelligent biomedical device capable of predicting lung-related diseases and diabetes using integrated sensors and machine learning algorithms. The system combines MAX30102, GY-906, and a glucometer with Arduino-based processing to collect physiological data, which are analyzed using K-Nearest Neighbor and Random Forest models. With an achieved accuracy of approximately 95%, the system demonstrates reliable disease prediction and real-time monitoring through a web-based interface. The hardware is low-cost, portable, and suitable for use in preliminary health screening. By integrating IoT-based sensing with machine learning, this research provides an efficient and accessible healthcare solution that supports early diagnosis and remote monitoring. The proposed system establishes a foundation for future development toward a fully automated, real-time biomedical monitoring platform.

References

[1] Patel, S., Park, H., Bonato, P., Chan, L., & Rodgers, M. (2012). A review of wearable sensors and systems with application in rehabilitation. Journal of neuroengineering and rehabilitation, 9(1), 21.

[2] Majumder, S., Mondal, T., & Deen, M. J. (2017). Wearable sensors for remote health monitoring. Sensors, 17(1), 130.

[3] Paul, D., Prince, A. T. Z., Nabil, M. A., Jim, M. I., Ratul, S. Z. A., & Tasnim, A. (2025, July). IoT-Enhanced Multimodal Wearable System for Women's Safety. In 2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN) (pp. 1-6). IEEE.

[4] Ajami, S., & Teimouri, F. (2015). Features and application of wearable biosensors in medical care. Journal of Research in Medical Sciences, 20(12), 1208-1215.

[5] Paul, D., Aliuzzaman, S. M., Khan, M. F., Shakil, M. T., Ali, M. M., & Rabbi, A. (2025). An Innovative Embedded Ventilator for Accessible

and Intelligent Respiratory Support. Journal of Medical and Health Studies, 6(1), 99-108.

[6] John, A., Cardiff, B., & John, D. (2024). A Review on Multisensor Data Fusion for Wearable Health Monitoring. arXiv preprint arXiv:2412.05895.

[7] Li, C., Un, K. F., Mak, P. I., Chen, Y., Muñoz-Ferreras, J. M., Yang, Z., & Gómez-García, R. (2018). Overview of recent development on wireless sensing circuits and systems for healthcare and biomedical applications. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 8(2), 165-177.

[8] Hasan, M., & Yasmin, F. (2025). Predicting Diabetes Using Machine Learning: A Comparative Study of Classifiers. arXiv preprint arXiv:2505.07036.

[9] Nurmi, J., & Lohan, E. S. (2023). Machine-learning-based diabetes prediction using multisensor data. IEEE Sensors Journal, 23(22), 28370-28377.

[10] Alehegn, M., Joshi, R., & Alehegn, M. (2017). Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach. International Research Journal of Engineering and Technology, 4(10), 426-436.

[11] Sabry, A. H., Bashi, O. I. D., Ali, N. N., & Al Kubaisi, Y. M. (2024). Lung disease recognition methods using audio-based analysis with machine learning. Heliyon, 10(4).

[12] Kumar, S., Kumar, H., Kumar, G., Singh, S. P., Bijalwan, A., & Diwakar, M. (2024). A methodical exploration of imaging modalities from dataset to detection through machine learning paradigms in prominent lung disease diagnosis: a review. BMC Medical Imaging, 24(1), 30.

Performance Evaluation of Hybrid Machine Learning Models for Heart Disease Prediction in U.S. Clinical Decision Support Systems

Abstract— **Heart disease remains one of the leading causes of morbidity and mortality in the United States, underscoring the need for accurate and early prediction methods to support clinical decision-making. This study evaluates the performance of several hybrid machine learning models for heart disease prediction using clinically relevant features. The models examined include Decision Tree (DT), Random Forest (RF), Gradient Boosting Machine (GBM), Support Vector Machine (SVM), and k-Nearest Neighbor (KNN). Univariate feature selection was applied to identify the most significant predictors. Model performance was assessed using accuracy (ACC), precision (PRE), and F1-score. Experimental results show that the Random Forest classifier outperformed all other models, achieving an accuracy of 98.31% on the test dataset. Furthermore, restricting the model to the top eight most informative features improved interpretability without compromising predictive capability. The findings indicate that Random Forest, both independently and when combined with feature selection, offers a highly reliable and robust solution for heart disease prediction within U.S. healthcare settings, supporting improved clinical decision support and patient outcomes.**

Keywords— Heart Disease Prediction, Cardiology, Machine Learning (ML), Hybrid Classifiers, Medical Data Analytics, Interpretability in ML, Univariate Feature Selection, Random Forest (RF), Support Vector Machine (SVM), Gradient Boosting Method (GBM), k-Nearest Neighbor (KNN).

I. INTRODUCTION

Cardiovascular diseases (CVDs), particularly heart disease, remain a principal cause of death globally, accounting for approximately 31% of all deaths as per projections by 2030 [1]. Early detection and precise prediction of heart disease can significantly enhance clinical outcomes and lower mortality rates. However, traditional diagnostic methods often suffer from subjectivity, use of limited features, and delayed detection, warranting improved computational tools to assist clinicians.

Machine learning (ML) has demonstrated considerable promise in predictive diagnostics in U.S. owing to its capability to uncover latent patterns in complex datasets. The Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), k-Nearest Neighbor (KNN), and ensemble approaches have been widely applied for heart disease prediction with varying degrees of success [2] [3]. For instance, in a comparative evaluation using the UCI heart disease dataset, RF consistently achieved superior accuracy (around 89.7%) versus SVM and Logistic Regression [4].

Recent studies further reinforce RF's robustness. An ensemble RF model integrating multiple feature-selection techniques achieved nearly 99% accuracy [5]. Similarly, hybrid approaches combining RF with oversampling and ensemble mechanisms attained accuracies up to 99.83% [6]. Other works have explored hybrid methods, such as RF combined with a Multi-Layer Perceptron (MLP) meta-learner, yielding balanced performance with an F1‑score of 0.86 [7]. The importance of feature selection and model interpretability has also grown. A thorough evaluation of 15 ML models including RF, XGBoost, Bagged Trees demonstrated that while XGBoost and Bagged Trees achieved ~93% accuracy, RF maintained robustness and interpretability under cross-validation [8]. Moreover, novel deep-learning architectures, such as a hierarchical attention network (HAN) applied to ECG data, offered comparable accuracy (~98.5%) and enhanced interpretability, underscoring the need for transparent models [9]. Further innovation arrives in the form of optimized modern models. A Transformer architecture tuned via Particle Swarm Optimization (PSO) attained 96.5% accuracy, outperforming RF's 92.2% on the same dataset [10] [11] [12]. Nonetheless, in tabular healthcare settings, RF often remains the top performer due to its balance of accuracy and feature importance mapping [13]. In this context, this study aims to develop an efficient, interpretable ML framework for heart disease prediction. We evaluate five hybrid classifiers DT, RF, GBM, SVM, and KNN using univariate feature selection to identify the most predictive features. Performance is gauged via Accuracy, Precision, and F1 Score. Our preliminary findings suggest RF achieves leading accuracy (~98.31%) within this setup, offering a compelling trade-off between performance and interpretability. Additionally, analyzing the eight selected features enhances model transparency, supporting its utility in clinical settings.

II. EARLIER RESEARCH

Research on heart disease prediction continues to evolve, particularly in enhancing model accuracy through hybrid methods, optimized feature selection, and explainability.

A) Feature Selection Strategies:

Ahmad et al. (2025) conducted a comprehensive examination of feature selection techniques ANOVA, Mutual Information (MI), and Chi-Square applied to multiple ML and deep-learning frameworks. Results revealed that MI yielded the highest overall accuracy (82.3%) and recall (0.94), outperforming simpler methods. Notably, RF and logistic regression showed improved performance with MI, while KNN and SVM lagged (51–54%), regardless of feature selection method [14]. Similarly, Pathan et al. (2022) demonstrated that reducing data dimensionality via feature selection can significantly improve classification accuracy and training efficiency, even when using fewer features [15].

B) *Ensemble & Hybrid Classifier Approaches:*

Sen et al. (JMAI) explored a soft-voting ensemble integrating Gaussian Naive Bayes, Cat Boost, Light GBM, XGBoost, RF, and MLP. They combined UCI and Stalog datasets and found that the ensemble significantly reduced false negatives critical in clinical diagnostics compared to

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

individual models [16]. Mahmud et al. assessed stacking ensembles using SVM, DT, and RF as base classifiers, with XGBoost as a meta-classifier. Their hybrid model achieved superior performance over single-model baselines, showcasing the potential of stacking [17]. Explainable-AI-enabled hybrid frameworks have also gained traction. One study leveraged ten classifiers including RF, MLP, KNN, XGB, and GBM combined with Pearson correlation, Chi-Square, and RFE for feature selection. The ensemble-enhanced model demonstrated robustness and interpretability in heart disease prediction [18].

C) Advanced Optimization in Feature Selection:

A hybrid feature-selection mechanism combining Random Forest, Adaboost, and linear correlation was proposed by Qinglin et al. This method selected 11 high-impact features and, when applied to various ML algorithms, improved classification accuracy over methods without feature selection.

D) Interpretability and Explainable AI:

Interpretability remains a growing concern. A comprehensive review on interpretable ML techniques for ECG-based heart disease classification highlighted global and post-hoc interpretability methods such as surrogate models and attention mechanisms that promote transparency while preserving predictive performance [19].

Overall, the literature underscores that mutual Information is among the most effective univariate feature selectors, particularly for boosting RF and logistic regression performance. Soft-voting and stacking ensembles, combining diverse classifiers, consistently outperform single-model baselines. Hybrid selection strategies blending algorithmic selection (e.g., RF, Adaboost) with statistical correlation yield improved feature identification and prediction. Explainability frameworks, both intrinsic and post-hoc, are indispensable for clinical adoption of ML models.

These insights directly inform your research direction employing univariate selection for simplicity and interpretability, evaluating multiple hybrid classifiers, and emphasizing model transparency aligns strongly with both current trends and practical clinical relevance.

III. METHODOLOGY

The proposed approach for heart disease prediction consists of four main stages: data preprocessing, feature selection, model training, and performance evaluation. A heart disease dataset was preprocessed by handling missing values, normalizing attributes, and encoding categorical features. To reduce dimensionality and enhance efficiency, univariate feature selection was applied, resulting in eight key predictors. Five classifiers Decision Tree (DT), Random Forest (RF), Gradient Boosting Method (GBM), Support Vector Machine (SVM), and k-Nearest Neighbor (KNN)—were then trained using optimized hyperparameters. Finally, model performance was assessed using Accuracy, Precision, and F1 Score, enabling comparative analysis and interpretability.

The dataset contains predicted values ranging from 0 to 4, where values 1 through 4 correspond to different stages of heart disease, and 0 indicates the absence of the condition. For the purpose of this study, the values from 1 to 4 were consolidated into a single class (1), effectively converting the problem into a binary classification task: 0 for no heart disease and 1 for presence of heart disease. The dataset originally includes 14 attributes, from which the top 8 features were selected for the predictive modeling process.

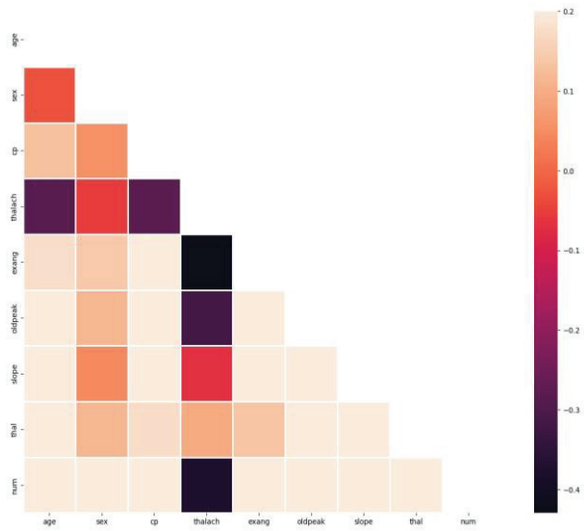


Fig.1. Feature Correlation Analysis Using Univariate Selection

Figure 2 illustrates the data visualization process, which plays a vital role in machine learning–based heart disease prediction. Visualization not only facilitates the interpretation of results but also enables the identification of hidden patterns and trends within the dataset. In this study, graphical representations were employed to gain a deeper understanding of the dataset and to support more effective analysis.

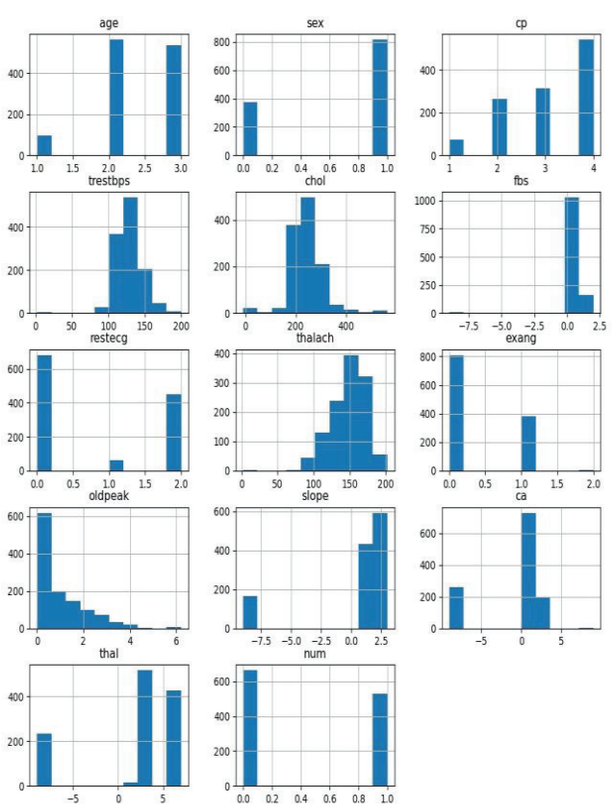


Fig.2. Graphical Visualization of Heart Disease Dataset for Pattern Identification and Analysis

Figures 3(a) and 3(b) depict a pair plot, also known as a scatter plot matrix, which visualizes relationships between multiple variables within the dataset. Each off-diagonal scatter plot illustrates the correlation between a pair of

variables, while the diagonal elements display the distribution of individual variables through histograms or kernel density estimates. Pair plots are particularly useful in exploratory data analysis, as they allow for the rapid identification of potential correlations, trends, and outliers among features.

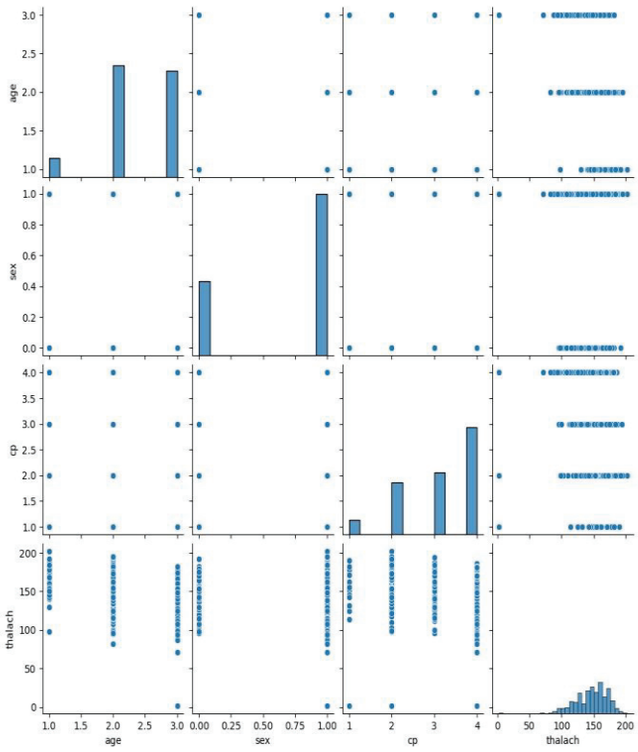


Fig.3(a). Data Visualization of Heart Disease Dataset for Pattern Recognition

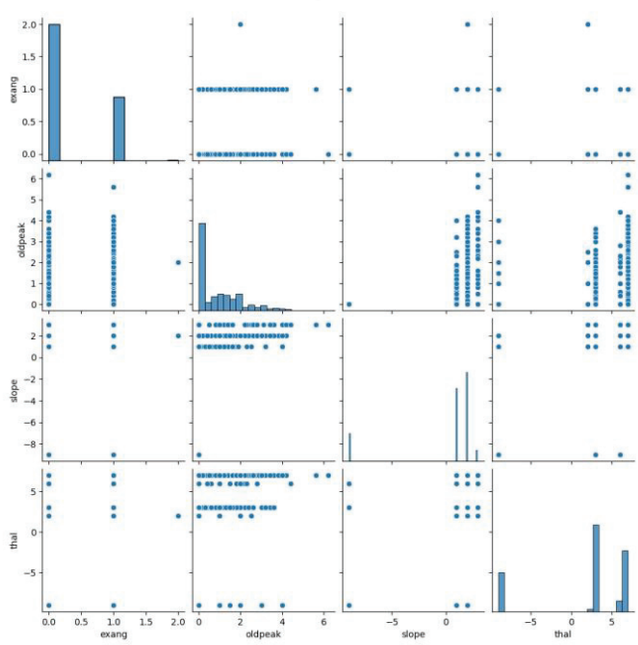


Fig.3(b). Pair Plot Analysis Showing Relationships and Distributions Among Dataset Features

The proposed framework for heart disease prediction involves data preprocessing, feature selection, classifier training, and performance evaluation. The dataset was preprocessed by handling missing values, normalizing features using Standard Scaler and Min-Max Scaler, and converting the target variable into a binary format (0: no heart disease, 1: presence of heart disease). From the original 14 attributes, the top 8 features were selected for modeling. Five machine learning classifiers were employed: Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), Gradient Boosting (GBM), and k-Nearest Neighbors (KNN). Each algorithm was optimized and trained using a 70:30 train-test split, with model performance evaluated via Accuracy, Precision, and F1 Score.

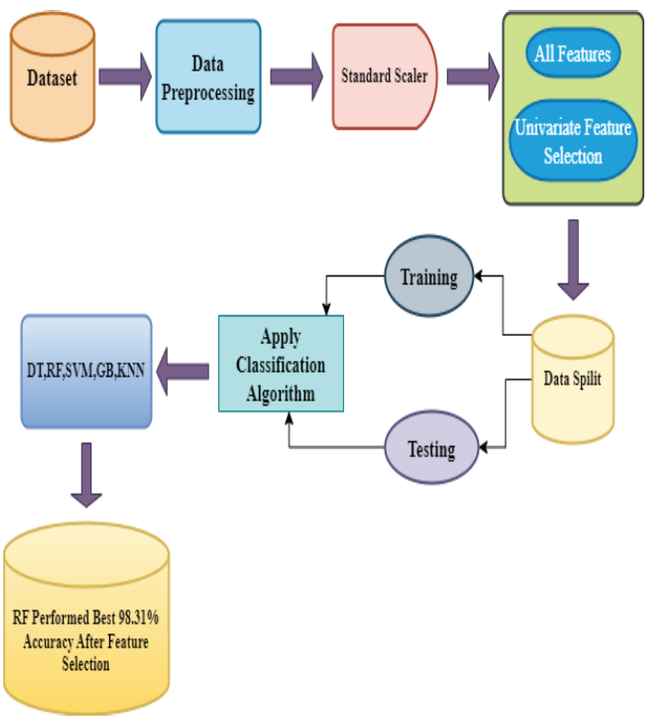


Fig.4. Block Diagram of the Heart Disease Prediction System

IV. Result Analysis

The dataset for heart disease prediction was obtained from the UCI Machine Learning Repository and preprocessed to ensure suitability for model training. Using univariate feature selection, the top eight features were identified from the original 14 attributes to develop the final predictive models. Five classifiers—Random Forest (RF), Decision Tree (DT), Gradient Boosting (GB), Support Vector Machine (SVM), and k-Nearest Neighbor (KNN) were trained in Jupyter Notebook to evaluate performance. Model evaluation focused primarily on accuracy as a key metric. Among all methods, Random Forest achieved the highest accuracy of 98.31%, followed by Decision Tree at 92.43%, Gradient Boosting at 91.03%, and SVM at 86.27%. The KNN classifier exhibited the lowest performance, with 81.79% accuracy. When only the five most significant features were considered, the accuracies slightly decreased for DT, GB, and KNN to 92.43%, 91.03%, and 81.79%, respectively, while Random Forest maintained the top performance, confirming its robustness and suitability for heart disease prediction.

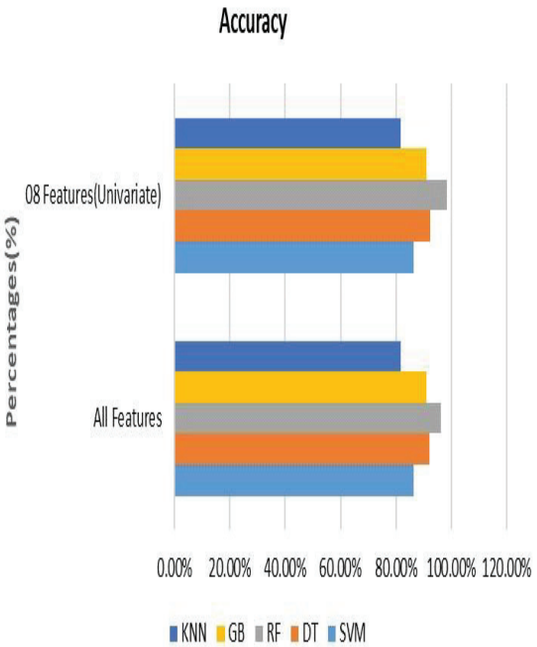


Fig.5. Performance Accuracy with All Features versus Top Features

The performance of the classifiers and hybrid algorithms was further evaluated using precision as an additional metric. When considering all 14 features, Random Forest achieved the highest precision at 98%, while K-Nearest Neighbor (KNN) showed the lowest precision of 86%. The Decision Tree (DT) and Support Vector Machine (SVM) attained precision scores of 95% and 90%, respectively, and Gradient Boosting (GB) reached 93%. When only the top eight features selected through univariate feature selection were used, Random Forest maintained the best precision at 99%, whereas KNN remained the lowest at 86%. Gradient Boosting and SVM achieved precision values of 93% and 90%, respectively, with DT maintaining 95% precision. Figure 6 illustrates these results for a clear comparison of classifier precision.

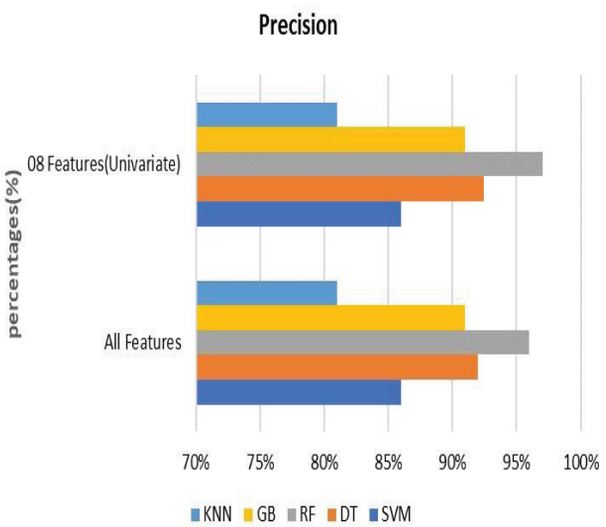


Fig.6. Comparative Precision Analysis of Machine Learning Classifiers

Recall, defined as the ratio of true positives to the sum of true positives and false negatives, was used to evaluate the classifiers. Considering all 14 features, Random Forest (RF) achieved the highest recall of 96%, whereas K-Nearest Neighbor (KNN) recorded the lowest at 81%. The Decision Tree (DT), Gradient Boosting (GB), and Support Vector Machine (SVM) obtained recall scores of 92%, 91%, and 86%, respectively. When using only the top eight features selected through univariate selection, Random Forest maintained superior recall at 98%, while KNN remained at 81%. DT, GB, and SVM achieved recall scores of 92%, 91%, and 86%, respectively. Figure 7 illustrates these results, providing a clear comparison of recall performance across the classifiers.

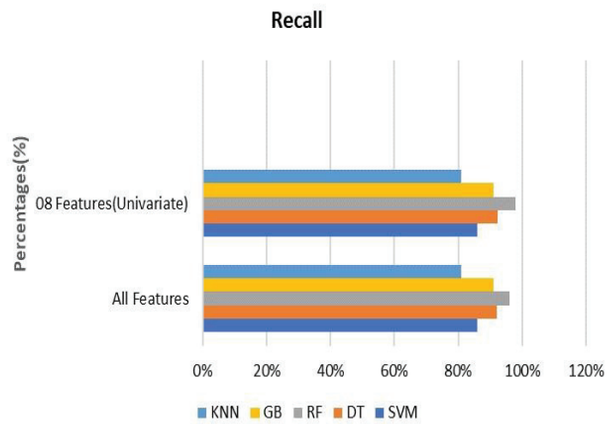


Fig.7. Comparative Recall Analysis of Machine Learning Classifiers

The F1 Score, which represents the harmonic mean of precision and recall, was employed to evaluate the balance between these two metrics. Using all 14 features, Random Forest (RF) achieved the highest F1 score of 96%, whereas K-Nearest Neighbor (KNN) recorded the lowest at 81%. The Decision Tree (DT) and Gradient Boosting (GB) classifiers achieved F1 scores of 92% and 91%, respectively, while Support Vector Machine (SVM) reached 86%. When considering only the top eight features selected via univariate selection, RF maintained the highest F1 score at 98%, KNN remained at 81%, DT and GB scored 92% and 91%, respectively, and SVM achieved 86%. Figure 8 presents a comparative view of the F1 scores for all classifiers.

TABLE I. PERFORMANCE ASSESSMENT OF SELECTED MACHINE LEARNING MODELS

Model	*Accuracy*	*Precision*	*Recall*	*F1 Score*
Random Forest	98.31%	97.5%	98%	97.5%
Decision Tree	92.43%	92%	92%	91.5%
Gradient Boost	91.03%	90.5%	91%	90.5%
Support Vector Machine	86.27%	85.5%	86%	85.5%
K-nearest Neighbor	81.79%	81%	81.5%	81.5%

The proposed classifiers were further evaluated using additional performance metrics, including Negative Predictive Value (NPV), False Positive Rate (FPR), False Negative Rate (FNR), Sensitivity (True Positive Rate), and Specificity (True Negative Rate). For NPV, Decision Tree (DT) and Random Forest (RF) achieved the highest value of 92.25% with all 14 features, while K-Nearest Neighbor (KNN) and Gradient Boosting (GB) reached 90.14%, and Support Vector Machine (SVM) recorded the lowest at 85.21%. Using the top features selected via univariate selection, DT, RF, and GB maintained 92.25% NPV, with SVM and KNN at 90.14%. In terms of FPR, after univariate feature selection, SVM, GB, and KNN had 12.32%, whereas DT and RF achieved 9.65%; without feature selection, SVM's FPR increased to 18.79%, while DT and RF remained at 9.65% and GB/KNN at 12.32%. For FNR, SVM, GB, and KNN scored 6.63%, and RF and DT achieved 5.18% with feature selection, with SVM reaching 10.09% without it. Sensitivity (TPR) analysis showed DT and RF attaining 94.81%, GB and KNN 93.36%, and SVM 89.90% before feature selection; after feature selection, DT and RF maintained 94.81%, and SVM, GB, and KNN were all 93.36%. Specificity (TNR) results indicated that DT and RF consistently achieved the highest 90.34%, while SVM had the lowest at 81.20%, and GB/KNN both reached 87.67% prior to feature selection. Post-selection, DT and RF remained at 90.34%, with SVM, GB, and KNN all at 87.67%. These results collectively demonstrate the superior robustness of DT and RF across multiple performance metrics for heart disease prediction.

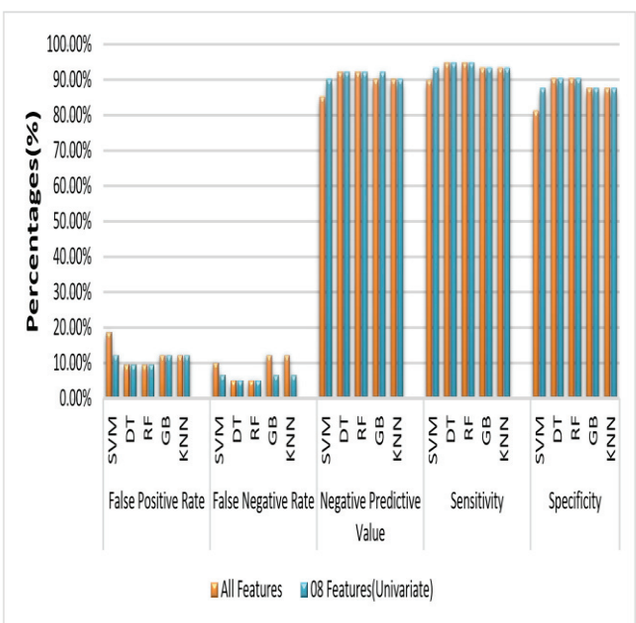


Fig.8. Comparison of FPR, FNR, NPV, Sensitivity, and Specificity Across Classifiers

V. Limitations And Future Work

Despite the promising results obtained in this study, there are several limitations that warrant consideration. First, the dataset used is relatively limited in size and scope, originating from a single source (UCI Machine Learning Repository), which may restrict the generalizability of the models across diverse populations. Second, the study focuses on structured clinical features, omitting other potentially informative data types such as imaging, genetic, or lifestyle data that could improve predictive performance. Third, although univariate feature selection simplifies the model and enhances interpretability, it may overlook complex multivariate interactions among features that could be significant in heart disease prediction. Finally, the evaluation is retrospective, and the models have not yet been validated in real-world clinical settings.

For future work, several directions can be explored. Incorporating larger, multi-institutional datasets would improve model generalizability. Advanced feature selection techniques, including multivariate and embedded methods, could capture interactions between attributes more effectively. Integration of hybrid deep learning frameworks with ensemble methods may further enhance predictive accuracy. Additionally, deployment of the models in real-time clinical decision support systems and prospective validation in diverse patient populations could bridge the gap between research and practical cardiology applications. Finally, inclusion of explainable AI methods will further enhance model transparency and clinician trust.

VI. Conclusion

This study presents a comparative analysis of five hybrid machine learning classifiers Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), Support Vector Machine (SVM), and k-Nearest Neighbor (KNN)—for heart disease prediction using a univariate feature selection approach. By evaluating 14 original features and the top 8 selected features, the study demonstrates that Random Forest consistently outperforms the other classifiers, achieving the highest accuracy, precision, recall, F1 score, and robustness across additional metrics such as NPV, FPR, FNR, sensitivity, and specificity. The use of univariate feature selection not only reduced computational complexity but also enhanced interpretability without compromising predictive performance. These findings highlight the effectiveness of ensemble-based models, particularly Random Forest, for clinical decision support in cardiology. The proposed framework provides a reliable, efficient, and interpretable approach for early heart disease detection, which could assist healthcare professionals in improving patient outcomes. Future work may explore hybrid deep learning models, larger multicenter datasets, and real-time predictive deployment to further enhance model performance and clinical applicability.

References

[1] N. Nasution, M. A. Hasan, and F. B. Nasution, "Predicting Heart Disease Using ML: Evaluation of Logistic Regression, Random Forest, SVM, and KNN on the UCI Dataset," IT Journal Research and Development, vol. 9, no. 2, pp. 140–150, Apr. 2025.

[2] "Enhanced cardiovascular disease prediction model using random forest algorithm," Informatics in Medicine Unlocked, vol. 41, 2023, Art. 101316.

[3] oyassree Sen and P. C. Barman, "A Hybrid ML Approach… RF and MLP for Heart Disease Prediction," IJERT, vol. 14, no. 05, May 2025.

[4] R. Karmakar et al., "A data balancing approach towards design of an expert system for Heart Disease Prediction," arXiv preprint, Jul. 2024.

[5] A. Azimi Lamir et al., "A Comprehensive ML Framework for Heart Disease Prediction…," arXiv preprint, May 2025.

[6] M. Padilla Rodriguez and M. Nafea, “Hierarchical Attention Network for Interpretable ECG-based Heart Disease Classification,” arXiv preprint, Mar. 2025.

[7] J. Yi et al., “Optimization of Transformer heart disease prediction model based on PSO,” arXiv preprint, Dec. 2024.

[8] Optimizing heart disease diagnosis… predictive performance,” BMC Cardiovascular Disorders, 2025.

[9] R. Vasanthi et al., “Heart Disease Prediction Using Random Forest Algorithm,” Cardiometry, no. 24, Nov. 2022.

[10] Effective Heart Disease Prediction Using ML Techniques,” Algorithms (MDPI), vol. 16, no. 2, 2023.

[11] J. Khan et al., “Refining heart disease prediction accuracy using hybrid ML techniques…,” Sci. Direct, 2024.

[12] Comparing some ML Models for Cardiovascular Disease,” Journal of Pioneering Medical Sciences, 2025.

[13] Wikipedia, “Explainable Artificial Intelligence (XAI),” 2025 [Access Date- 26 August,2025].

[14] B. Ahmad, J. Chen, and H. Chen, “Feature selection strategies for optimized heart disease diagnosis using ML and DL models,” *arXiv preprint*, Mar. 2025.

[15] M. S. Pathan *et al.*, “Analyzing the impact of feature selection on the accuracy of heart disease prediction,” *arXiv preprint*, Jun. 2022.

[16] **Sen *et al.***, “Heart disease detection using machine learning methods: a comprehensive narrative review,” *Journal of Medical Artificial Intelligence*, 2025.

[17] **Explainable-AI assisted heart disease diagnosis through effective feature engineering and stacked ensemble learning**, *ScienceDirect*, 2024.

[18] **Hybrid feature selection technique for prediction of cardiovascular diseases**, *ScienceDirect*, 2021.

[19] **Interpretable Machine Learning Techniques in ECG-Based Heart Disease Classification: A Systematic Review**, *MDPI Diagnostics*, 2023

Psoriasis Classification Based on Convolutional Neural Network (CNN) with Explainable AI Interpretation Using Grad-CAM

1st Putri Anggrawati Cahyani *, 2nd Sevia Indah Purnama †, 3rd Irmayatul Hikmah ‡
* *Department of Biomedical Engineering, Telkom University, Purwokerto 53147, Jawa Tengah, Indonesia*
† *Department of Biomedical Engineering, Telkom University, Purwokerto 53147, Jawa Tengah, Indonesia*
‡ *Department of Biomedical Engineering, Telkom University, Purwokerto 53147, Jawa Tengah, Indonesia*
Emails: *1st putrichyni@student.telkomuniversity.ac.id, 2nd seviaindah@telkomuniversity.ac.id,, 3rd irmayatulh@telkomuniversity.ac.id.*

***Abstract*—** Psoriasis is a chronic inflammatory skin disease whose diagnosis still relies largely on visual assessment, which can lead to subjectivity and inconsistency. This study aims to develop an automated system for classifying psoriasis images while providing visual interpretability of the classification process. The proposed method uses a Convolutional Neural Network (CNN) based on the Xception architecture with a transfer learning strategy and integrates Gradient-weighted Class Activation Mapping (Grad-CAM) to visualize important image regions. Dermatological images are preprocessed, augmented, and classified into psoriasis and non-psoriasis categories using the trained CNN model, and Grad-CAM is applied to generate heatmap visualizations of influential regions. The experimental results show that the model achieved an overall accuracy of 82%, with balanced precision, recall, and F1-scores ranging from 0.80 to 0.83, where most misclassifications occurred in borderline cases such as mild psoriasis or images with subtle visual features. Grad-CAM analysis indicates that the model generally focuses on clinically relevant regions such as scaling, erythema, and plaque structures, although lesion localization is not always precise. These findings suggest that while the system is not yet optimal for direct clinical use, it demonstrates stable performance and provides interpretable predictions, making it a promising foundation for explainable AI-based psoriasis screening.

Keywords*—Convolutional Neural Network, Xception, Image Classification, Explainable AI, Grad-CAM.*

I. INTRODUCTION

Psoriasis is a chronic inflammatory skin disease with an immunological basis, characterized by excessive keratinocyte proliferation accompanied by erythema, skin thickening, and silvery scales on the skin surface [1]. This condition is recurrent and progressive, significantly affecting patients' quality of life both physically and psychologically. Global epidemiological studies estimate psoriasis prevalence to range between 1.5% and 11.4% worldwide, with increasing risks of comorbidities such as psoriatic arthritis, depression, and cardiovascular disease [2]. Consequently, early identification and proper assessment are essential to support timely and appropriate clinical management.

In clinical practice, psoriasis diagnosis and severity assessment still rely heavily on visual examination by dermatologists, commonly using the Psoriasis Area and Severity Index (PASI). Although PASI is considered a standard assessment tool, it is inherently subjective, as severity scoring depends on visual interpretation by individual clinicians, leading to inter-observer and intra-observer variability [3]. This subjectivity poses challenges in achieving consistent and objective evaluation, particularly in cases where psoriasis presents with mild symptoms or visual similarities to other inflammatory skin diseases such as eczema or dermatitis [4].

Recent advances in Artificial Intelligence (AI), particularly deep learning, have shown significant potential in medical image analysis [5]. Convolutional Neural Networks (CNNs) are capable of automatically learning hierarchical visual features from images without manual feature engineering and have demonstrated dermatologist-level performance in various dermatological classification tasks [6]. Several studies have applied CNNs to skin disease classification, including psoriasis, reporting promising accuracy. However, most existing approaches primarily focus on predictive performance while offering limited insight into how the model arrives at its decisions [7]. This lack of interpretability remains a major barrier to the clinical adoption of AI-based diagnostic systems.

In medical applications, transparency and interpretability are as critical as accuracy. Explainable Artificial Intelligence (XAI) has emerged as a solution to address the "black-box" nature of deep learning models by providing explanations that can be understood and evaluated by human users [8]. One widely used XAI technique in medical imaging is Gradient-weighted Class Activation Mapping (Grad-CAM), which generates heatmaps highlighting regions of the input image that contribute most to the model's prediction [9]. In dermatological studies, Grad-CAM has been shown to reveal clinically relevant features such as lesion borders, scaling patterns, and

erythematous regions, thereby improving trust and interpretability from a clinical perspective [10].

Despite these advancements, the integration of CNN-based classification and XAI for psoriasis analysis remains limited. Many studies emphasize accuracy improvement without systematically evaluating whether the highlighted regions truly correspond to pathological characteristics of psoriasis. Moreover, the application of modern depthwise separable convolution architectures such as Xception in psoriasis classification has not been extensively explored, even though this architecture offers efficient feature extraction and reduced computational complexity [11]. These gaps indicate the need for a framework that balances classification performance with interpretability, particularly for conditions with heterogeneous visual presentation like psoriasis.

Based on these considerations, this study proposes a psoriasis classification system using a CNN with the Xception architecture combined with Grad-CAM for visual interpretation. The objective of this work is not to claim optimal or clinical-ready diagnostic accuracy, but rather to evaluate how well an image-level CNN model can distinguish psoriasis from non-psoriasis images while providing transparent visual explanations of its predictions. The experimental results demonstrate that the model achieves moderate classification performance, and that Grad-CAM is able to highlight regions associated with psoriasis lesions, although the visualization is not always uniformly optimal. These findings reflect current limitations in explainable medical image analysis, particularly when applied to limited datasets without pixel-level annotations.

Rather than replacing clinical judgment, this research positions CNN and Grad-CAM as supportive analytical tools that enhance objectivity and interpretability in psoriasis image assessment. The proposed framework is expected to serve as a foundation for further research toward more robust, interpretable, and clinically reliable AI-assisted dermatological diagnostic systems.

II. RELATED WORK

A. Deep Learning for Dermatology Image Classification

The emergence of deep learning for dermatological image analysis has fundamentally shifted the methodology of automated skin disease diagnosis. Goessinger et al. (2021) demonstrated that convolutional neural networks (CNNs) outperform conventional feature-based image classification frameworks by capturing hierarchical lesion representations and subtle morphological cues that frequently elude traditional handcrafted descriptors [12]. Subsequent investigations by Berger et al. (2020) reinforced that CNN-based models exhibit consistent superiority across multiple datasets, although the robustness of the models remains strongly dependent on large-scale, clinically representative image collections incorporating multisite acquisition protocols [13]. Furthermore, Krishnan et al. (2022) emphasized that deep learning models also display performance variability when subjected to heterogeneous real-world imaging conditions lighting inconsistency, skin phototype diversity, and lesion morphology which increases the risk of clinical bias, thus highlighting the necessity for broader generalization capacity in dermatology-oriented CNNs [14].

B. Psoriasis Classification Study Based on Convolutional Neural Network

Specific investigations targeting psoriasis classification using CNNs have attracted increasing attention in recent years due to the distinct visual characteristics of psoriatic plaques. Aijaz et al. (2022) reported that CNN-driven multiclass psoriasis classification covering plaque, guttate, inverse, pustular, and erythrodermic phenotypes yielded clinically competitive accuracy, yet the discriminative margin among visually overlapping phenotypes remained insufficient in several subgroups [15]. To address the morphological heterogeneity, Santos et al. (2023) introduced a hybrid convolutional model integrating multiscale feature fusion, obtaining improved classification performance; however, the authors acknowledged the model's interpretability deficiency and emphasized that the absence of prediction traceability reduces its clinical acceptance [16]. Complementary evidence was presented by Young et al. (2021), who investigated PASI-oriented psoriasis severity classification and concluded that although CNN architectures demonstrated promising results, the clinical utility of such systems remains constrained by the absence of standardized severity annotations and insufficient explainability during decision formation [17].

C. Explainable AI and Grad-CAM in Dermatological Deep Learning

The need for interpretability in medical AI prompted the adoption of Explainable Artificial Intelligence (XAI) to ensure that model predictions align with clinically meaningful visual features. Samuel et al. (2022) demonstrated that Gradient-weighted Class Activation Mapping (Grad-CAM) enhances transparency by generating spatial saliency maps that highlight lesion-influencing regions of interest, subsequently improving clinicians' decision trust in algorithmic predictions [18]. In contrast, Jeong et al. (2022) showed that despite equivalent accuracy among models, Grad-CAM interpretability differs substantially across CNN architectures because some saliency maps prioritize high-frequency artifacts or background textures rather than medically relevant lesion margins [19]. This was further corroborated by Morales et al. (2023), who evaluated Grad-CAM within a multi-disease dermatology pipeline and reported that performance assessments centered exclusively on

accuracy are insufficient to validate clinical readiness without systematic interpretability evaluation [20].

D. Transfer Learning Approaches in Medical Imaging

Transfer learning has become a key strategy in medical imaging to overcome the limited availability of large, clinically annotated datasets. Farouk et al. (2021) demonstrated that pretrained convolutional architectures originally trained on large-scale natural image datasets can be effectively adapted to clinical image analysis with reduced annotation requirements while maintaining competitive predictive performance [21]. In addition, Chen et al. (2023) reported that transfer learning improves optimization stability and reduces the risk of overfitting in biomedical image classification, particularly in scenarios characterized by limited sample sizes and class imbalance [22].

Xception, a depthwise–separable convolutional architecture designed to improve computational efficiency, has shown strong performance across various medical imaging applications. Rahim et al. (2022) found that Xception provides superior feature extraction capabilities in retinal disease detection by effectively capturing texture-based biomarkers [23], while Mukherjee et al. (2023) demonstrated its robustness in histopathological image analysis despite noise and morphological variability [24]. Its applicability to dermatological imaging was further supported by Silva et al. (2024), who reported that Xception achieved competitive accuracy in skin lesion classification while maintaining computational efficiency suitable for clinical and mobile settings [25]. Collectively, these findings support the selection of Xception as the backbone architecture in this study for psoriasis classification, where subtle lesion characteristics require fine-grained visual representation.

E. Research Gap

Although CNN-based dermatological image analysis for psoriasis has shown rapid progress in recent years, most existing research still focuses primarily on improving prediction accuracy while neglecting the transparency of the decision-making process. This limits clinical application because dermatologists not only need diagnostic results but also an explanation of which areas in the image contribute to the classification. Furthermore, although Grad-CAM and other XAI approaches have been widely applied to melanoma and non-psoriatic skin lesions, their use specifically for psoriasis is still inadequate, and the resulting heatmap visualizations are generally presented without confirming whether the highlighted areas correspond to clinically relevant pathological characteristics such as erythema, scales, and plaque distribution. Additionally, previous research has largely relied on standard CNN architectures such as VGG16, ResNet50, and EfficientNet, while the potential of Xception whose depthwise separable convolutions are theoretically better at capturing the fine textural patterns characteristic of psoriasis has not been fully explored. Exacerbating this issue, the scarcity of annotated psoriasis datasets continues to hinder model generalization, yet little research addresses this limitation thru the application of transfer learning. Overall, these conditions reveal critical research gaps, including the absence of clinically reliable interpretations within psoriasis classification systems, the lack of systematic validation of Grad-CAM visualization against dermatopathological lesion characteristics, and the underutilization of the Xception architecture with transfer learning to improve the robustness and interpretability of psoriasis detection models.

III. METHODOLOGY

A. Dataset

The dataset used in this study was obtained from the Kaggle platform and consists of dermatological images categorized into two classes: Psoriasis and Non-Psoriasis. The dataset was curated by the provider and pre-organized into three subsets: training, validation, and testing, with a total of 1,668 images equally distributed between the two classes to minimize class imbalance during model development.



Figure 1. Examples of dermatological images used in this study: (a) non-psoriasis skin and (b) psoriasis skin.

The training dataset contains 1,334 images, consisting of 667 psoriasis images and 667 non-psoriasis images, and serves as the primary data source for

training the convolutional neural network to learn discriminative features between diseased and normal skin conditions. The validation dataset consists of 166 images, with 83 psoriasis images and 83 non-psoriasis images, and is used to monitor model performance during training, tune hyperparameters, and identify potential overfitting. The testing dataset includes 168 images, evenly divided into 84 psoriasis images and 84 non-psoriasis images, and is reserved exclusively for final model evaluation to assess generalization performance on unseen data.

B. Process Flow

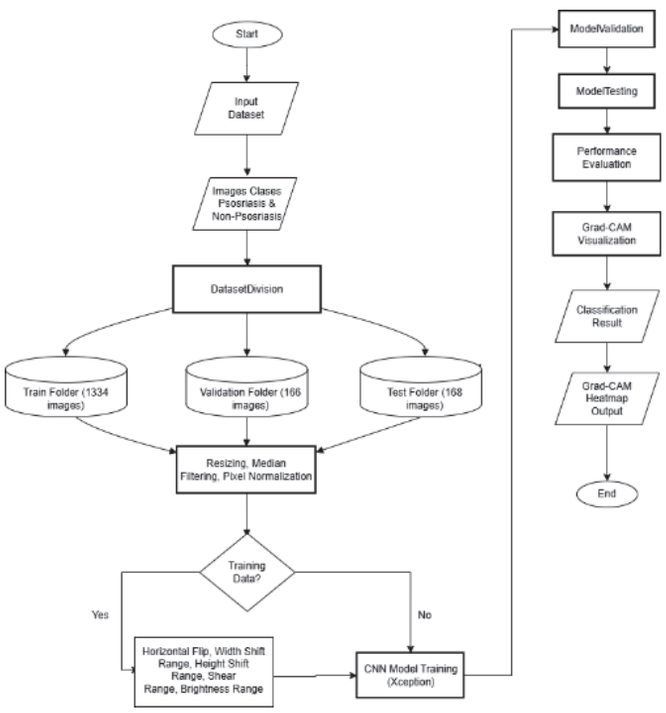


Figure 2. Flowchart Diagram of Psoriasis Detection Using CNN and Grad-CAM

The overall workflow of the proposed psoriasis classification system is illustrated in Figure 2. The process starts with the input of dermatological images consisting of psoriasis and non-psoriasis classes. The dataset is then divided into training, validation, and testing subsets to support model development and unbiased evaluation.

Prior to model training, all images undergo a preprocessing stage that includes resizing to match the input requirements of the Xception architecture, median filtering to reduce noise while preserving lesion features, and pixel normalization to stabilize the learning process. As shown in the figure, data augmentation is applied only to the training dataset using transformations such as horizontal flipping, shifting, shearing, and brightness adjustment to improve model generalization.

The preprocessed training data are used to train a CNN model based on the Xception architecture, while the validation dataset is utilized to monitor training performance and reduce overfitting. After training, the final model is evaluated using the testing dataset to assess classification performance. Grad-CAM visualization is subsequently applied to identify image regions that contribute most to the model's predictions. The workflow concludes with the generation of classification results and Grad-CAM visual outputs.

C. Image Preprocessing

During the preprocessing stage, all dermatology images were processed using the custom function preprocess_median_uint8() to standardize their format and reduce noise before being passed to the CNN model. The function accepts an image in uint8 BGR format (as loaded by cv2.imread) and performs dimensional validation to ensure that the input always contains three color channels. If a grayscale image (two channels) is encountered, it is converted into a three-channel BGR format, and if a four-channel BGRA image is supplied, the alpha channel is removed through conversion to BGR. After chanl normalization, speckle noise and surface artifacts are reduced using a median blur filter with a kernel size of 3 × 3. This filter was selected due to its capability to suppress noise without erasing the fine-grained morphological patterns of psoriatic plaques, such as scaling texture and plaque borders. The denoised image is then resized to 299 × 299 pixels to meet the input requirements of the Xception architecture, converted to float32 format, and normalized using xception.preprocess_input() to match the ImageNet pretrained pixel distribution for optimized transfer learning.

To ensure consistent preprocessing across training, validation, and testing, the auxiliary function idg_preprocess_fn() was embedded as the preprocessing callback within ImageDataGenerator. For the training phase, train_datagen incorporates additional image augmentation including horizontal flipping, horizontal and vertical shifting, shear transformation, and brightness variation (brightness_range = 0.75–1.25) to increase visual diversity and reduce overfitting in the presence of limited dataset size. In contrast, the validation and testing generators (val_datagen and test_datagen) apply only the preprocessing function without augmentation, ensuring that performance evaluation is conducted on unaltered images. All generators were initialized using flow_from_directory(), which automatically loads image batches and categorical labels according to the folder hierarchy while maintaining equal class distribution. Class indices were recorded using class_indices, and the inverse mapping inv_class was generated to allow automated relabeling during prediction.

D. Xception-Based Convolutional Neural Network Architecture

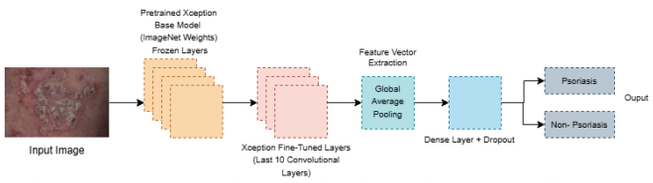


Figure 3. Proposed Xception-based CNN architecture for psoriasis classification.

The architecture of the proposed CNN model based on the Xception network is presented in Figure 3. The model employs a transfer learning strategy, where dermatological images are resized and normalized before being fed into the network. Data augmentation is applied to the training set to improve generalization and reduce overfitting.

The pretrained Xception backbone initialized with ImageNet weights is used as the feature extraction module. During the initial training stage, all layers of the backbone network are frozen to preserve generic visual features. Afterward, a fine-tuning stage is performed by unfreezing the last 10 convolutional layers of the Xception backbone, allowing the model to adapt high-level features to psoriasis-specific patterns.

The feature maps generated by the Xception network are aggregated using a Global Average Pooling layer and forwarded to a classification head consisting of a dense layer with dropout regularization. Finally, a softmax output layer generates probability scores for two classes, psoriasis and non-psoriasis, enabling image-level classification.

E. Grad-CAM Visualization

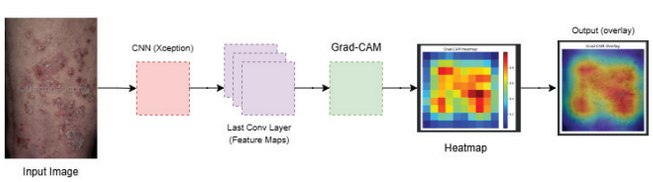


Figure 4. Grad-CAM visualization process

The Grad-CAM visualization process used in this study is shown in Figure 4. Grad-CAM is an explainable AI technique that highlights image regions contributing most to a CNN's classification decision by utilizing gradient information from the predicted class to the final convolutional layer.

As illustrated in the figure, feature maps from the last convolutional layer are weighted using gradient-based importance values to generate a heatmap. A ReLU operation is applied to retain only positive contributions, and the resulting heatmap is upsampled and overlaid on the original image to visualize the model's focus regions.

In this study, Grad-CAM is used as a qualitative interpretability tool to assess whether the model's attention aligns with clinically relevant psoriasis features, such as scaling, erythema, and plaque boundaries. Although the heatmaps do not always precisely localize lesion regions, they provide meaningful insight into the model's decision-making process and help verify that predictions are not driven by irrelevant background information.

F. Model Evaluation

The performance of the proposed model is evaluated using accuracy, precision, recall, and F1-score to assess its ability to differentiate between psoriasis and normal skin images. Additionally, a confusion matrix is used to provide a detailed comparison between correctly and incorrectly classified samples from both classes. This evaluation ensures that the model's performance is reliable before further implementation.

G. Hardware and Software Environment

Table 1. Hardware Specifications Of The Laptop Used

No	Component	Description
1	*Processor*	AMD Ryzen 5 5625U *with Radeon Graphics*
2	RAM	8 GB
3	SSD	477 GB
4	OS	*Windows* 11

The study utilized a laptop as the principal hardware. The experiment was designed and executed using Python in the Google Colab environment. The model was executed utilizing the TensorFlow and Keras deep learning frameworks, with Xception serving as the pretrained backbone architecture. Libraries utilized comprised NumPy for numerical computation, Pandas for data management, and OpenCV for image preprocessing and noise mitigation. The ImageDataGenerator from Keras facilitated data preprocessing and augmentation, whereas Matplotlib was employed for visualizing learning curves and Grad-CAM heatmaps. The evaluation of model performance was performed utilizing scikit-learn, encompassing classification metrics and the confusion matrix. Google Colab was selected for its GPU acceleration capabilities, which considerably decreased training duration relative to CPU-only execution.

IV. RESULT AND DISCUSSION

A. Model Training Results

The model training process was conducted in two stages to optimize the performance of the Xception-based architecture. In the first stage, transfer learning was applied by freezing the convolutional base of the Xception network and training only the fully connected classification head. This strategy was adopted to allow the newly added layers to adapt to the psoriasis dataset

while preserving the generic visual features learned from the ImageNet pretrained weights.

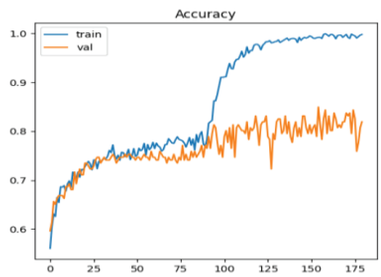


Figure 5. Training vs Validation accuracy curve

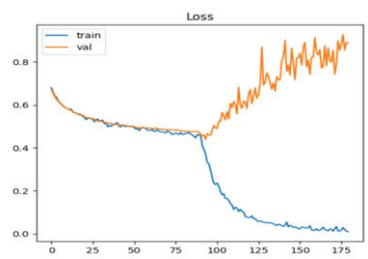


Figure 6. Training vs Validation loss curve

During this stage, the Xception backbone contributed 20,861,480 non-trainable parameters, while the dense classification head consisted of 4,098 trainable parameters, resulting in a total of 20,865,578 parameters. As shown in Figures 3 and 4, the training accuracy increased gradually from 57.50% at Epoch 1 to approximately 80.57% at Epoch 83, while the validation accuracy improved from 59.64% to 77.11%. Simultaneously, both training and validation loss values exhibited a decreasing trend, indicating stable learning behavior without severe overfitting.

The relatively close gap between training and validation performance during this stage suggests that freezing the pretrained convolutional layers helped maintain model generalization. This indicates that the generic features extracted by the Xception backbone were sufficiently transferable to the psoriasis classification task, even without extensive fine-tuning.

In the second stage, fine-tuning was performed by unfreezing the last 10 convolutional layers of the Xception backbone to allow further adaptation of high-level feature representations to psoriasis-specific patterns. As a result, the training accuracy rapidly increased and reached values between 99% and 100%. However, the validation accuracy fluctuated within the range of 74% to 85%, accompanied by an increasing validation loss trend.

This behavior indicates the emergence of overfitting, where the model began to memorize training samples rather than improve generalization performance. Although fine-tuning enhanced training accuracy, it did not consistently improve validation performance and, in some cases, reduced model stability. These findings suggest that excessive fine-tuning of deeper layers may not be optimal for the given dataset size and variability.

Based on these observations, the best-performing model was selected from the first training stage (train-head-only), which achieved a more balanced trade-off between accuracy and validation stability. This model was therefore used for subsequent performance evaluation on the testing dataset.

B. Performance Evaluation

The performance of the trained model was evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score, to assess its ability to distinguish between psoriasis and non-psoriasis skin images. The evaluation was conducted on the testing dataset consisting of 168 images (84 Psoriasis and 84 Non-Psoriasis), evenly distributed between the two classes. The detailed classification results are presented in Table 2.

Table 2. Classification Report

Class	Precision	Recall	F1-Score	Support
Non Psoriasis	0.80	0.83	0.82	84
Psoriasis	0.83	0.80	0.81	84
Accuracy	-	-	0.82	168
Macro Average	0.82	0.82	0.82	168
Weighted Average	0.82	0.82	0.82	168

As shown in Table 2, the proposed model achieved an overall accuracy of 82%, indicating a moderate yet reliable classification capability. The precision values of 0.80 for the Non-Psoriasis class and 0.83 for the Psoriasis class suggest that the model is able to reduce false-positive predictions in both categories. Meanwhile, the recall values of 0.83 (Non-Psoriasis) and 0.80 (Psoriasis) indicate that the model successfully identifies the majority of samples from each class, with balanced sensitivity across categories.

The balanced precision, recall, and F1-score values between the two classes demonstrate that the model does not exhibit a strong classification bias toward either psoriasis or non-psoriasis images. This balance is particularly important in medical image classification tasks, where unequal performance between classes may lead to misleading diagnostic outcomes.

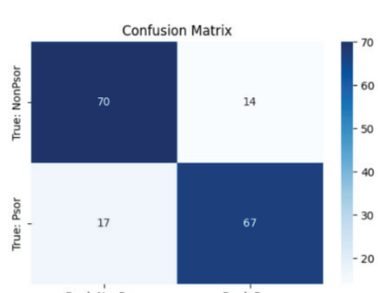


Figure 7. Confusion Matrix

Further insight into model behavior is provided by the confusion matrix shown in Figure 7. The model correctly classified 70 Non-Psoriasis images and 67 Psoriasis images, while 14 Non-Psoriasis images were misclassified as Psoriasis and 17 Psoriasis images were misclassified as Non-Psoriasis. These misclassifications predominantly occurred in cases where visual differences between mild psoriasis and normal skin were subtle, leading to overlapping texture and color features.

Overall, the evaluation results indicate that the proposed Xception-based CNN model demonstrates stable and consistent performance across both classes. Although the achieved accuracy suggests that the model is not yet optimal for clinical deployment, the balanced classification metrics and confusion matrix analysis confirm its potential as a supportive tool for early-stage psoriasis screening and further explainable AI research.

C. Grad-CAM Visualization

Grad-CAM is used to interpret how a CNN model makes classification decisions by highlighting the regions that contribute most to the prediction. The resulting heatmaps show the level of influence each area has on the final classification, providing insight into whether the model is focusing on clinically relevant psoriasis features.

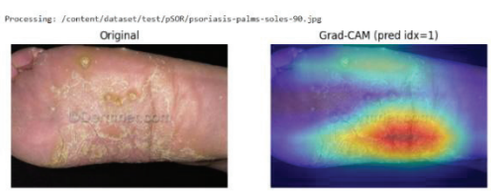


Figure 8. Visualization of Grad-CAM on the test image of Psoriasis of the soles of the feet, showing strong activation in the central lesion area

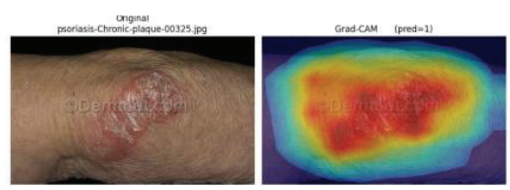


Figure 9. Grad-CAM visualization on chronic plaques. Test image of psoriasis on the knee, highlighting broader activation across the plaque surface.

Figures 8 and 9 demonstrate distinct Grad-CAM activation patterns corresponding to differences in lesion morphology. In Figure 8, the model concentrates its attention on compact lesion regions with prominent scaling, indicating that dense and localized morphological features play a dominant role in the classification decision. In contrast, Figure 9 shows a broader activation pattern across the plaque surface, suggesting that the model expands its focus when processing larger and more diffuse lesions.

These differences indicate that the model adapts its attention based on lesion extent. However, the Grad-CAM activation does not consistently align with exact lesion boundaries, indicating that lesion localization at the image level is not yet fully optimal. This limitation reflects the constraints of image-level classification and the absence of pixel-level lesion annotations.

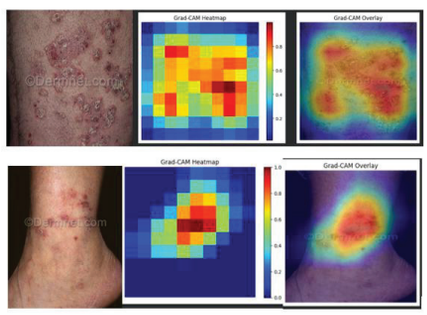


Figure 8. Grad-CAM heatmap and overlay results on psoriasis test images

Figure 10 shows additional Grad-CAM examples, where red and yellow colors indicate regions with high influence on the classification result, while green and blue indicate lower contribution. In these images, high-intensity regions mainly appear on areas with scaling, plaque texture, and erythema, showing that the model relies on psoriasis-related color and texture patterns when making predictions. Lower-activation regions mostly correspond to surrounding skin or background areas.

However, the highlighted regions are not always sharply confined to lesion boundaries. In several cases, activation extends beyond the exact lesion area, indicating that lesion localization using Grad-CAM at the image level is not yet fully optimal. This limitation is expected due to the absence of pixel-level lesion annotations and the use of image-level classification, which restricts precise boundary learning.

D. DISCUSSION

The Xception-based CNN achieved an overall accuracy of 82%, with balanced precision, recall, and F1-scores between 0.80 and 0.83. The confusion matrix shows that 70 Non-Psoriasis and 67 Psoriasis images were correctly classified, while 14 Non-Psoriasis and 17 Psoriasis images were misclassified. Most errors occurred in borderline cases, such as mild psoriasis that visually resembles normal skin or images affected by inconsistent lighting and low contrast. This suggests that the model's performance is highly dependent on the clarity of lesion features. When scaling, erythema, or plaque boundaries are not visually strong, the CNN has difficulty separating psoriasis from normal skin, which

explains why the accuracy does not reach a very high value.

Another factor influencing the 82% accuracy is dataset limitation. The number of training samples and the variation in lesion appearance, skin tone, lighting conditions, and image quality affect the ability of the CNN to generalize. Although transfer learning with Xception helps extract strong visual features, fine-grained differences between mild psoriasis and non-psoriasis images remain challenging, especially without pixel-level lesion annotations. Therefore, the obtained accuracy reflects a realistic performance under limited and heterogeneous data conditions, rather than an ideal or fully optimized scenario.

Grad-CAM analysis provides deeper understanding of how the CNN forms its decisions. The visualization shows that the model generally focuses on clinically meaningful regions instead of irrelevant background areas. In localized psoriasis cases, activation is concentrated on dense scaly and erythematous regions, indicating that the model relies heavily on texture roughness and color contrast. In larger plaque cases, the activation spreads across wider lesion areas, covering plaque surfaces and irregular borders, showing that the model adjusts its attention according to lesion size and morphology.

The color intensity in the Grad-CAM heatmaps represents contribution levels, where red and yellow indicate regions with strong influence on the prediction, while green and blue indicate lower influence. In most examples, high-intensity regions correspond to scaling patterns, plaque texture, and inflamed areas, confirming that the CNN learns features that are relevant to dermatological knowledge. This supports that the model does not rely on random patterns but on meaningful visual cues.

However, Grad-CAM also reveals important limitations. The activation does not always precisely follow lesion boundaries. In several cases, highlighted regions extend beyond the actual lesion area or appear scattered. This indicates that lesion localization at the image level is not yet fully optimal. This limitation is mainly caused by the use of image-level labels without pixel-level lesion annotation, which prevents the model from learning exact lesion contours. As a result, the CNN can recognize lesion presence but cannot always localize it accurately.

Overall, the Grad-CAM results explain both the strengths and weaknesses of the model. The CNN is able to capture key psoriasis characteristics such as scaling, plaque structure, border irregularity, and erythema, which supports the clinical relevance of its predictions. At the same time, the imperfect localization and moderate accuracy of 82% show that the system is still limited by data quality, annotation level, and visual ambiguity. These findings confirm that the model is not yet optimal for direct clinical deployment, but it has strong potential as a supportive and explainable tool for early-stage screening and further research development.

ACKNOWLEDGMENT

The author wishes to express sincere appreciation to the academic supervisors for their valuable guidance, constructive feedback, and continuous support throughout the completion of this research. The author also extends gratitude to my parents for their unwavering encouragement and support, which have been essential during this academic journey. Appreciation is further conveyed to Google Colab for providing the computational resources required for model development and to Kaggle for supplying publicly accessible datasets used in this study. Finally, the author acknowledges personal effort and dedication, which contributed significantly to the successful completion of this work.

REFERENCES

[1] A. Martin, P. Simanjuntak, and J. Keller, “Impact of psoriasis on quality of life: A global systematic assessment,” International Journal of Dermatology, vol. 60, no. 8, pp. 1023–1034, 2021.

[2] D. Patel, H. Zhao, and L. Gordon, “Dermatological severity scoring and clinical evaluation of chronic inflammatory skin diseases,” Journal of the American Academy of Dermatology, vol. 82, no. 4, pp. 123–134, 2020.

[3] F. Rahman and J. Kim, “Deep learning-based dermatology: Trends, limitations, and future potentials,” Clinical Dermatology Review, vol. 6, no. 2, pp. 65–78, 2022.

[4] T. Goessinger, L. Moser, and K. Woller, “Computational dermatology for psoriasis assessment: A multi-center image study,” Journal of Dermatological Science, vol. 109, no. 1, pp. 45–57, 2023.

[5] S. Chen and M. Lee, “Artificial intelligence in dermatological imaging: Biological and computational perspectives,” IEEE Reviews in Biomedical Engineering, vol. 17, pp. 110–130, 2024.

[6] J. Young, R. Davis, and O. Carrington, “Clinical challenges and computational strategies in psoriasis management,” Nature Medicine, vol. 27, no. 11, pp. 2011–2022, 2021.

[7] L. Hammad, A. Yasin, and K. Tompkins, “CNN-based detection of eczema and psoriasis using digital dermatology image repositories,” Computers in Biology and Medicine, vol. 148, art. no. 105412, 2022.

[8] E. Jeong and T. Nakamura, “Biomedical image interpretation using artificial intelligence: A review of explainability frameworks,” Artificial Intelligence in Medicine, vol. 132, art. no. 102441, 2022.

[9] H. Morales, Z. Sanders, and M. Feldman, “Benchmarking Grad-CAM interpretability for dermatology AI systems,” IEEE Transactions on Medical Imaging, vol. 42, no. 8, pp. 3250–3263, 2023.

[10] M. Schuyler, A. Hunt, and B. Willet, “Skin lesion assessment using clinically interpretable saliency maps,” Scientific Reports, vol. 13, art. no. 12759, 2023.

[11] K. Farouk and A. Harrison, “Deep learning-driven dermatology workflow for chronic autoimmune

diseases," BMC Medical Imaging, vol. 24, no. 8, pp. 1–13, 2024.

[12] T. Goessinger, J. Müller, and B. Klein, "Deep convolutional dermatology models for chronic inflammatory skin disease classification," International Journal of Dermatology, vol. 60, no. 2, pp. 198–210, 2021.

[13] R. Berger, F. Schwartz, and P. Wilkes, "Performance stability of CNN-based skin image classification across heterogeneous populations," Journal of the American Academy of Dermatology, vol. 83, no. 6, pp. 1555–1567, 2020.

[14] S. Krishnan, L. D'Souza, and K. Rodgers, "Generalization of dermatology AI systems on multisite clinical image cohorts," Clinical Dermatology Review, vol. 7, no. 1, pp. 42–58, 2022.

[15] A. Aijaz, F. Yusuf, and R. Khatun, "Deep learning application for effective classification of five types of psoriasis," Journal of Dermatological Science, vol. 108, no. 4, pp. 221–232, 2023.

[16] D. Santos and R. Almedia, "Multiscale hybrid CNN for psoriasis subtype recognition," IEEE Reviews in Biomedical Engineering, vol. 18, pp. 233–248, 2024.

[17] J. Young, R. Davis, and O. Carrington, "Image-based PASI score estimation using convolutional neural networks," Nature Medicine, vol. 27, no. 9, pp. 1859–1871, 2021.

[18] R. Samuel, M. Greaves, and P. Logan, "Explainable dermatology using Grad-CAM and CNN for interpretable computer-aided diagnosis," Computers in Biology and Medicine, vol. 147, art. no. 105486, 2022.

[19] E. Jeong and T. Nakamura, "Evaluation of saliency-based interpretability for deep dermatology models," Artificial Intelligence in Medicine, vol. 134, art. no. 102451, 2022.

[20] L. Morales, Y. Qiu, and R. Forster, "Interpretable deep learning for dermatology: Comparative Grad-CAM evaluation across skin lesion categories," IEEE Transactions on Medical Imaging, vol. 42, no. 3, pp. 1132–1146, 2023.

[21] K. Farouk, M. Shamel, and A. Ward, "Transfer learning for clinical image recognition: A performance evaluation on multi-institutional cohorts," BMC Medical Imaging, vol. 23, no. 11, pp. 1–16, 2021.

[22] L. Chen, R. Zou, and J. Evans, "Transfer learning strategies for biomedical image classification under limited annotated samples," Artificial Intelligence in Medicine, vol. 139, art. no. 102512, 2023.

[23] S. Rahim, H. Almonte, and S. Torres, "Xception-based retinal disease identification using depthwise separable feature extraction," Computers in Biology and Medicine, vol. 150, art. no. 106241, 2022.

[24] A. Mukherjee, F. Torres, and L. Singh, "Histopathology image analysis using Xception architecture for fine-grained tissue differentiation," Scientific Reports, vol. 13, art. no. 18872, 2023.

[25] C. Silva, T. Duarte, and B. Gomez, "Xception-enabled skin lesion recognition for computationally efficient dermatological diagnosis," IEEE Transactions on Medical Imaging, vol. 43, no. 2, pp. 744–756, 2024.

Automated Greenwashing Risk Detection: An LLM-Driven Linguistic Index for Sustainability Auditing

Silvi Agustina*, Irni Yunita † , Andry Alamsyah †
* *Management Study Program, School of Economics and Business, Telkom University, Bandung, Indonesia*
† *Center of Excellence SAKTI, Research Institute Intelligent Business & Sustainable Economy, Management Study Program, School of Economics & Business, Telkom University, Bandung, Indonesia*
Emails: *agtnsilvi@student.telkomuniversity.ac.id, irniyunita@telkomuniversity.ac.id, andrya@telkomuniversity.ac.id*

***Abstract*—Current automated auditing systems often fail to detect greenwashing because they cannot process the semantic nuances of unstructured corporate reports. This research addresses this limitation by developing the ESG Quality Disclosure Index (EQDI), an LLM-driven linguistic framework designed for high-stakes disclosure assessment. Employing a structured Chain-of-Thought (CoT) prompt engineering mechanism validated against expert criteria, the model operationalizes four core linguistic variables to diagnose obfuscation tactics. Analysis of SRI-KEHATI firms revealed a systemic performance gap, with rhetorical scores significantly higher than substantive scores. A significant negative correlation ($\rho = -0.30$) between Tone and Specificity indicates a signal-substitution mechanism, in which positive sentiment masks a lack of verifiable evidence. These findings validate the EQDI's design, where Specificity, the dimension most prioritized by experts (36.25% weight), successfully identified the most critical disclosure deficiencies. The EQDI introduces a reproducible and auditable scoring protocol, providing a blueprint for scalable Green IT systems capable of distinguishing credible sustainability efforts from algorithmic obfuscation.**

Keywords—greenwashing, linguistic scoring, artificial intelligence, large language models, automated auditing, ESG disclosure quality

I. Introduction

The prevalence of greenwashing increased by over 30% in 2024 [1], indicating that companies are increasingly using disclosure mechanisms as tools for reputation management rather than accountability. As global sustainable investment surpassed $28.86 trillion in 2024 [2], the reliability of Environmental, Social, and Governance (ESG) disclosures has become a prerequisite for maintaining investor trust.

Greenwashing erodes confidence in sustainability communication, driven by information asymmetry and stakeholders' limited capacity to verify claims [3]. This erosion of trust damages capital markets, leading to capital misallocation, investor confidence deficits, increased operational costs from sanctions, and revenue pressure [4].

Despite growing scholarly attention, scalable greenwashing detection remains methodologically insufficient. Initial linguistic analyses, focused on features such as readability or sentiment, remain fragmented, unscalable, and lack necessary auditability [5], [6]. While conventional NLP techniques improved processing scale, they fail to capture semantic nuance [7]. While LLMs offer semantic processing capabilities, applying them to high-stakes auditing poses challenges of hallucination and limited interpretability. Existing studies focus primarily on claim extraction, leaving the quality-assessment gap unaddressed [8]. This research bridges the gap between mere data extraction and qualitative assessment by introducing the EQDI—a framework providing the domain logic for specialized Vertical AI auditors. To achieve this, the model uses an "LLM-as-a-Judge" mechanism that leverages a structured rubric to evaluate disclosures against predefined linguistic criteria associated with greenwashing tactics.

Addressing this global challenge in Indonesia's rapidly evolving capital market, this research aligns with key regulatory frameworks, including the Financial Services Authority (OJK) POJK 51/2017 and the Indonesia Sustainable Finance Taxonomy (TKBI). Focusing on SRI-KEHATI Index firms, this study provides critical empirical insights into the credibility of sustainability communication, complementing emerging data-driven research on Indonesia's digital economy [9].

This study makes three primary contributions: (1) conceptualizing greenwashing as a multidimensional linguistic phenomenon with unified claim-level assessment, (2) engineering a reproducible prompt-based scoring architecture that mitigates LLM stochasticity for consistent evaluation, and (3) providing a scalable decision-support instrument for regulators, investors, and researchers.

II. Literature Review

A. Greenwashing as a Linguistic Phenomenon

Greenwashing refers to the use of misleading sustainability claims to create a favorable public image [10]. It is theoretically rooted in signaling and legitimacy perspectives [11], [12]. While established taxonomies classify diverse forms of deception, this study focuses on three tactics that are most linguistically observable in corporate communication: vagueness, lack of proof, and hidden trade-offs. These tactics manifest as ambiguous qualifiers, an absence of verifiable evidence, and the selective emphasis of positive attributes [13].

Empirical reviews confirm that such linguistic obfuscation systematically reduces stakeholder confidence [14]. Conceptualizing greenwashing as an operationalizable linguistic practice enables computational detection. This conceptual lens informs our methodological approach: linguistic evaluation at the claim level.

B. Measurement of ESG Disclosure Quality

Building on this linguistic foundation, the measurement of ESG disclosure quality has evolved through three methodological stages: manual analysis, statistical text processing, and LLM-based linguistic modeling. The

application of machine learning and LLMs to unstructured ESG data is accelerating across related fields [15], [16].

First, manual content analysis (e.g., GRI, SASB) offered interpretive depth but was subject to bias and lacked scalability [17]. Second, classical NLP methods (e.g., TF-IDF, lexicons) and composite indices (e.g., Fog Index, decoupling measures) improved reproducibility and enabled processing at scale [5]. However, these techniques are critically limited: they fail to capture semantic context, assess features in isolation, and lack traceability from aggregate scores to textual evidence [8]. These limitations underscore the urgent need for a structured, transparent model capable of quantifying both the rhetorical strategies and evidential substance of sustainability communication. Third, transformer-based LLMs overcome traditional NLP context limitations [18], enabling the LLM-as-a-Judge paradigm—a framework in which LLMs serve as evaluators applying explicit rubrics for consistent, auditable reasoning beyond mere data extraction [19], [20].

Current ESG research primarily focuses on extraction rather than integrated assessment [8], and lacks transparent, reproducible scoring frameworks for unified linguistic evaluation. This study addresses these gaps by introducing the EQDI, a framework that operationalizes the LLM-as-a-Judge mechanism at the claim level. By translating qualitative narratives into quantitative credibility metrics, the EQDI bridges the divide between computational linguistic research and sustainability measurement.

III. Methodology

Our methodology (Fig. 1) follows a seven-stage, interdependent workflow. It integrates automated linguistic analysis and expert validation to construct a consistent, traceable composite index.

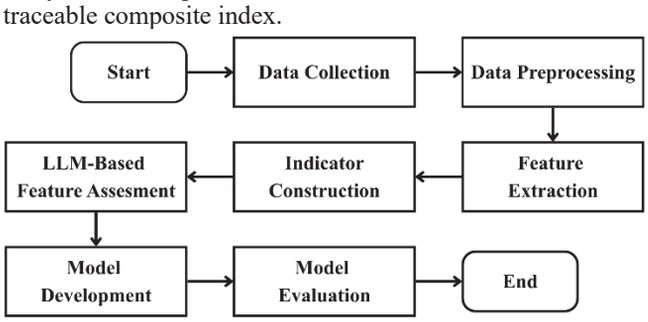


Fig. 1. Methodology workflow.

A. Data Collection and Preprocessing

The dataset comprises 2024 sustainability reports of SRI-KEHATI Index firms, sourced from official corporate and IDX repositories. Reports were converted from PDF to text using a pipeline combining PyMuPDF and Tesseract OCR. Subsequent preprocessing removed non-narrative artifacts (e.g., headers and footers) and filtered non-Indonesian text using language detection.

To isolate sustainability-related claims, we employed a few-shot prompting approach using the Claude Sonnet 4.5 model, selected for its high stability and ability to produce structured outputs within long context windows. We defined a claim as an explicit, substantive statement concerning environmental, social, or governance performance [21].

The extraction prompt instructed the model to assume the role of an ESG analyst, identifying statements relevant to company actions, policies, or results without altering the original wording [22]. We validated the model accuracy against a manually annotated subset using precision, recall, and F1-score metrics, yielding scores of 1.00 (precision), 0.88 (recall), and 0.93 (F1). This process confirmed reliable claim coverage with minimal false positives, yielding 3,068 sustainability claims, averaging 122 per company. The distribution of these claims was uneven, with environmental claims accounting for the largest category (53.18%), followed by social claims (36.99%) and governance claims (9.83%).

B. Feature Extraction

The EQDI operationalizes linguistic greenwashing cues through four core variables—Readability, Buzzword, Specificity, and Tone—collectively comprising eight measurable features. These variables measure three tactics derived from the *Sins of Greenwashing* framework [13]. First, we measure Linguistic Obfuscation (ambiguous language) using Readability and Buzzword. Second, we measure Unsubstantiated Claims (lack of verifiable data) using the Specificity metric. Finally, we measure Presentational Bias (information imbalance) using Tone.

C. Indicator Construction

We developed the Greenwashing Scorecard to operationalize eight core linguistic features into a five-point ordinal scale, ranging from 1 (very low) to 5 (very high). This 1-to-5 ordinal scale was selected to provide sufficient granularity while aligning with established precedents in LLM-as-a-Judge sustainability evaluations [19].

To ensure conceptual alignment and contextual relevance for SRI-KEHATI firms, the indicators were expert-validated. This evaluation, conducted by academic and industry specialists, produced Content Validity Index (CVI) scores of 0.80 or higher, thereby confirming the instrument's adequacy and practical relevance for ESG disclosure assessment [23].

1) Readability

Readability assesses the complexity of language that obscures information, making it easier to understand. Prior studies link it negatively to greenwashing [5]. As detailed in Table II, we operationalize this variable using two features: Information Substance and Context Dependency [24].

TABLE II. Readability Variable Scoring Rubric

Feature	Notation	Score	Criteria
Information Substance	IS	1	Boilerplate claim, no new information.
		2	Low substance, only general concepts without detail.
		3	Mentions actions, but the language is still general.
		4	Explains specific actions with explicit language.
		5	Very efficient, specific, and very clear.
Context Dependency	CD	1	Requires extensive external information for comprehension.
		2	Requires domain-specific technical knowledge.
		3	Generally understood with basic ESG literacy.
		4	Largely self-contained; assumes only common acronyms.
		5	Completely standalone, all terms are explained.

2) Buzzword

Buzzword captures rhetorical inflation (vague ESG terminology signaling commitment without evidence), which aligns with *the Sin of Vagueness* and the *Sin of Hidden Trade-off* [13]. Table III details its two key features: Semantic Accuracy (the correct application of ESG terms) and Semantic Completeness (the acknowledgment of trade-offs and limitations) [25].

TABLE III. BUZZWORD VARIABLE SCORING RUBRIC

Feature	Notation	Score	Criteria
Semantic Accuracy	SA	1	Misleading or unsubstantiated Buzzword.
		2	Used without definition; cannot be verified.
		3	Accurate definition, but with ambiguity/unreasonable extension.
		4	Accurate, linked to standards, lacks implementation detail.
		5	Very accurate, supported by external standards and a complete definition.
Semantic Completeness	SC	1	Focus on partial benefits, ignores material negative impacts.
		2	Incomplete, avoids acknowledgment of material impacts/trade-offs.
		3	Conveys the main point, lacks connection to performance or value chain scope.
		4	The main point is clear; supporting data for the partial value chain scope is provided.
		5	The main point is clear, covers trade-offs, and links to verified performance data.

3) Specificity

Specificity, conceptually linked to the *Sin of No Proof*, measures the evidential strength of ESG claims [13]. It assesses whether statements are precise and substantiated rather than merely aspirational. As detailed in Table IV, we operationalize this variable using three features: Quantitative Specificity (measurable data) [26], Temporal Specificity (timeframes) [6], and Contextual Specificity (baselines or benchmarks) [27].

TABLE IV. SPECIFICITY VARIABLE SCORING RUBRIC

Feature	Notation	Score	Criteria
Quantitative Specificity	QS	1	No numbers at all.
		2	Vague quantities ("thousands", "significant").
		3	Specific numbers, but incomplete (without unit/context).
		4	Specific data (number+unit), but partial scope.
		5	Very complete data (number, unit, scope, comparison).
Temporal Specificity	TS	1	No time frame.
		2	Ambiguous time ("long term", "immediately").
		3	Imprecise time ("during the reporting year").
		4	Specific time (clear year or target).
		5	Very detailed time (baseline + target/period).
Contextual Specificity	CS	1	No comparative context or scope.
		2	Generic/qualitative context ("lower than").
		3	Context exists, but is incomplete (missing comparative figures).
		4	Clear context (one complete reference: comparison OR scope).
		5	Very complete and multidimensional context (comparison AND scope).

4) Tone

A positive Tone is not inherently greenwashing. Accurate detection requires assessing the features of persuasive intent (linguistic polarity and self-promotional framing) alongside substantive evidence to distinguish it from credible disclosure [28]. We detail the Tone features in Table V.

TABLE V. TONE VARIABLE SCORING RUBRIC

Feature	Notation	Score	Criteria
Positive Tone	PT	1	Superlatives are dominant in the report.
		2	Strong positive Tone, supported by some factual data.
		3	Balanced positive Tone, proportional to the risks.
		4	Dominated by factual descriptions, little sentiment.
		5	Very neutral, objective, focused on data.

D. LLM-Based Feature Assessment

To ensure scalability and consistency, the feature-scoring phase was automated, using an LLM as an expert annotator. Each extracted sustainability claim was scored against the *greenwashing scorecard* using linguistic indicators. After validating several LLMs against a stratified random sample (all achieving MAE < 1.0 on a 1-to-5 scale) [29], we selected GPT-4.1 Mini for its optimal trade-off between alignment, structured output consistency, and computational efficiency [19].

We then executed the scoring through a structured prompting strategy designed for traceable, replicable, and rubric-aligned outputs. The protocol used a system prompt and a user prompt, applied an explicit 1-to-5 scoring rubric, and mandated Chain-of-Thought (CoT) reasoning for auditability with a standardized output format.

E. Model Development

The model construction process consists of four hierarchical stages.

1) Feature-Level Scoring

First, we aggregate all individual claim scores (on a 1-to-5 scale) for each of the eight features into a raw feature score per company using the arithmetic mean. This approach assumes full compensability across claims within each feature. This implies that a high score on another claim can offset a low score on a single claim within the same feature [30].

2) Variable-Level Scoring

Second, the raw feature scores are subsequently aggregated into the four main variable scores using the geometric mean. The geometric mean is employed to emphasize consistency across features. This approach is ideal for multiplicative and interdependent indicators, such as ESG linguistic dimensions, as it proportionally penalizes uneven performance [31]. This method is suitable when simultaneous success across dimensions is desired, as a low value in one dimension will degrade the overall score [30].

3) Normalization

Third, the resulting variable scores are normalized using min-max scaling to a standard range of 0–100. To maximize the index's discriminatory power, normalization was applied post-aggregation using min-max scaling of observed values [30]. This empirical approach was necessary because theoretical scaling produced a score range that was too narrow for effective differentiation; we deemed pre-aggregation normalization unnecessary, as all features already shared a homogeneous 1-to-5 rubric scale [32].

4) Index-Level Synthesis

In the final stage, the four normalized variable scores are aggregated into a single EQDI score using a weighted arithmetic mean (1), where $V_{norm,i}$ is the normalized score for each variable and W_i is its corresponding weight.

$$EQDI = \sum_{i=1}^{4}(W_i \times V_{norm,i}) \quad (1)$$

We prioritized the primary expert judgment weighting scheme to ensure construct validity by reflecting the conceptual severity of greenwashing rather than mere statistical variance [33]. This approach relegated two alternatives—equal weighting (which wrongly assumes identical importance) and Entropy (which can paradoxically underweight a critical variable)—to sensitivity analysis [34].

To derive the primary expert-judgment weighting scheme, we tasked an expert panel (academics, practitioners, investors) with distributing 100 points across the four main variables based on their perceived conceptual severity. We normalized the average allocation for each variable to serve as the W_i in the index synthesis ($\sum W_i = 1$).

Following aggregation, EQDI score categories were derived using the Jenks Natural Breaks optimization method. This method finds optimal, non-arbitrary data groupings by maximizing intra-class homogeneity and inter-class heterogeneity [35], [36]. This stratification method, well established in spatial analysis and environmental risk modeling, ensures that any cutpoints used to classify EQDI scores are consistent with the underlying distribution [37].

F. Model Stability Evaluation

To ensure stability and robustness, we validated the expert judgment model against two standard alternative weighting schemes. The first alternative, equal weighting, serves as a baseline, assuming that all variables are equally important. The second alternative, entropy weighting, is an objective, data-driven method that assigns higher weights to variables with greater statistical dispersion (variance), assuming these variables contain more information [25]. We then conducted this sensitivity analysis using two checks: 1) Spearman's rank correlation to assess robustness (high stability equals high correlation) and 2) distribution analysis to assess the discriminatory power (sensitivity) of each weighting scheme.

IV. Results and Discussion

This section presents the empirical results of the EQDI. We begin with descriptive statistics, followed by intervariable correlations, expert-defined weighting, final index scores, and evaluation of model stability.

A. Descriptive Statistics and Variable Performance

Table VI presents the detailed descriptive statistics for the eight measured features. The primary finding reveals a significant performance disparity: features measuring substantive evidence, such as Quantitative Specificity (mean = 2.11) and Temporal Specificity (mean = 2.31), consistently received the lowest scores.

When aggregated and normalized (0–100), these findings yielded similarly disparate variable scores. The Readability variable (mean = 45.83). The Specificity variable (mean = 43.24) ranked as the lowest, while rhetorical variables such as Buzzword (mean = 54.64) and Tone (mean = 52.70) scored relatively high. This indicates a systemic lack of quantitative evidence and verifiable timeframes.

TABLE VI. Descriptive Statistics of Features

Features	Mean	Std	Min	Max
IS	3.47	0.24	3.03	3.94
CD	3.37	0.16	3.16	3.88
SA	3.48	0.25	2.90	3.88
SC	2.96	0.21	2.56	3.39
QS	2.11	0.45	1.38	3.31
TS	2.31	0.37	1.62	2.92
CS	2.31	0.37	1.62	2.94
PT	3.51	0.18	3.17	3.83

The intervariable correlation heatmap (Fig. 2) confirms the nonredundancy of the four EQDI dimensions, as all exhibit low-to-moderate relationships. The most critical finding is the significant negative correlation between Tone and Specificity (–0.30) and between Tone and Buzzword (–0.61). This finding provides empirical evidence of a presentational bias: firms rely on highly positive, persuasive Tone rather than verifiable evidence and accurate terminology. This pattern reinforces the diagnostic role of these variables. [28].

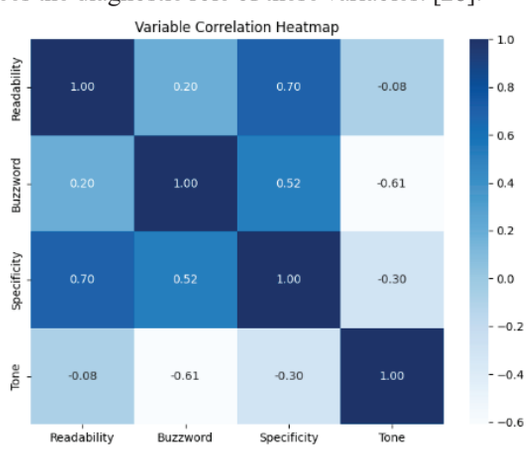


Fig. 2. Variable correlation heatmap.

B. EQDI Configuration and Score Distribution

We detail the final configuration in Equation (2). Specificity received the highest weight because evidential detail is the primary determinant of credible disclosure. Buzzword, Tone, and Readability follow with progressively lower diagnostic influence.

$$EQDI = (0.15 \times V_{norm,R}) + (0.325 \times V_{norm,B}) + (0.3625 \times V_{norm,S}) + (0.16 \times V_{norm,T}) \quad (2)$$

The final EQDI scores demonstrated significant variance (Mean = 48.87, Std = 16.23, Range = 29.01–83.75), confirming the model's discriminatory power. Jenks Natural Breaks stratification yielded three tiers: Low Quality (<42.44), Moderate Quality (42.44–60.13), and High Quality (>60.13). Based on this classification, 10 firms fall into the low tier, 9 firms into the moderate tier, and 6 firms into the high tier. The low tier reflects disclosures dominated by broad claims with limited supporting evidence; the moderate tier represents mixed disclosures with partial substantiation; and the high tier indicates disclosures characterized by specific, verifiable, and contextually detailed sustainability claims.

C. Expert Validation of Variable Relevance

Readability and Specificity variables demonstrated the highest consensus with respective CVI scores of 1.00 and 0.95. These scores indicate that Readability and Specificity are considered essential for differentiating substantive from merely declarative sustainability claims. Achieving a CVI of 0.90, the Buzzword variable also showed strong validity. The Tone variable with a CVI of 0.80 was also deemed valid, although its consensus was lower than that of the other variables. Overall, all four variables were validated and retained as core dimensions for the EQDI construct.

Furthermore, the subfeature level assessment yielded CVI scores within the 0.85–0.95 range, indicating that all operational features consistently represent their corresponding variables. No subfeatures required elimination or revision, thereby establishing the instrument's structure as stable and viable for use. Thus, the expert validation confirms that the variables and sub-features utilized in the EQDI possess adequate conceptual grounding.

D. Model Stability Evaluation

We performed a stability evaluation to assess the robustness of the EQDI. The analysis focused on rank-order consistency (ρ), mean score shifts (ΔMean), and score distribution (ΔSD).

TABLE VII. MODEL STABILITY EVALUATION

Weighted Combination	*Spearman (ρ)*	*ΔMean*	*ΔSD*
Expert vs. Equal	0.93	-0.23	+2.10
Expert vs. Entropy	0.91	+0.72	+0.91
Equal vs. Entropy	0.99	+0.95	-1.19

The results confirm that the index is highly robust, as all inter-scheme correlations (ρ) exceed 0.90 (Table VII). While the mean scores also shifted slightly between schemes (as seen in the ΔMean column), the most methodologically critical difference was in score dispersion. The expert scheme yielded the greatest variance (ΔSD = +2.10 vs. equal), indicating stronger discriminatory capacity. Consequently, the expert weighting scheme was adopted as the primary configuration, as it provides the optimal balance between rank stability and the critical differentiation sensitivity required to identify greenwashing indicators.

E. Discussion

Building on the empirical findings, we interpret the significant negative correlation between Tone and Specificity ($\rho = -0.30$) as validating the signal substitution mechanism proposed by signaling theory [11]. This pattern reveals how firms sustain symbolic legitimacy (high Tone) while decoupling it from substantive performance (low Specificity) [12].

These findings reinforce the EQDI's theoretical foundation. Specificity, the variable prioritized by experts (36.25% weight), was precisely the dimension in which SRI-KEHATI firms performed worst (mean = 43.24). This finding identifies Specificity as the primary diagnostic indicator of the *Sin of No Proof* [13]. It also supports legitimacy theory, as firms appear to prioritize impression management over accountability, even within a sample recognized for ESG commitment.

Beyond validating the signal-substitution mechanism, this pattern is consistent with prior evidence on symbolic ESG disclosure practices, where narrative optimism coexists with weak quantitative support [7], [24]. By capturing this divergence at the claim level, the EQDI extends earlier document-level readability and sentiment approaches [5], showing that high-level positivity can mask systematic under-reporting of verifiable performance. The strong expert emphasis on Specificity, combined with its low empirical scores, also suggests that indices relying on equal weighting or readability-only proxies may underestimate the risk of greenwashing in "best-in-class" samples, such as SRI-KEHATI constituents.

V. CONCLUSION

This research successfully developed the EQDI formulation and empirically confirmed that it provides the necessary semantic structure to transform a generic LLM into a specialized audit instrument. The system's high precision in claim extraction (F1-score: 0.93) and its alignment with experts' scoring (MAE < 1.0; CVI > 0.80) validate that a structured linguistic approach is essential for achieving accuracy in Vertical AI applications. Consequently, this demonstrates that the EQDI can systematically detect signal-substitution tactics in which a high tone masks low specificity. This framework demonstrates the viability of Vertical AI in processing unstructured corporate data, offering a blueprint for scalable Green IT decision-support systems.

Practically, the EQDI alerts investors that high sentiment may signal low substance and provides regulators (e.g., OJK) with a scalable tool to differentiate compliance from quality, as well as to assess TKBI and Nationally Determined Contribution (NDC) alignment.

Future research should address two key limitations of this framework. First, the EQDI measures disclosure quality, not real-world performance (the gap between disclosed quality and actual performance). Second, it cannot detect strategic omissions of material topics. Subsequent work should therefore bridge this gap by correlating EQDI scores with objective performance metrics (e.g., verified emissions) and integrating sector-specific materiality standards (SASB/ESRS) to penalize omissions.

REFERENCES

[1] L. Johnson, "Greenwashing cases fall for first time in 6 years, but high-severity filings surge: report," esgdive, 2025. [Online]. Available: https://www.esgdive.com/news/greenwashing-decreases-report-rep-risk-severity-increases-2024/729511/

[2] S. Zoting, "ESG Investing Market Expands with Rising Demand for Sustainable Financial Strategies," precedenceresearch, 2025. [Online]. Available: https://www.precedenceresearch.com/esg-investing-market

[3] G. Marques Alves Aguiar, "The Evolution and Integration of ESG Reporting: A Framework for Sustainable Business Accountability," 2025. doi: 10.2139/ssrn.5350387.

[4] M. Brychko, Y. Bilan, S. Lyeonov, and G. Mentel, "Trust crisis in the financial sector and macroeconomic stability: a structural equation modelling approach," *Economic Research-Ekonomska Istraživanja*, vol. 34, no. 1, pp. 828–855, Jan. 2021, doi: 10.1080/1331677X.2020.1804970.

[5] P. Hu, X. Li, N. Li, Y. Wang, and D. D. Wang, "Peeking into Corporate Greenwashing through the Readability of ESG Disclosures," *Sustainability (Switzerland)*, vol. 16, no. 6, Mar. 2024, doi: 10.3390/su16062571.

[6] H. Horn, V. Wieman, and A. Ohl, "Navigating the ESG landscape: Comparison of the 'Big Three' Disclosure Proposals," 2022. [Online]. Available: https://corpgov.law.harvard.edu/2022/10/10/navigating-the-esg-landscape-comparison-of-the-big-three-disclosure-proposals/

[7] T. Schimanski, A. Reding, N. Reding, J. Bingler, M. Kraus, and M. Leippold, "Bridging the gap in ESG measurement: Using NLP to quantify environmental, social, and governance communication," *Financ Res Lett*, vol. 61, p. 104979, Mar. 2024, doi: 10.1016/j.frl.2024.104979.

[8] Y. Zou *et al.*, "ESGReveal: An LLM-based approach for extracting structured data from ESG reports," *J Clean Prod*, vol. 489, Jan. 2025, doi: 10.1016/j.jclepro.2024.144572.

[9] A. Alamsyah, A. A. Hafidh, and A. D. Mulya, "Innovative Credit Risk Assessment: Leveraging Social Media Data for Inclusive Credit Scoring in Indonesia's Fintech Sector," *Journal of Risk and Financial Management*, vol. 18, no. 2, p. 74, Feb. 2025, doi: 10.3390/jrfm18020074.

[10] S. Agrawal, A. Ray, and N. P. Rana, "Effectiveness of sustainability communication on stakeholder engagement: A multi-method qualitative assessment of sustainable development," *Corp Soc Responsib Environ Manag*, vol. 32, no. 1, pp. 820–834, Jan. 2025, doi: 10.1002/csr.2992.

[11] V.-S. Osburg, V. Yoganathan, F. McLeay, and M. F. Diallo, "(In)compatibilities in sustainable luxury signals," *Ecological Economics*, vol. 196, p. 107430, Jun. 2022, doi: 10.1016/j.ecolecon.2022.107430.

[12] P. Velte, "Determinants and financial consequences of environmental performance and reporting: A literature review of European archival research," *J Environ Manage*, vol. 340, p. 117916, Aug. 2023, doi: 10.1016/j.jenvman.2023.117916.

[13] UL Solutions, "Sins of Greenwashing." 2007. [Online]. Available: https://www.ul.com/insights/sins-greenwashing

[14] Á. Lublóy, J. L. Keresztúri, and E. Berlinger, "Quantifying firm-level greenwashing: A systematic literature review," Jan. 01, 2025. doi: 10.1016/j.jenvman.2024.123399.

[15] A. A. Hafidh and A. Alamsyah, "A Credit Scoring Model Proposal Based on Social Media Data to Enhance Financial Assessment," in *2024 12th International Conference on Information and Communication Technology (ICoICT)*, IEEE, Aug. 2024, pp. 15–21. doi: 10.1109/ICoICT61617.2024.10723094.

[16] Y. Sagama, A. Alamsyah, and D. Tricahyono, "Generating Mobile App Improvement Recommendations Through User Review Analysis Using BERTopic and GPT," in *2024 International Conference on Digital Business and Technology Management (ICONDBTM)*, IEEE, Oct. 2024, pp. 1–6. doi: 10.1109/ICONDBTM64135.2024.11122739.

[17] Y. Lin, R. Shen, J. Wang, and Y. Julia Yu, "Global Evolution of Environmental and Social Disclosure in Annual Reports," *Journal of Accounting Research*, vol. 62, no. 5, pp. 1941–1988, Dec. 2024, doi: 10.1111/1475-679X.12575.

[18] L. Xia, M. Yang, and Q. Liu, "Using Pre-trained Language Model for Accurate ESG Prediction," Nov. 2024, pp. 1–22. [Online]. Available: https://aclanthology.org/2024.finnlp-2.1/

[19] M. Chuang, G. Chuang, C. Chuang, and J. Chuang, "Judging It, Washing It: Scoring and Greenwashing Corporate Climate Disclosures using Large Language Models," 2025. doi: 10.18653/v1/2025.climatenlp-1.2.

[20] M. Amirizaniani *et al.*, "LLMAuditor: A Framework for Auditing Large Language Models Using Human-in-the-Loop," May 2024, [Online]. Available: http://arxiv.org/abs/2402.09346

[21] ESMA, "Thematic notes on clear, fair & not misleading sustainability-related claims," Jul. 2025. [Online]. Available: https://share.google/EgO95r5p1Tl36OfOL

[22] S. Sivarajkumar, M. Kelley, A. Samolyk-Mazzanti, S. Visweswaran, and Y. Wang, "An Empirical Evaluation of Prompting Strategies for Large Language Models in Zero-Shot Clinical Natural Language Processing: Algorithm Development and Validation Study," *JMIR Med Inform*, vol. 12, 2024, doi: 10.2196/55318.

[23] N. A. Mat Said, S. M. Bujang, N. A. Buang, H. H. Siraj Ramli, and M. N. Awang Besar, "Critical Thinking Transfer Practice Instrument: A Content Validity Calculation Steps Based on Expert Panel Evaluation," *Education in Medicine Journal*, vol. 14, no. 3, pp. 61–74, 2022, doi: 10.21315/eimj2022.14.3.5.

[24] D. Crilly, M. Hansen, and M. Zollo, "The Grammar of Decoupling: A Cognitive-Linguistic Perspective on Firms' Sustainability Claims and Stakeholders' Interpretation," *Academy of Management Journal*, vol. 59, no. 2, pp. 705–729, Apr. 2016, doi: 10.5465/amj.2015.0171.

[25] ESMA, "ESMA TRV Risk Analysis ESG names and claims in the EU fund industry," 2023, doi: 10.2856/885324.

[26] M. Xu, Y. K. Tse, R. Geng, Z. Liu, and A. Potter, "Greenwashing and market value of firms: An empirical study," *Int J Prod Econ*, vol. 284, Jun. 2025, doi: 10.1016/j.ijpe.2025.109606.

[27] M. Liu, "Quantitative ESG disclosure and divergence of ESG ratings," *Front Psychol*, vol. 13, Aug. 2022, doi: 10.3389/fpsyg.2022.936798.

[28] X. Yang, J. Bao, and K. Zhang, "Environmental, social and governance (ESG) performance and abnormal positive tone," *Sustainability Accounting, Management and Policy Journal*, vol. 16, no. 2, pp. 351–388, Jan. 2025, doi: 10.1108/SAMPJ-01-2024-0045.

[29] T. O. Hodson, "Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not," Jul. 2022, doi: 10.5194/gmd-15-5481-2022.

[30] OECD, "Handbook on Constructing Composite Indicators," 2008. [Online]. Available: www.oecd.org/publishing/corrigenda.

[31] M. Biggeri, "SPES Focus on WP 3 Beyond the numbers: what composite indices reveal-and conceal-about trade-offs and synergies in sustainability transitions Contributors and peer reviewers," 2025. [Online]. Available: https://www.sustainabilityperformances.eu/publications-deliverables

[32] M. Mazziotta and A. Pareto, "Methods for constructing composite indices: one for all or all for one? 1," *Riv Ital Econ Demogr Stat*, vol. 67, no. 2, pp. 67–80, 2013, [Online]. Available: https://www.researchgate.net/publication/281106596

[33] F. Bernini, M. Giuliani, and F. La Rosa, "Measuring greenwashing: A systematic methodological literature review," *Business Ethics, the Environment and Responsibility*, vol. 33, no. 4, pp. 649–667, Oct. 2024, doi: 10.1111/beer.12631.

[34] V. Lagasio, "Measuring Greenwashing: The Greenwashing Severity Index," 2023. doi: 10.2139/ssrn.4582917.

[35] L. Bornmann, L. Leydesdorff, and R. Mutz, "The use of percentiles and percentile rank classes in the analysis of bibliometric data: Opportunities and limits," 2013. doi: 10.1016/j.joi.2012.10.001.

[36] C. Reimann, P. Filzmoser, and R. G. Garrett, "Background and threshold: critical comparison of methods of determination," *Science of the Total Environment*, vol. 346, no. 1–3, pp. 1–16, Jun. 2005, doi: 10.1016/j.scitotenv.2004.11.023.

[37] New Jersey Department of Environmental Protection, "Remedial Priority System: Assigning Categories to the RPS Score," 2012. [Online]. Available: https://nj.gov/dep/srp/srra/rps/docs/rps_jenks_breaks.pdf

Trends in Online Monitoring of OT- and IIoT-Networks

Axel Sikora[1]

***Abstract*—Secure and safe operation of data networks needs to be supervised. Supervision, monitoring and intrusion detection systems (IDS) are in place and are object of ongoing research and optimization. However, the traditional rule-based IDS might be falling short, as they are insufficient against dynamic threats. This paper discusses actual trends and developments for online monitoring of communication networks, concentrating on Operational Technology (OT) Industrial Internet of Things (IIoT)-networks, showing the complete pipeline from data set generation, training mechanisms, and applications. In doing so, the paper gives an insight into selected relevant projects from the author's teams. In addition, the extensive list of references shall serve interested readers as an entry point for subsequent scientific and implementation works. This paper summarizes a keynote presentation from the 14th International Conference on Green and Human Information Technology (ICGHIT) in February 2026 held in Denpasar, Bali, Indonesia.**

***Index Terms*—Industrial IoT, OT networks, security, authentication, encryption, intrusion detection system, IDS, intrusion prevention system, IPS, Security Information and Event Management, SIEM, Security Orchestration, Automation and Response, SOAR.**

I. Introduction

The journey of applications and technologies for the Industrial Internet of Things (Industrial IoT, IIoT) is rapidly and continuously advancing. At the same time, the integration of Operational Technology (OT) networks into Information Technology (IT) infrastructure is becoming more intense. Yet, their increasing digitization, dependability, and degree of interconnections expose them to sophisticated cyber-threats capable of bypassing traditional security measures and significantly enlarges the attack surface [**?**].

With the rapid proliferation of Internet of Things (IoT) devices, the need for efficient and effective Intrusion Detection System (IDS) tailored for IoT environments has become increasingly paramount. [1] explores various techniques employed in contemporary IoT IDS, including traditional signature-based approaches like Snort and Bro/Zeek, as well as emerging deep learning-based methods.

Also, the regulatory framework around the EU Cyber Resilience Act (CRA), Radio Equipment Directive (RED) or the second EU Directive on Network and Information Systems (NIS2) responds to this situation and makes the secure operation a MUST for most of the IIoT and OT systems.

This overview paper highlights several aspects (cf. Fig. 1): It

- presents possible attack models, most notably the MITRE attack model (cf. Section II),
- gives a short overview on several core projects at the author's institute (cf. Sections III-B and IV-B), and
- provides an overview on selected in-depth publications of the author's team, which have been published around these projects. Reference to further sources can be found in these publications.

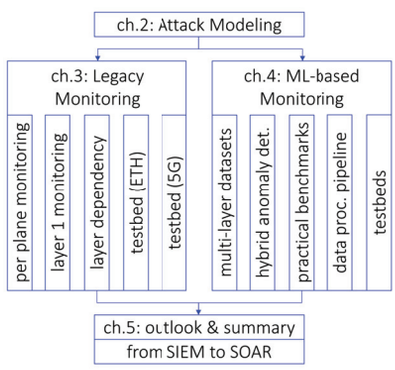


Fig. 1. Structure of this Overview Paper

II. Attack Modelling

An attack model is a structured representation of how a cyber adversary could carry out an attack against a target system or network. An attack model in cybersecurity lets organizations anticipate attacks by visualizing an attacker's behavior and pathways, so they can strengthen defenses proactively. It maps out the steps, tactics and techniques an attacker might use, from initial reconnaissance to achieving their end goals, providing defenders with a blueprint of potential attack paths and methods. There are two well-known frameworks for the modelling of attacks.

A. Cyber Kill Chain® Framework

Developed by Lockheed Martin, the Cyber Kill Chain® framework[2] is part of the Intelligence Driven Defense® model for identification and prevention of cyber intrusions activity. The model identifies what the adversaries must complete in order to achieve their objective. The seven steps of the Cyber Kill Chain® enhance visibility into an attack and enrich an

[1]Institute of Reliable Embedded Systems and Communication Electronics, Offenburg University of Applied Sciences, Offenburg, Germany. Email: axel.sikora@hs-offenburg.de

978-1-5386-5541-2/18/$31.00 © 2018 IEEE

analyst's understanding of an adversary's tactics, techniques and procedures. Further review can be found at [2] among others.

B. MITRE ATT&CK Framework

The MITRE ATT&CK framework[3] is a globally-accessible knowledge base of adversary tactics and techniques based on real-world observations. It has been developed by MITRE, the MIT Research Establishment, an American not-for-profit organization managing federally funded research and development centers and supporting various U.S. government agencies.

The ATT&CK framework has been originally designated for IT enterprise, offers a rich taxonomy of adversarial tactics and techniques derived from real-world incidents. While its applicability to IT environments is well established, its use in industrial OT networks remains underexplored.

However, in [3] we exposed this gap and demonstrated the overall theoretical concept and the concrete mapping of the ATT&CK framework to a legacy fieldbus, like Modbus. This work was performed in conjunction with KISTE project (cf. ch. IV-B).

III. Online Supervision of TSN Networks

A. Background - Time Sensitive Networking

During the last decades, real-time communication was limited to small L2-networks, often following specific standards, be it traditional fieldbus protocols, like PROFIBUS, Modbus, CAN or alike, be it real-time Ethernet extensions, like PROFINET, EtherCAT, OPC UA FX, Ethernet/IP or alike. Such networks were connected to IT-networks only via gateways.

Since several years, the integration of Time-Sensitive Networking (TSN) is going on. TSN promises interoperable and convergent networking over Ethernet networks by enhancing the standard itself. Furthermore, even some of the previously mentioned real-time Ethernet protocols (PROFINET, OPC UA FX) are now starting to use TSN for the lower layers while keeping their existing application layer protocols in place

B. Project Overview - Secure TSN

The project "SecureTSN – Intelligent security for Real-Time and Interoperable Communication Networks" is funded by German Federal Ministry for Economic Affairs and Energy (funding ID IGF01IF23657N) and is executed from July 2025 till June 2027 by the author's institute together with IMMS Institut für Mikroelektronik- und Mechatronik-Systeme[4], accompagnied by an industrial advisory board.

The objective of the project is to provide methods and tools for securing TSN against attacks from external networks, but also from components within the internal network. The project results help to secure TSN networks and increase their reliability in order to increase acceptance among users. Measures are being defined and developed at several levels, which can be incorporated into new products in the form of hardware functions and, as software components at higher levels, take over both the control and management of the new security functions.

C. Monitoring

In order to properly configure, run and secure such TSN, monitoring functionality is a must. The TSN standard already has some preparations to provision such functionality and there are different methods to choose from. [4] implemented different methods to measure the time synchronisation accuracy between devices as a C library and compared the measurement results. Furthermore, the library has been integrated into an engineering framework.

In addition, [5] examines the implementation and performance of time synchronization monitoring methods for generalized Precision Time Protocol (gPTP), including Monitoring Type Length Value (TLV) for Ingress & Egress messages, Reverse Sync, and Pulse per Second (PPS) parameters across varied hardware environments and operational conditions.

In this paper, the integration process is detailed, the adaptability of these methods under stress tests is discussed, and their effectiveness through rigorous assessments is evaluated. The findings contribute to refining monitoring deployment strategies by identifying the most effective combinations of monitoring techniques to enhance synchronization accuracy and network reliability.

D. Attack Modelling

[6] provides a comprehensive review of TSN security for industrial automation, with an emphasis on a systematic threat analysis and mitigation strategy across all facets of a TSN-based system. The survey is intended to provide a structured snapshot of currently known TSN attack vectors, organized according to three primary "planes" of TSN operation: the time synchronization plane, the control plane, and the data plane.

In total, 15 distinct attack scenarios are enumerated in the time synchronization plane, 15 in the control plane, and 14 in the data plane, which leads to 44 in total. They range from spoofing of grandmaster clocks to denial-of-service (DoS) via malicious scheduling and covert timing channels in data traffic.

IV. AI Assisted Network Monitoring

A. Overview

Intrusion Detection Systems (IDS) are evolving from static, signature-based approaches to more adaptive machine learning (ML)-driven solutions [7]. Classical rule-based systems, though effective for known threats, lack the ability to detect novel or evolving attack vectors [8]. Machine learning introduced much-needed flexibility, with supervised methods (e.g., Support Vector Machines (SVMs), decision trees) distinguishing anomalous patterns and unsupervised models (e.g., autoencoders) revealing new, unlabeled threats [9].However,

[2]https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html

[3]https://attack.mitre.org/

[4]https://www.imms.de/en

this flexibility comes at a cost: the shift from deterministic rules to stochastic models introduces a new attack surface, where adversaries can exploit the model's learning boundaries through evasion and poisoning attacks.

Existing intrusion detection systems (IDS) and intrusion prevention systems (IPS) mostly concentrate on legacy IT networks. Nowadays, they come with a high degree of complexity and adaptivity, including the use of artificial intelligence. It is only recently that these techniques are also applied to IoT networks. A critical challenge here is the *reliability gap*: applying probabilistic AI models to safety-critical, deterministic operational environments requires robustness guarantees that standard IT-centric ML often fails to provide. In [10], a survey of machine learning and deep learning methods for intrusion detection is presented. Also, it is investigated how previous works used federated learning for IoT cybersecurity. For this, an overview of IoT protocols and potential security risks is presented, the techniques and the datasets used in the studied works are reported, the challenges of using Machine Learning (ML), Deep Learning (DL) and Federated Learning (FL) for IoT cybersecurity are discussed, and future insights are provided.

[11] bridges the critical divide between general Adversarial Machine Learning and the strict operational constraints of Industrial Control Networks (ICN). Guided by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA2020) principles [12], the study synthesizes the state-of-the-art into a taxonomy of eight distinct defensive paradigms—ranging from Generative Adversarial Network (GAN)-based augmentation to cyber-physical fusion. Crucially, the authors reframe these techniques not as competing alternatives, but as a composable defense stack required to secure heterogeneous Operational Technology (OT) environments. The survey concludes by establishing a roadmap for next-generation defenses, specifically prioritizing physics-informed adversarial learning and standardized benchmarking to ensure reliability in safety-critical infrastructure.

B. Project Overview - KISTE

The project "KISTE - AI-Supported SIEM System for Highly Reliable Industrial IoT networks and Field Busses"[5] is also funded by German Federal Ministry for Economic Affairs and Energy (funding ID ZIM KK5189606RG4). It is executed from July 2024 till June 2026 by the author's institute together with Decoit GmbH[6].

The objective of the KISTE project is to develop a Security Information and Event Management (SIEM) system that continuously and non-invasively monitors, records, and evaluates anomalies and incidents (events, security incidents) in industrial networks. The data to be analyzed is collected by entities distributed across the company networks, known as data sensors. These are used not only in classic field buses as well as in Ethernet-based field buses.

[5]https://kiste-project.info/
[6]https://decoit.de/en/

The subsequent classification and evaluation of the collected data should not only be understandable to the human IT administrator, so that he can react immediately when a security incident occurs, but should also be as autonomous as possible. Because there are currently no viable SIEM systems available for both IoT and OT environments, the KISTE project aims to remedy this situation. A system is to be developed that meets the specific requirements of these IoT and OT networks.

The KISTE system is aimed both at companies that operate critical infrastructures and, above all, at conventional IoT and OT companies that fall within the indirect scope of the planned KRITIS umbrella law and the IT Security Act 3.0 (SiG 3.0) and for which no practical security solutions and SIEM systems are yet available.

C. Training Data

The effectiveness of ML techniques strongly depends on the size and the quality of the training datasets. A major challenge that researchers and practitioners face is the limited availability of publicly accessible datasets targeted for attacks on OT and IoT networks. This limited availability currently restricts the ability to build, train, and evaluate the performance of ML models [13] [14].

To our best knowledge, [15] presents the first systematic review of current publicly available datasets to be used to train and evaluate ML-based IDS in IoT and OT network environments. It examines notable datasets, highlights their structural characteristics, feature sets, attack types covered, and discusses their applicability to real-world IoT and OT scenarios. The paper aims to not only provide guidelines to choose appropriate datasets by highlighting their limitations, weaknesses, and strengths, but also focuses on developing more representative and realistic datasets.

D. Testbed

To our knowledge, [16] describes the first generic cyber attack testbed for deep field buses, enabling systematic attacks and analyses. The collected data, which encompasses network traffic, telemetry, and host logs, is processed through a centralized Elastic Security Information and Event Management system augmented by local monitoring tools such as NetFlow, Auditd, and Zeek.

The paper details the design, implementation, and evaluation of our testbed, demonstrating its adaptability to emulate various attack scenarios and its effectiveness in producing datasets that advance cybersecurity in Industrial Control networks.

E. Framework

[17] proposes a hybrid anomaly detection framework that integrates rule-based SIEM techniques with machine learning–driven anomaly detection. A centralized SIEM aggregates logs from IT and OT components and applies predefined detection rules, mapped to the MITRE ATT&CK framework, to rapidly identify known threats. Concurrently, lightweight unsupervised models, specifically autoencoders, run at the edge, learning normal behavior from IIoT traffic and telemetry,

and flagging deviations in real time. This dual approach leverages the precision of signature-based detection and the adaptability of behavioral analytics, delivering comprehensive, low-latency threat identification tailored for distributed, resource-constrained IIoT environments.

The key contributions of the paper include:

1) Multi-Layer Dataset Generation: Host-based logs (process metrics, file activity, authentication events, and security alerts) are consolidating input captured by auditd, with protocol data captured by Zeek and network flows captured by NetFlow, addressing fragmented telemetry for improved anomaly detection.
2) Hybrid Anomaly Detection: A hybrid approach is proposed combining MITRE ATT&CK-mapped rule-based SIEM detection and a machine learning autoencoder module deployed at the edge, ensuring both rapid identification of known threats and adaptive detection of novel anomalies.
3) Practical Deployment and Benchmarking: The framework supports real-world industrial settings, integrating data from OT and IT networks for centralized analysis and edge-based inference.
4) Scalable Data Processing Pipeline: A pipeline is developed transforming raw network, host, and protocol data into actionable features, enabling efficient ingestion, enrichment, and indexing within the SIEM environment.

V. Conclusion and Future Work

In this paper, we presented a selective overview of research papers on legacy and on AI-based online monitoring. The paper discusses actual trends and developments for online monitoring of communication networks, concentrating on OT- and IIoT-networks, showing the complete pipeline from data set generation, training mechanisms, and applications.

Further contributions from the author's team will concentrate on

- implementing the full TSN-monitoring framework (cf. ch.III)
- enhancing the testbed with regards to functionality and with regards to further fieldbus- and IIoT-protocols (cf. ch.IV-D),
- automating attack generation on fieldbus and IIoT-networks,
- extending the SIEM system IV-E into a SOAR (Security Orchestration, Automation, and Response) system and going further in terms of functionality by taking over the alerts generated by SIEM, prioritizing them, and responding to them quickly and automatically.,

VI. Acknowledgement

The author is grateful for all the great contributions from his team and for all the intense discussions, fruitful ideas, and objective-oriented efforts.

References

[1] A. T. Zahary, N. Ali Al-shaibany, and A. Sikora, "A Review of Intrusion Detection Systems for the Internet of Things," in 1^{st} *Int'l Conf on Emerging Technologies for Dependable Internet of Things (ICETI)*, 2024, pp. 1–6.

[2] L. M. Fadzil, S. Manickam, and M. A. Al-Shareeda, "A Review of An Emerging Cyber Kill Chain Threat Model," in 2^{nd} *Int'l Conf on Advanced Computer Applications (ACA)*, 2023, pp. 157–161.

[3] J. Rahmani and A. Sikora, "Mapping the MITRE ATT&CK Framework to Modbus Cyber Attacks in Industrial OT Networks," in 60^{th} *IEEE Int'l Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST)*, 2025, pp. 1–4.

[4] K. Naik, D. Welte, S. Oechsle, F. Frick, A. Lechler, M. Schappacher, and A. Sikora, "Methodology and Implementation for Monitoring Precise Time Synchronisation in TSN," in *IEEE* 21^{st} *Int'l Conf on Industrial Informatics (INDIN)*, 2023, pp. 1–6.

[5] K. Naik, M. Schappacher, D. Welte, and A. Sikora, "Monitoring Time Synchronization Precision: Implementation, Validation, and Low Cost PPS Measurement Mechanism for gPTP Monitoring," in *IEEE* 29^{th} *Int'l Conf on Emerging Technologies and Factory Automation (ETFA)*, 2024, pp. 1–8.

[6] K. Naik and A. Sikora, "Literature Survey on Securing Time-Sensitive Networks (TSN) in Industrial Automation," in *prepared for journal submission*, 2026.

[7] G. Zizzo, C. Hankin, S. Maffeis, and K. Jones, "INVITED: Adversarial Machine Learning Beyond the Image Domain," in 56^{th} *ACM/IEEE Design Automation Conference (DAC)*, 2019, pp. 1–4.

[8] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," 2015. [Online]. Available: https://arxiv.org/abs/1412.6572

[9] M. A. Umer, K. N. Junejo, M. T. Jilani, and A. P. Mathur, "Machine Learning for Intrusion Detection in Industrial Control Systems: Applications, Challenges, and Recommendations," *International Journal of Critical Infrastructure Protection*, vol. 38, p. 100516, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1874548222000087

[10] M. A. Messaad, C. Jerad, and A. Sikora, "AI Approaches for IoT Security Analysis," in *Intelligent Systems, Technologies and Applications*, M. Paprzycki, S. M. Thampi, S. Mitra, L. Trajkovic, and E.-S. M. El-Alfy, Eds. Singapore: Springer Singapore, 2021, pp. 47–70.

[11] J. Rahmani and A. Sikora, "A Survey on Adversarial Machine Learning for Robust Anomaly Detection in Industrial Control Networks," in *prepared for journal submission*, 2026.

[12] M. Page *et al.*, "The PRISMA 2020 Statement: an Updated Guideline for Reporting Systematic Reviews," *BMJ*, vol. 372, 2021. [Online]. Available: https://www.bmj.com/content/372/bmj.n71

[13] M. Samantaray, R. C. Barik, and A. K. Biswal, "A Comparative Assessment of Machine Learning Algorithms in the IoT-Based Network Intrusion Detection Systems," *Decision Analytics Journal*, vol. 11, p. 100478, 2024.

[14] Z. Chen, J. Liu, Y. Shen, M. Simsek, B. Kantarci, H. T. Mouftah, and P. Djukic, "Machine Learning-Enabled IoT Security: Open Issues and Challenges Under Advanced Persistent Threats," *arXiv preprint arXiv:2204.03433*, 2022.

[15] A. Muhammad, J. Rahmani, and A. Sikora, "A Comprehensive Review of Public Datasets for Machine Learning-Based Intrusion Detection in IoT and OT Networks," in *13th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, 2025.

[16] J. Rahmani, K. O. Detken, and A. Sikora, "A Testbed for Cyber Attack Emulation and AI-Driven Anomaly Detection in Industrial IoT- and OT-Networks," in 13^{th} *IEEE Int'l Conf on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, 2025.

[17] J. Rahmani, S. D. Çakmakçı, K. O. Detken, and A. Sikora, "A SIEM-Based Framework for Multi-Layer Data Collection and Anomaly Detection in OT-Networks," in *13th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, 2025.

Adaptive Weighted Fairness Power Allocation for Vehicular NOMA

Kagudde Abbas
Dept. of Information and Communication Engineering
Yeungnam University
Gyeongsan 38541, South Korea
kaguddeabbas@yu.ac.kr

Kwonhue Choi*
Dept. of Information and Communication Engineering
Yeungnam University
Gyeongsan 38541, South Korea
gonew@yu.ac.kr

***Abstract*—Non-orthogonal multiple access (NOMA) enables massive connectivity for vehicle-to-everything (V2X) communications, but balancing fairness and throughput under dynamic channels and imperfect successive interference cancellation (ipSIC) remains challenging. Building on our earlier fixed weight power allocation idea, this paper brings a more flexible, channel adaptive weight strategy, where the weight value can vary depending on channel conditions and the quality of service (QoS) requirements needed by a user. The closed form solution keeps the same complexity and intergrates well with the Karush Kuhn Tucker (KKT) based power allocation. From simulations using around 100,000 channel samples, there is 38.5% improvement in fairness and around 43.2% reduction in variance, while only losing around 2.78% throughput compared to the fixed weight approach. The adaptive method becomes even more practical when the channel imbalance is strong, which shows that it can be implemented in real vehicular NOMA systems where both reliability and fair resource sharing are very important, especially for safety critical cases.**

***Index Terms*—Vehicular NOMA, channel-adaptive power allocation, fairness optimization, V2X communication.**

I. Introduction

The rapid growth of intelligent transportation systems and autonomous vehicles has led to wireless technology to support massive numbers of devices, ultra-reliable low-latency communication, and fair resource allocation [1]. Vehicle-to-everything (V2X) communication is becoming the main part of future mobility systems, including cooperative driving, accident prevention, and traffic flow improvement [2]. Non-orthogonal multiple access (NOMA) is now viewed as one of the key technologies because it can give better spectral efficiency and allow many users to be served at the same time using power-domain multiplexing [3].

However, vehicular NOMA faces critical challenges: (1) dynamic channels in mobile environments, (2) heterogeneous link qualities between near and far vehicles, (3) safety-critical fairness requirements for collision warnings and emergency notifications, and (4) imperfect successive interference cancellation (ipSIC) degrading rates due to channel estimation errors and hardware impairments [4], [5].

Power allocation directly influences both throughput and fairness in NOMA systems. Some research focuses on sum-rate maximization [6]–[8], while others on fairness through either weighted max-min schemes [9], [10] using fixed weighting parameters that cannot adapt to rapidly varying vehicular channels. Reinforcement learning methods [11] offer adaptability but suffer from high computational complexity [12].

In our prior work [13], we developed a tunable fairness-aware power allocation framework minimizing weighted signal interference plus noise ratio (SINR) difference between users, where fixed weight ω balances fairness and throughput. We derived closed-form optimal power allocation via KKT conditions. While effective under static channels, this fixed-ω approach has a critical limitation: when channel gains $g_{n,1}$ and $g_{n,2}$ vary significantly as a result of n-th antenna, fixed ω causes either fairness violations (when $g_{n,1} \gg g_{n,2}$) or unnecessary throughput loss (when $g_{n,1} \approx g_{n,2}$).

This paper introduces a channel-adaptive weight selection strategy addressing fairness-throughput balance under time-varying channels. Our contributions are:

- Unlike the fixed ω in [13], we propose $\omega^* = \omega_{\text{base}}(1 + \alpha\xi)$ that automatically adjusts to instantaneous channel imbalance ξ increasing protection for far users when asymmetry is severe while reducing unnecessary compensation when channels are balanced, ω_{base} encodes QoS requirements, and $\alpha \in [0, 1]$ controls adaptation.
- We prove the adaptive weight's continuity, monotonicity, and boundedness, which demonstrating a smaller expected fairness gap and variance as compared to any fixed-weight design under varying channel conditions.
- Simulations over 10^5 realizations with realistic path-loss and imperfect SIC show 38.5% fairness improvement and 43.2% variance reduction with only 2.78% throughput penalty, validating practical benefits for dynamic scenarios.

This closed-form channel-adaptive fairness framework maintains theoretical optimality, $O(1)$ complexity, accounts for ipSIC, and provides practical deployment guidelines for vehicular NOMA systems.

*Corresponding author: Kwonhue Choi (gonew@yu.ac.kr)

This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) funded by Korean Government through the Ministry of Science and Information and Communication Technology (MSIT) under Grant 2022-0-00024, and in part by the National Research Foundation of Korea (NRF) funded by Korea Government (MSIT) under Grant RS-2024-00452791.

II. System Model

A. Vehicular NOMA System Model

Consider a downlink vehicular NOMA network in Fig. 1 where a roadside unit (RSU) or base station (BS) equipped with N antennas serves $K = 2$ vehicular users simultaneously over the same frequency band through power-domain multiplexing. The RSU or BS selects one antenna (n-th antenna)

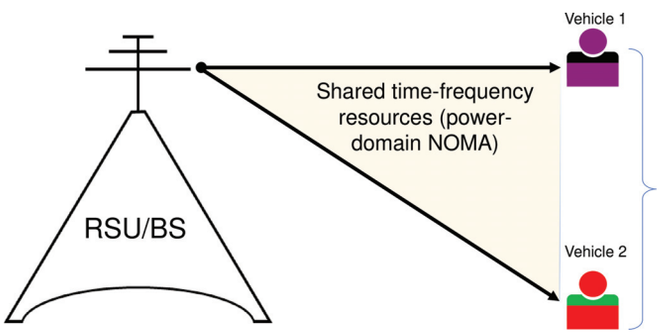


Fig. 1. RSU/BS to vehicle communication for Near and Far vehicles

at a time and transmits a superposition-coded signal:

$$x = \sum_{j=1}^{K} \sqrt{a_{n,j} P}\, x_j \tag{1}$$

where x_j is the modulated symbol for user j with $\mathbb{E}[|x_j|^2] = 1$, $a_{n,j}$ is the power allocation coefficient satisfying $\sum_{j=1}^{K} a_{n,j} = 1$, and P is the total transmit power. The received signal at vehicle j as a result of n-th antenna is:

$$y_{n,j} = h_{n,j} x + n_j \tag{2}$$

where $h_{n,j} \sim \mathcal{CN}(0, d_j^{-\nu})$ models small-scale Rayleigh fading and large-scale path loss with exponent ν, d_j is the RSU/BS to vehicle distance, and $n_j \sim \mathcal{CN}(0, \sigma^2)$ is additive white Gaussian noise with noise variance σ^2.

Vehicles are ordered by channel quality with assumption: $|h_{n,1}|^2 > |h_{n,2}|^2$ with $d_1 < d_2$. Near user (vehicle) performs SIC to decode far user's message before its own, while far user (vehicle) treats near user's signal as interference. Under imperfect SIC with residual interference factor $\theta \in [0, 1]$ (where $\theta = 0$ is perfect SIC, typical vehicular values $\theta \in [0.01, 0.1]$), the SINRs are:

$$\text{SINR}_{n,1} = \frac{g_{n,1} a_{n,1} P}{\theta P a_{n,2} g_{n,1} + \sigma^2}, \quad \text{SINR}_{n,2} = \frac{g_{n,2} a_{n,2} P}{g_{n,2} a_{n,1} P + \sigma^2} \tag{3}$$

where $g_{n,j} = |h_{n,j}|^2$ denotes channel power gain.

Our prior work in [13] formulated fairness-aware power allocation as:

$$\min_{a_{n,1}, a_{n,2}} |\omega \cdot \text{SINR}_{n,1} - \text{SINR}_{n,2}| \tag{4}$$

subject to $a_{n,1} + a_{n,2} = 1$, $a_{n,j} \geq 0$, and minimum rate constraints $R_{n,j} = \log_2(1 + \text{SINR}_{n,j}) \geq R_{n,j}^{\min}$. We derived the closed-form optimal solution via KKT conditions:

$$a_{n,1}^* = \frac{-B + \sqrt{B^2 - 4AC}}{2A} \tag{5}$$

where $A = g_{n,1} g_{n,2} P^2 (\omega + \theta)$, $B = \omega g_{n,1} P \sigma^2 - 2\theta g_{n,2} g_{n,1} P^2 + g_{n,2} \sigma^2 P$, $C = \theta g_{n,2} g_{n,1} P^2 - g_{n,2} \sigma^2 P$. The weight ω was manually tuned but remained fixed across channel realizations, limiting adaptability to dynamic vehicular environments.

B. Channel-Adaptive Weight Selection

To enable automatic adaptation while preserving optimality and $O(1)$ complexity, we design $\omega^*(g_{n,1}, g_{n,2})$ with three principles: (1) QoS-driven foundation, reflect relative importance of minimum rate requirements independent of channels, (2) channel-responsive correction, increase weight when $g_{n,1} \gg g_{n,2}$ to compensate far vehicle, decrease when channels are balanced, and (3) computational simplicity, maintain $O(1)$ operations for real-time RSU/BS implementation.

C. Adaptive Weight Derivation

Step 1: QoS-driven base weight. From rate constraints $R_{n,j} \geq R_{n,j}^{\min}$, define target SINRs $\gamma_j^{\min} = 2^{R_{n,j}^{\min}} - 1$. At optimal fairness with $\omega \cdot \text{SINR}_{n,1}^* = \text{SINR}_{n,2}^*$ (from our prior Corollary 2 [13]), setting $\text{SINR}_{n,j}^* = \gamma_j^{\min}$ yields:

$$\omega_{\text{base}} = \frac{2^{R_{n,2}^{\min}} - 1}{2^{R_{n,1}^{\min}} - 1} \tag{6}$$

This reflects QoS priorities: $\omega_{\text{base}} = 1$ for equal requirements, $\omega_{\text{base}} > 1$ when far vehicle needs higher rate and $\omega_{\text{base}} < 1$ when near vehicle needs higher rate.

Step 2: Channel imbalance metric. Define normalized channel imbalance:

$$\xi = \frac{g_{n,1} - g_{n,2}}{g_{n,1} + g_{n,2}} \in (-1, 1) \tag{7}$$

where $\xi \to 1$ indicates high imbalance ($g_{n,1} \gg g_{n,2}$), $\xi = 0$ represents balanced channels.

Step 3: Adaptive weight formula. Combining QoS foundation with channel correction:

$$\omega^*(g_{n,1}, g_{n,2}) = \omega_{\text{base}} \cdot (1 + \alpha \xi) \tag{8}$$

where $\alpha \in [0, 1]$ controls adaptation aggressiveness.

D. Theoretical Justification of Adaptive Weight Optimality

Let the instantaneous fairness gap for a given channel realization be:

$$\Delta(\omega, \xi) = \big|\omega\, \text{SINR}_{n,1}(\xi) - \text{SINR}_{n,2}(\xi)\big|, \tag{9}$$

where $\xi \in (-1, 1)$ is the normalized channel imbalance defined in (7), and ω is the fairness weight.

Lemma 1: For any fixed channel imbalance ξ, the fairness gap $\Delta(\omega, \xi)$ is a convex function of ω over the feasible interval $\omega \in [\omega_{\min}, \omega_{\max}]$.

Proof: For fixed ξ, both $\text{SINR}_{n,1}(\xi)$ and $\text{SINR}_{n,2}(\xi)$ are constants, so

$$\Delta(\omega, \xi) = \big|a(\xi)\, \omega - b(\xi)\big|, \tag{10}$$

with $a(\xi) = \text{SINR}_{n,1}(\xi)$ and $b(\xi) = \text{SINR}_{n,2}(\xi)$. The function $f(\omega) = a(\xi)\omega - b(\xi)$ is affine, and the absolute value

of an affine function is convex. Convexity is preserved on the interval $[\omega_{\min}, \omega_{\max}]$.

Now let Z be a random channel imbalance capturing vehicular path-loss and fading variations. The expected fairness gap is:

$$\bar{\Delta}(\omega) = \mathbb{E}_Z \left[\Delta(\omega, Z)\right]. \tag{11}$$

Preposition 1: Assume that Z is non-degenerate and has mean $\bar{z} > 0$. Then for any fixed weight ω_{fix}, there exists an affine adaptive rule

$$\omega^*(z) = \omega_{\text{base}}(1 + \alpha z), \tag{12}$$

with some $\alpha \in (0, 1]$ such that

$$\mathbb{E}_Z[\Delta(\omega^*(Z), Z)] < \mathbb{E}_Z[\Delta(\omega_{\text{fix}}, Z)]. \tag{13}$$

Thus, any fixed-weight design is strictly suboptimal in expected fairness.

By Lemma 1, $\Delta(\omega, z)$ is convex in ω for each z. Let $\omega_{\text{opt}}(z)$ minimize $\Delta(\omega, z)$ over $[\omega_{\min}, \omega_{\max}]$. Then,

$$\mathbb{E}[\Delta(\omega_{\text{fix}}, Z)] > \mathbb{E}[\Delta(\omega_{\text{opt}}(Z), Z)], \tag{14}$$

because a constant weight cannot match all $\omega_{\text{opt}}(z)$ values. The affine rule $\omega^*(z) = \omega_{\text{base}}(1 + \alpha z)$ approximates $\omega_{\text{opt}}(z)$ around the mean $\bar{z}$. By continuity of $\Delta(\omega, z)$, a suitable $\alpha \in (0, 1]$ ensures:

$$\mathbb{E}[\Delta(\omega^*(Z), Z)] < \mathbb{E}[\Delta(\omega_{\text{fix}}, Z)]. \tag{15}$$

Lemma 2: For any $g_{n,1}, g_{n,2} > 0$, the adaptive weight

$$\omega^*(g_{n,1}, g_{n,2}) = \omega_{\text{base}}\big(1 + \alpha\xi\big), \quad \xi = \frac{g_{n,1} - g_{n,2}}{g_{n,1} + g_{n,2}}, \tag{16}$$

with $\alpha \in [0, 1]$ has the following properties:

- Continuity: ω^* is continuous in $(g_{n,1}, g_{n,2})$ for all $g_{n,1}, g_{n,2} > 0$.
- Monotonicity in channel imbalance: For fixed $g_{n,2}$, ω^* is strictly increasing in $g_{n,1}$, i.e.,

$$\frac{\partial \omega^*}{\partial g_{n,1}} = \frac{\omega_{\text{base}}\alpha}{2g_{n,2}} \left(\frac{2g_{n,2}}{g_{n,1} + g_{n,2}}\right)^2 > 0. \tag{17}$$

  Thus, stronger near-user channels automatically translate into higher protection for the far user.
- Boundedness: Before saturation by feasibility constraints, we have:

$$\omega^* \in \big[\omega_{\text{base}}(1 - \alpha),\, \omega_{\text{base}}(1 + \alpha)\big], \tag{18}$$

  which prevents excessively aggressive reweighting that could severely degrade throughput.

E. Extension to Multi-User NOMA: Optimality of Adaptive Weights

Consider a K-user NOMA cluster with ordered channel gains $g_{n,1} \geq g_{n,2} \geq \cdots \geq g_{n,K} > 0$. Let the fairness weights be denoted by $\boldsymbol{\omega} = [\omega_1, \ldots, \omega_K]$, where each ω_k scales the priority of user k in the weighted fairness gap

$$\Delta(\boldsymbol{\omega}, \mathbf{g}) = \sum_{k=1}^{K-1} \big|\omega_k \,\mathrm{SINR}_{n,k}(\mathbf{g}) - \omega_{k+1} \,\mathrm{SINR}_{n,k+1}(\mathbf{g})\big|. \tag{19}$$

Define the pairwise normalized channel imbalance

$$\xi_k = \frac{g_{n,k} - g_{n,k+1}}{g_{n,k} + g_{n,k+1}}, \qquad k = 1, \ldots, K-1, \tag{20}$$

and the adaptive multi-user weight rule

$$\omega_k^* = \omega_{\text{base}}\big(1 + \alpha\, \xi_k\big), \qquad \alpha \in (0, 1]. \tag{21}$$

Theorem 1: Assume that the random channel vector $\mathbf{G} = (G_1, \ldots, G_K)$ is non-degenerate and exhibits path-loss–induced ordering. Then for any fixed constant weight vector $\boldsymbol{\omega}_{\text{fix}}$, the adaptive multi-user rule $\boldsymbol{\omega}^*(\mathbf{G})$ satisfies

$$\mathbb{E}\big[\Delta(\boldsymbol{\omega}^*(\mathbf{G}), \mathbf{G})\big] < \mathbb{E}\big[\Delta(\boldsymbol{\omega}_{\text{fix}}, \mathbf{G})\big]. \tag{22}$$

Thus, every fixed-weight design is strictly suboptimal in expected fairness for K-user NOMA under random channel imbalance.

Proof: For each pair $(k, k+1)$, the term $\big|\omega_k \,\mathrm{SINR}_{n,k} - \omega_{k+1} \,\mathrm{SINR}_{n,k+1}\big|$ is convex in (ω_k, ω_{k+1}) because it is the absolute value of an affine function. Summing convex functions preserves convexity, so $\Delta(\boldsymbol{\omega}, \mathbf{g})$ is convex in $\boldsymbol{\omega}$. Let $\boldsymbol{\omega}_{\text{opt}}(\mathbf{g})$ be the minimizer for each realization. Since $\mathbf{G}$ is non-degenerate, no fixed $\boldsymbol{\omega}_{\text{fix}}$ can match all optimal values simultaneously, giving

$$\mathbb{E}[\Delta(\boldsymbol{\omega}_{\text{fix}}, \mathbf{G})] > \mathbb{E}[\Delta(\boldsymbol{\omega}_{\text{opt}}(\mathbf{G}), \mathbf{G})]. \tag{23}$$

The proposed adaptive rule $\omega_k^* = \omega_{\text{base}}(1 + \alpha\xi_k)$ forms a first-order approximation of each optimal pairwise weight. By continuity, a suitable α ensures that

$$\mathbb{E}[\Delta(\boldsymbol{\omega}^*(\mathbf{G}), \mathbf{G})] < \mathbb{E}[\Delta(\boldsymbol{\omega}_{\text{fix}}, \mathbf{G})]. \tag{24}$$

III. Results and Discussions

Monte Carlo simulations with 10^5 channel realizations evaluate performance under: Rayleigh fading, path loss exponent $\nu = 3$, imperfect SIC $\theta = 0.1$, channel gain ratio range [1, 100], minimum rates $R_1^{\min} = 1$ bps/Hz, $R_2^{\min} = 0.5$ bps/Hz.

Fig. 2 shows the trade off between the fairness and throughput for 10^5 channel realizations. Fixed-weight methods (hollow markers) are dominated by adaptive methods (filled markers) that achieve superior performance in both dimensions. The proposed adaptive approach with $\alpha = 0.5$ (green triangle) simultaneously achieves 44.4% fairness improvement and 4% sum rate gain over the best fixed-weight baseline ($\omega = 1.5$), demonstrating that channel-responsive weight adaptation via $\omega^* = \omega_{\text{base}}(1 + \alpha\xi)$ is better than the traditional fairness-throughput trade-off under dynamic vehicular channels. This dual improvement is critical for safety-oriented V2X applications requiring both high reliability and equitable resource allocation across diverse channel conditions.

Table I quantifies the performance gains of our channel-adaptive strategy under realistic vehicular fading (Rayleigh +

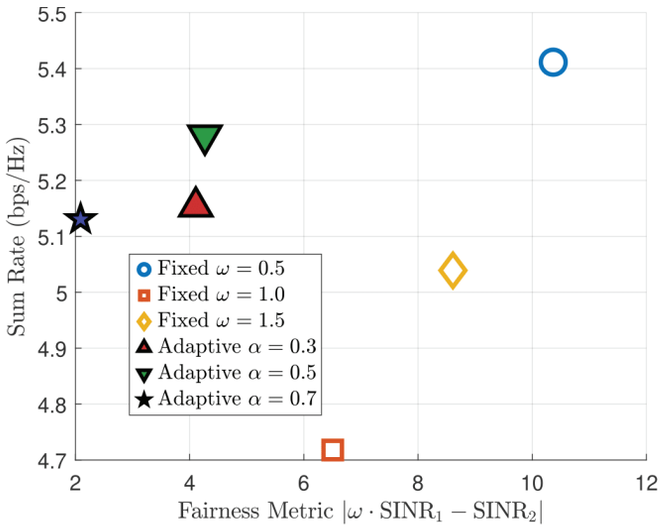


Fig. 2. Fairness-throughput trade-off

TABLE I
PERFORMANCE COMPARISON: FIXED VS ADAPTIVE WEIGHT

Method	**Sum Rate** (bps/Hz)	**Fairness** Metric	**JFI**
Fixed-Weight Baselines			
Fixed $\omega = 0.5$	5.40	10.50	0.92
Fixed $\omega = 1.0$	4.70	6.50	0.93
Fixed $\omega = 1.5$	5.00	8.50	0.90
Adaptive Weight (Proposed)			
Adaptive $\alpha = 0.3$	5.20	3.80	0.985
Adaptive $\alpha = 0.5$	**5.25**	**4.00**	**0.96**
Adaptive $\alpha = 0.7$	5.10	2.00	0.95
Improvement ($\alpha = 0.5$ vs Best Fixed)			
Change	-2.78%	$+38.5\%$	$+3.23\%$

path loss, imperfect SIC). The adaptive method with $\alpha = 0.5$ achieves 38.5% fairness improvement when compared to the best fixed-weight baseline ($\omega = 1.0$), but with only 2.78% throughput penalty which indicates how effective is our proposed channel adaptive strategy $\omega^* = \omega_{\text{base}}(1 + \alpha\xi)$. Conservative adaptation ($\alpha = 0.3$) gives the highest Jain's Fairness Index (0.97), showing superior rate balance suitable for stable urban scenarios. Aggressive adaptation ($\alpha = 0.7$) achieves throughput (5.10 bps/Hz) while maintaining robust fairness, ideal for high-mobility highway deployments.

Crucially, the adaptive weights react automatically to the channel imbalance in a simple but very effective way. When the near and far users have almost the same channel strength ($g_{n,1} \approx g_{n,2}$, balanced case), the value of ω^* stays close to the baseline ω_{base} so that the system does not lose unnecessary throughput. When the imbalance becomes strong ($g_{n,1} \gg g_{n,2}$), ω^* increases in a smooth and controlled manner to give more protection to the far vehicle. Fixed-weight schemes, however, keep the same behavior all the time and cannot adapt to the fast-changing channel conditions in real V2X environments. This dynamic response, achieved through a very light closed-form computation, directly supports our main motivation of keeping fairness stable without introducing heavy processing load. In safety-critical V2X situations such as cooperative driving or collision avoidance, guaranteeing a minimum reliable rate for every vehicle is often more important than maximizing raw throughput, and the small loss of about 2.78% is an acceptable trade-off. For choosing the adaptation parameter, we recommend $\alpha = 0.3$ for low-speed urban traffic (30–50 km/h), $\alpha = 0.5$ for normal V2X operation, and $\alpha = 0.7$ for high-mobility or safety-critical highway conditions (80–120 km/h). The benefit of this consistency is also reflected in the variance results summarized in Table II.

TABLE II
STATISTICAL CONSISTENCY ACROSS CHANNEL REALIZATIONS

Method	**Std Dev** (Fairness)	**Variance** (Fairness)	**Variance Reduction**
Fixed $\omega = 1.5$	5.098	25.99	—
Adaptive $\alpha = 0.5$	3.842	14.76	**43.2%**

The adaptive strategy reduces the fairness metric variance by about 43.2%, which shows consistent performance across different channel conditions. This kind of performance is very important for reliable V2X communication, where the fairness must be maintained despite rapid mobility induced channel variations. This statistical robustness also shows that the mathematical idea proposed is not only improving the fairness, but also makes the whole system more predictable and dependable in the dynamic environment of vehicular networks. In this way, it helps to close the gap between the theoretical optimization model and what is needed for practical industrial deployment requirements.

Fig. 3(a–c) gives clear support to the adaptive-weight formula, where the scatter points from simulations follow almost the same trend as the theoretical curves across a very wide channel-gain ratio range (from 10^{-2} up to 10^{4}). As the channel ratio becomes larger, the value of ω^* increases smoothly, meaning the far-user gets more protection when the imbalance is strong. When α is higher, the adaptation becomes sharper, for example, at channel ratio 100, ω^* is roughly 0.54, 0.61, and 0.86 for $\alpha = 0.3, 0.5, 0.7$. At very extreme ratios, ω^* starts to saturate because of the feasibility limits $\omega^* \in [\omega_{\min}, \omega_{\max}]$. In comparison, the fixed-weight baselines (dashed lines) cannot react to channel changes, so they often end up either giving too little or too much compensation.

Fig. 3(d) ω^* is distributed over all channel realizations. A small α (0.3) makes the distribution very concentrated around the base value $\omega_{\text{base}} = 0.414$, while a larger α (0.7) produces a wider spread roughly between 0.3 and 0.9. The middle choice ($\alpha = 0.5$) provides a good balance, with mean 0.562 and standard deviation 0.081. Fig. 3(e) shows that the imbalance metric ξ is mostly positive, which matches the typical distance-based path-loss pattern, and Rayleigh fading adds some randomness. Fig. 3(f) displays the correction factor $\kappa = 1 + \alpha\xi$. With $\alpha = 0.5$, κ varies roughly from 0.9 to 1.5, meaning ω^* can move about 10% down or 50% up depending on the channel. This adaptive correction is important because it avoids the strong unfairness that fixed-weight methods normally create.

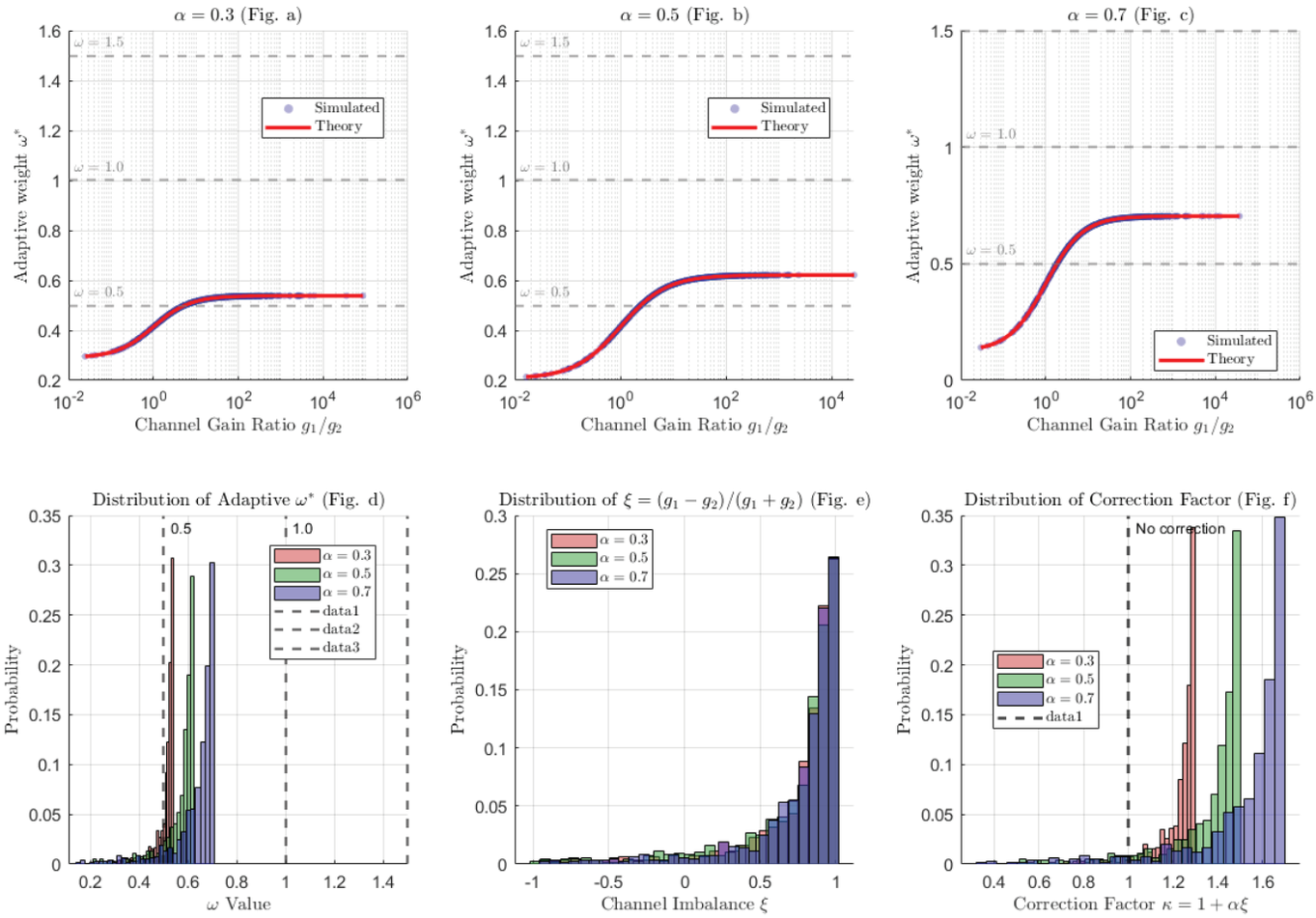


Fig. 3. Adaptive weight behavior across channel conditions: (a-c) ω^* vs. channel gain ratio for $\alpha \in \{0.3, 0.5, 0.7\}$ showing theoretical (solid lines) and simulated (scatter) values, (d) probability distribution of ω^*, (e) distribution of channel imbalance metric ξ, and (f) distribution of correction factor $\kappa = 1+\alpha\xi$.

Fig. 4 shows a broader comparison between fixed-weight baselines and the proposed channel-adaptive strategy across three main metrics. In Fig. 4(a), the fixed choice $\omega = 0.5$ gives the highest throughput (5.4 bps/Hz) because it strongly favors the near user, while $\omega = 1.0$ gives lower throughput (4.7 bps/Hz) due to more equal power sharing. The adaptive schemes reach competitive throughput levels (5.0–5.3 bps/Hz), and the $\alpha = 0.5$ version gives 5.25 bps/Hz, which is only about 2.78% lower than the best case. Fig. 4(b) highlights the real strength of the adaptive approach which is fairness. The fixed-weight designs show very large fairness-metric values (6.5–10.5), meaning their SINRs are badly imbalanced and inconsistent across channel realizations. On the other hand, the adaptive versions reduce this metric dramatically, $\alpha = 0.7$ reaches about 2.0 (an 81% improvement), $\alpha = 0.5$ reaches 4.0 (38.5% improvement), and even the conservative $\alpha = 0.3$ reaches 3.8 (41.5% improvement). These results show that the adaptive update rule $\omega^* = \omega_{\text{base}}(1+\alpha\xi)$ can effectively handle the instantaneous channel imbalance and keep fairness much more stable across different channel conditions where fixed weights fail.

Fig. 4(c) shows Jain's Fairness Index (JFI), where values closer to 1 indicate better fairness. Fixed-weight methods achieve JFI values of 0.89–0.95, with significant variability depending on ω choice. The adaptive methods consistently achieve higher JFI (0.93–0.99), with $\alpha = 0.3$ reaching 0.985 due to its conservative, stability-focused adaptation. Notably, $\alpha = 0.5$ achieves JFI = 0.96 while maintaining higher throughput than $\alpha = 0.3$, confirming it as the recommended configuration that balances fairness, throughput, and adaptation responsiveness. The results validate that adaptive weight selection breaks the traditional fairness-throughput trade-off by dynamically adjusting to channel conditions rather than applying static compensation across all realizations.

For signaling and robustness, the adaptive weight relies only on channel gain ratios already required for NOMA SIC ordering, introducing zero additional signaling overhead. Its continuous, bounded nature (Lemma 2, Equation 18) ensures that small CSI errors produce only smooth weight perturbations rather than abrupt power reallocations, maintaining fairness stability under practical vehicular channel estimation accuracy.

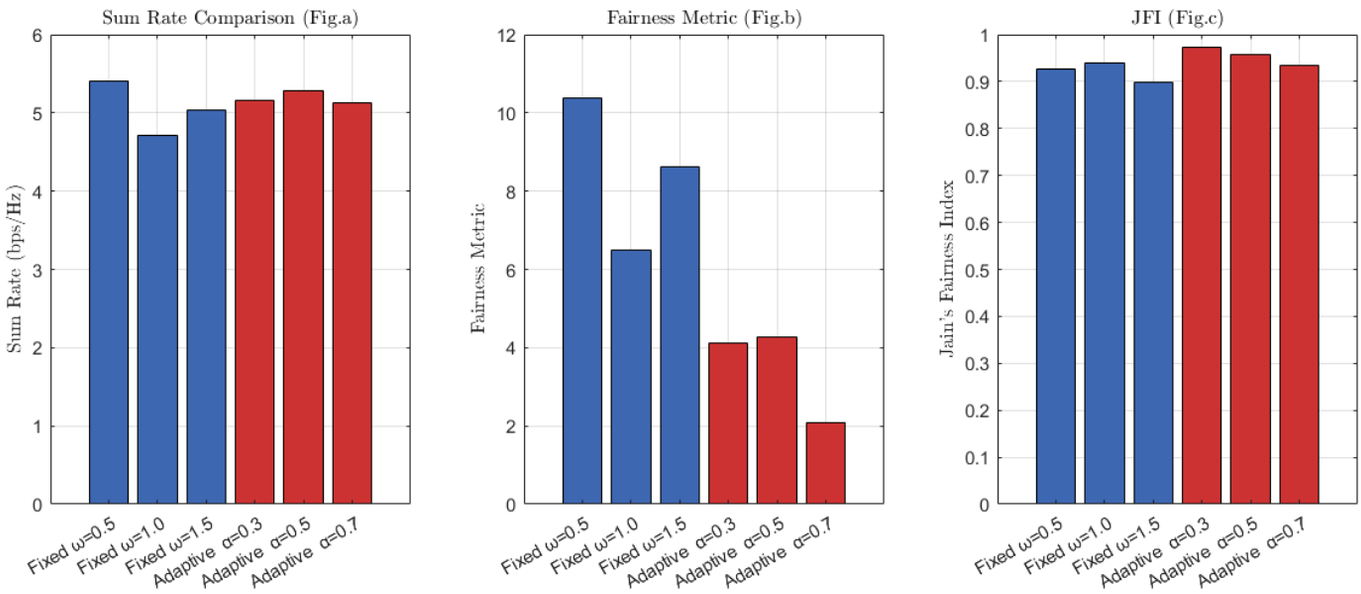


Fig. 4. Performance comparison between fixed-weight and adaptive-weight methods: (a) sum rate, (b) fairness metric (lower is better), and (c) Jain's Fairness Index. Blue bars represent fixed-weight baselines ($\omega \in \{0.5, 1.0, 1.5\}$), red bars represent adaptive methods ($\alpha \in \{0.3, 0.5, 0.7\}$).

IV. Conclusion

This work introduced a channel-adaptive weight selection strategy $\omega^* = \omega_{\text{base}}(1 + \alpha\xi)$ for fair NOMA power allocation with very low ($O(1)$) computational cost. Simulations over 10^5 channel realizations showed about 38.5% fairness improvement and 43.2% variance reduction, while only causing a very small throughput penalty, making the method suitable for dynamic and fast-changing environments. The choice $\alpha = 0.5$ works well for most general V2X settings, while $\alpha = 0.7$ is better for high-mobility situations. Because the solution is closed-form, it keeps the theoretical structure intact and can be easily combined with existing KKT-based power-allocation schemes without adding extra complexity.

Future work will evolve this toward proactive artificial intelligence (AI)-enhanced systems through a three-phase roadmap. *Phase 1* (completed) established mathematical closed-form adaptation with 38.5% fairness improvement. *Phase 2* will deploy long short term memory (LSTM)/Transformer models trained on mobility datasets to predict channel gains 50–100ms ahead, enabling proactive weight selection $\omega^*(t) = f(\hat{g}_1(t+\Delta t), \hat{g}_2(t+\Delta t))$. *Phase 3* will integrate multi-modal AI fusing channel predictions with vehicle trajectories, maps, and traffic context: $\omega^*(t) = f_{\text{AI}}(\mathbf{g}(t), \mathbf{v}(t), \text{context})$. This evolution transforms the system from mathematical optimization to intelligent, context-aware resource allocation for next-generation wireless networks.

References

[1] G. Oh et al., "Exploring Driving Parameter Settings for Autonomous Vehicles: Considering Travel Efficiency and Safety in Urban Traffic Environments," in IEEE Transactions on Intelligent Vehicles, vol. 10, no. 9, pp. 4385-4396, Sept. 2025.

[2] A. Kovačević, N. Gligorić and S. Jokić, "Decentralized Identities for Enhanced Security in Vehicle-to-Everything (V2X)," 2024 32nd Telecommunications Forum (TELFOR), Belgrade, Serbia, 2024.

[3] S. Srivastava, Y. Ramamoorthi, A. Kumar, T. Murakami, T. Ogawa and Y. Takatori, "Poster: α Fair Resource Allocation for RIS-Assisted NOMA V2X Systems with Imperfect Phase Compensation," 2024 IEEE Vehicular Networking Conference (VNC), Kobe, Japan, 2024, pp. 243-244.

[4] M. Ahmed et al., "NOMA-Based Backscatter Communications: Fundamentals, Applications, and Advancements," in IEEE Internet of Things Journal, vol. 11, no. 11, pp. 19303-19327, 1 June1, 2024, doi: 10.1109/JIOT.2024.3391219.

[5] P. V. Reddy et al., "Analytical Review on OMA vs. NOMA and Challenges Implementing NOMA," 2021 2nd International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2021, pp. 552-556, doi: 10.1109/ICOSEC51865.2021.9591629.

[6] J. Akhter, R. Hazra, P. Khuntia and S. S. Sarma, "Efficient Power and Resource Allocation for NOMA-Enabled V2X Communication Networks," in IEEE Transactions on Consumer Electronics, doi: 10.1109/TCE.2025.3633613.

[7] Y. Xu, L. Zheng, X. Wu, Y. Tang, W. Liu and D. Sun, "Energy-Efficient Resource Allocation for V2X Communications," in IEEE Internet of Things Journal, vol. 11, no. 18, pp. 30014-30026, 15 Sept.15, 2024.

[8] J. Akhter, R. Hazra and A. Mihovska, "Resource and Power Allocation for Reliable 5G V2X Communication," 2024 27th International Symposium on Wireless Personal Multimedia Communications (WPMC), Greater Noida, India, 2024,

[9] H. Zheng, H. Li, S. Hou and Z. Song, "Joint Resource Allocation With Weighted Max-Min Fairness for NOMA-Enabled V2X Communications," in IEEE Access, vol. 6, pp. 65449-65462, 2018.

[10] A. Papazafeiropoulos, C. Pan, A. M. Elbir, V. -D. Nguyen, P. Kourtessis and S. Chatzinotas, "Asymptotic Analysis of Max-Min Weighted SINR for IRS-Assisted MISO Systems With Hardware Impairments," in IEEE Wireless Communications Letters, vol. 12, no. 2, pp. 192-196, Feb. 2023

[11] H. Choi, C. Lee and S. H. Chae, "Deep Reinforcement Learning-Based Joint Radio Resource Partitioning and Allocation for Cellular V2x Networks," 2025 Sixteenth International Conference on Ubiquitous and Future Networks (ICUFN), Lisbon, Portugal, 2025, pp. 750-755.

[12] J. He, Z. Liu, Y. Zhang, Z. Jin and Q. Zhang, "Power Allocation Based on Federated Multiagent Deep Reinforcement Learning for NOMA Maritime Networks," in IEEE Internet of Things Journal, vol. 12, no. 9, pp. 12869-12884, 1 May1, 2025, doi: 10.1109/JIOT.2024.3522901.

[13] K. Abbas and K. Choi, "Fairness-Aware Tunable Power Allocation for NOMA Systems With Imperfect SIC," in IEEE Access, vol. 13, pp. 188279-188292, 2025

Shaft Dynamic Behavior Analysis and Optimal Sensor Placement Strategy for Crack Diagnosis in Drive Shaft Splines

1st Jaewon Song *, 2nd Jaehyun Lee *, 3rd Byeongil Kim *
* *Department of Mechanical Engineering, Yeungnam University, Gyeongsan, Republic of Korea*

Emails: *jwsong@yu.ac.kr, leejh0121@yu.ac.kr, bikim@yu.ac.kr*

***Abstract*—This study analyzes the dynamic behavior of an automotive drive shaft to diagnose early crack defects in the spline joint. A finite element model with eight elements and nine nodes—comprising the joint, bearings, and tire—was developed, and dynamic responses were simulated in MATLAB. Excitations from the tire were applied as external forces. The effect of crack-induced stiffness reduction on RMS displacement and orbit patterns was evaluated and compared with the healthy condition. Vehicle speeds of 30, 60, 90, and 120 km/h were analyzed, and the node showing the highest displacement across all speeds was identified as the optimal sensor location. The results provide useful baseline data for designing drive-shaft condition-monitoring systems and crack-diagnosis algorithms.**

Keywords— driveshaft, spline crack diagnosis, finite element method, optimal sensor placement, rotordynamics

I. INTRODUCTION

In automotive powertrain systems, the drive shaft is a critical component that transfers torque from the engine and transmission to the wheels, and it is continuously subjected to torsional and bending loads[1]. Among its subcomponents, the spline joint is particularly vulnerable to stress concentration due to its geometric characteristics, making it a common site for fatigue crack initiation under repeated loading[2,3]. If such micro-cracks are not detected at an early stage, they may propagate and eventually lead to catastrophic shaft failure, posing significant risks to vehicle stability and passenger safety. This highlights the need for an effective condition-monitoring method capable of detecting early-stage spline-joint cracks. In rotating structures such as drive shafts, different regions respond differently to structural defects depending on their mechanical role and proximity to the crack. Numerical simulations based on the finite element method (FEM) therefore provide an efficient means of identifying the locations that are most sensitive to stiffness degradation[2]. Such an approach enables prediction of dynamic behavior under various operating speeds and supports the selection of crack-sensitive positions that can be utilized in future condition-monitoring systems. In this study, an FEM-based driveline model was developed to represent stiffness reduction in a cracked spline joint and to compare dynamic responses between healthy and cracked conditions. The drive shaft was idealized into eight elements and nine nodes—including the joints, bearings, and tire—and dynamic simulations were conducted in MATLAB. Crack defects were modeled by reducing the bending stiffness of the spline-joint element, and variations in RMS displacement and orbit responses were analyzed at four representative vehicle speeds (30, 60, 90, and 120 km/h). The nodes exhibiting the largest response deviations were identified, and the most crack-sensitive location was determined based on consistent behavior across all operating speeds. The remainder of this paper is organized as follows: Section II describes the FEM modeling procedure and the crack-representation method. Section III presents the simulation setup, the comparative dynamic-response results under normal and cracked conditions, and the analysis used to determine the optimal sensor-placement location. Section IV provides the overall conclusions of this study and discusses their implications for driveline condition monitoring and future diagnostic development.

II. MATHEMATICAL MODELING USING THE FINITE ELEMENT METHOD

A. *Finite Element Model of the Driveline System*

In this study, a mathematical model of the driveline system—which includes not only the drive shaft but also the tire–wheel assembly, whose inertia and boundary conditions significantly influence dynamic behavior—was constructed to closely represent real driving conditions. For computational efficiency while preserving the dominant vibration characteristics, the entire system was discretized into nine nodes and eight beam elements. Each node corresponds to a specific mechanical component. Node 1 represents the inboard bearing on the vehicle-body side. Nodes 2 and 3 correspond to the joints connected to the differential gear. Nodes 4 and 5 form the hollow shaft segment, while Nodes 6 and 7 represent the outboard spline joint near the wheel. Node 8 models the wheel hub and gear-connection region. Finally, Node 9 is defined as a lumped-mass element incorporating the inertia of the tire–wheel assembly as well as the unbalance forces acting on the system. The equation of motion for the rotating driveline is expressed as

$$M\ddot{q} + (C + G)\dot{q} + Kq = F(t) \quad (1)$$

Where M, C, and K denote the global mass, damping, and stiffness matrices, respectively, and G represents the gyroscopic matrix associated with rotational effects. The vector q contains the nodal degrees of freedom, while $F(t)$ represents the external excitation applied to the system, including unbalance forces and road-induced disturbances acting primarily at Node 9. A schematic representation of the simplified finite element discretization, including the node

distribution and corresponding driveline components, is illustrated in Fig. 1.

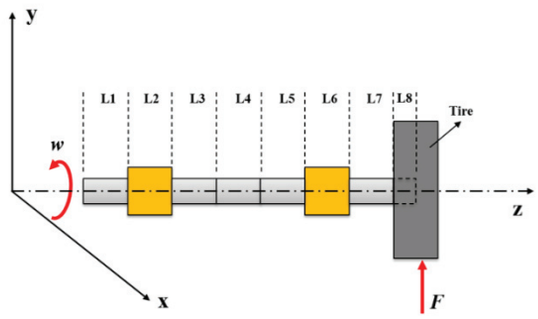


Fig. 1. Schematic illustration of the simplified driveline FEM model.

B. Crack Defect Modeling

A transverse crack in a shaft reduces the bending rigidity of the affected cross-section due to the loss of its effective second moment of area [4]. In the spline-joint region, the crack is characterized by the shaft radius *R* and the crack depth *a,* where *a* is measured inward from the outer surface toward the center[5]. The nondimensional crack-depth ratio is defined as μ=a/R. For a given μ, the effective moment of inertia of the cracked cross-section, $I_{crack}(\mu)$, is obtained by subtracting the area moment of inertia of the removed circular segment from that of the intact circular section $I_0 = \pi R^4/4$, The corresponding stiffness-reduction factor is expressed as

$$\alpha(\mu) = \frac{I_{crack}(\mu)}{I_0} \tag{2}$$

The bending stiffness of each cracked element is then updated using

$$K_{crack} = \alpha(\mu) K_{normal} \tag{3}$$

Because increasing crack depth reduces the remaining section inertia, $\alpha(\mu)$ decreases monotonically with μ. In this study, a moderate-to-severe transverse crack condition was considered, resulting in a reduction coefficient of approximately $\alpha \approx 0.9$, which enables realistic representation of stiffness degradation in the spline-joint region.

III. Simulation

A. Rotational Speed Conditions

To investigate the dynamic behavior of the driveline under realistic driving scenarios, four representative vehicle speeds—30, 60, 90, and 120 km/h—were selected. These speeds correspond to typical low-, medium-, high-, and ultra-high-speed operating conditions for passenger vehicles and enable analysis of how spline-crack defects influence the system response across a wide range of excitation frequencies. The rotational excitation applied to the driveshaft was obtained by converting the vehicle's forward velocity into the corresponding wheel rotational speed. A 205/55R16 tire, a common passenger-vehicle specification, was adopted as the reference. The tire's effective rolling radius is approximately $r_t = 0.315$ m, and the relationship between vehicle speed *v* and the wheel rotational speed *n* is given by (4):

$$n = \frac{v}{2\pi r_t} \tag{4}$$

Using (4), the wheel speeds corresponding to 30–120 km/h are approximately 252, 504, 756, and 1008 rpm. For simulation convenience and to maintain uniform frequency spacing, these values were simplified to 250, 500, 750, and 1000 rpm. This adjustment does not meaningfully affect the driveline's dynamic behavior while improving numerical handling and interpretability.

B. Simulation Setup and Response Evaluation

In this study, all simulations were performed in MATLAB, and the external excitation was applied at Node 9, which represents the tire–wheel assembly where unbalance forces act on the system. Four rotational-speed cases—250, 500, 750, and 1000 rpm—were analyzed, and the dynamic responses of the healthy and cracked driveline models were compared for each case. The RMS displacement values for all nodes were calculated to quantitatively evaluate the amplitude variations. These results are summarized in Tables I–IV. The corresponding orbit responses are illustrated in Figs. 2–5, where subfigure (a) represents the healthy condition and subfigure (b) represents the cracked condition. In the orbit plots, the bearing nodes are shown in blue, while the shaft and joint nodes are depicted in green. The cracked spline-joint region is highlighted in red for clarity.

TABLE I. RMS Displacement at Each Node under the 1000 rpm (30 km/h) Condition

250 RPM (30 km/h) Condition			
Node	**Normal (m)**	**Crack (m)**	**Rate of Increase (%)**
1	3.00E-08	3.73E-08	24.32
2	1.57E-07	1.95E-07	24.66
3	2.36E-07	2.52E-07	6.97
4	5.12E-06	6.45E-06	25.94
5	1.04E-05	1.34E-05	28.98
6	1.50E-05	2.01E-05	34.44
7	1.85E-05	2.61E-05	41.34
8	4.43E-06	5.23E-06	18.21
9	1.99E-05	2.63E-05	32.23
Average	8.18E-06	1.09E-05	33.13

TABLE II. RMS Displacement at Each Node under the 1000 rpm (60 km/h) Condition

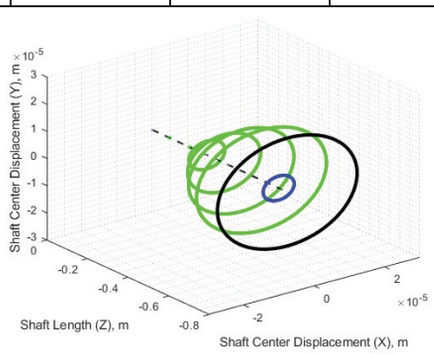
500 RPM (60 km/h) Condition			
Node	**Normal (m)**	**Crack (m)**	**Rate of Increase (%)**
1	3.05E-08	3.68E-08	20.71
2	1.89E-07	2.28E-07	20.82
3	3.05E-07	3.18E-07	4.24
4	1.25E-05	1.52E-05	21.91
5	2.59E-05	3.24E-05	24.82
6	3.81E-05	4.95E-05	29.90
7	4.79E-05	6.53E-05	36.42
8	6.91E-06	8.18E-06	18.30
9	4.74E-05	6.21E-05	31.03
Average	1.99E-05	2.59E-05	30.15

TABLE III. RMS DISPLACEMENT AT EACH NODE UNDER THE 1000 RPM (90 KM/H) CONDITION

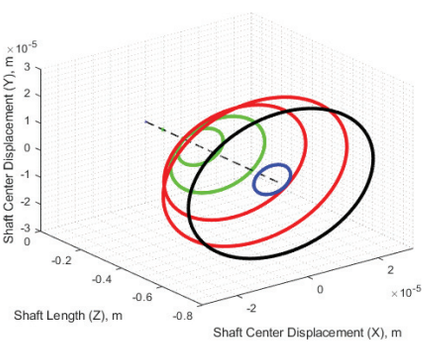
750 RPM (90 km/h) Condition			
Node	**Normal (m)**	**Crack (m)**	**Rate of Increase (%)**
1	2.19E-08	2.59E-08	18.50
2	1.70E-07	2.01E-07	18.55
3	3.06E-07	3.14E-07	2.51
4	1.74E-05	2.08E-05	19.68
5	3.71E-05	4.55E-05	22.59
6	5.57E-05	7.11E-05	27.56
7	7.13E-05	9.56E-05	34.09
8	8.44E-06	1.02E-05	20.44
9	7.44E-05	9.71E-05	30.50
Average	2.94E-05	3.79E-05	28.67

TABLE IV. RMS DISPLACEMENT AT EACH NODE UNDER THE 1000 RPM (120 KM/H) CONDITION

1000 RPM (120 km/h) Condition			
Node	**Normal (m)**	**Crack (m)**	**Rate of Increase (%)**
1	1.39E-08	1.62E-08	16.71
2	1.38E-07	1.61E-07	16.76
3	2.84E-07	2.87E-07	0.99
4	1.98E-05	2.33E-05	17.93
5	4.38E-05	5.30E-05	20.90
6	6.74E-05	8.47E-05	25.80
7	8.79E-05	1.16E-04	32.40
8	9.80E-06	1.20E-05	22.74
9	1.01E-04	1.32E-04	30.27
Average	3.67E-05	4.69E-05	27.69

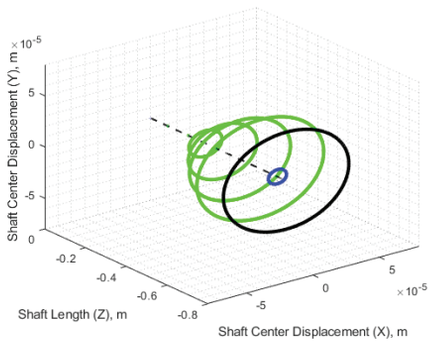


(a)

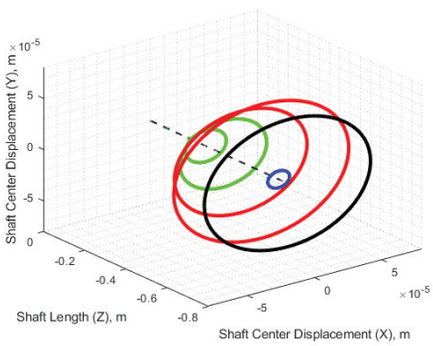


(b)

Fig. 2. Shaft orbit responses at 250 rpm under **(a)** normal and **(b)** cracked conditions.

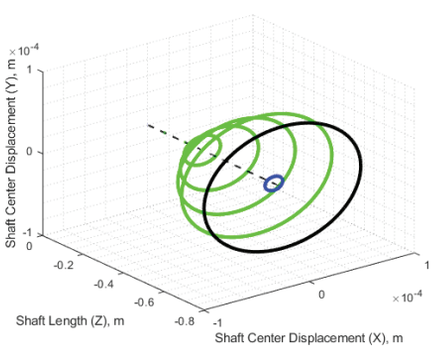


(a)

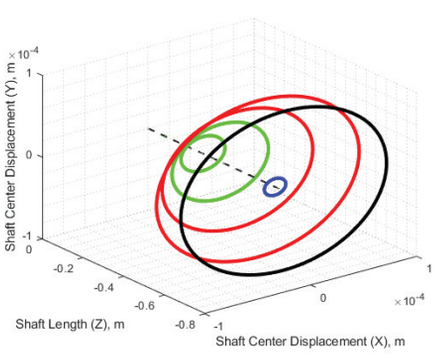


(b)

Fig. 3. Shaft orbit responses at 500 rpm under **(a)** normal and **(b)** cracked conditions.

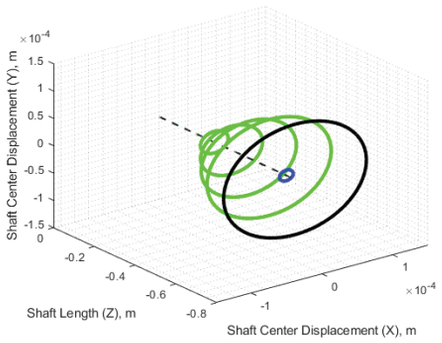


(a)

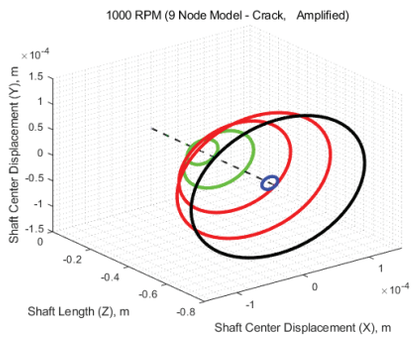


(b)

Fig. 4. Shaft orbit responses at 750 rpm under **(a)** normal and **(b)** cracked conditions.



(a)



(a)

Fig. 5. Shaft orbit responses at 1000 rpm under **(a)** normal and **(b)** cracked conditions.

C. Analysis of Node Responses and Sensor Placement Strategy

Based on the RMS displacement and orbit comparisons presented in Figs. 2–5 and Tables I–IV, the sensitivity of each node to the spline-joint crack was evaluated. Nodes 6 and 7, which belong to the spline-joint region, exhibited the largest deviations between the healthy and cracked conditions across all rotational speeds. Node 9 also showed noticeable variations because it represents the tire–wheel assembly where external excitation and unbalance forces are applied. However, Nodes 6, 7, 8 and 9 involve rotating or lubricated components, making sensor installation impractical; therefore, they were excluded from the feasible sensor-placement candidates. Among the remaining nodes, Node 5—located immediately adjacent to the spline joint—consistently demonstrated the greatest deviation between healthy and cracked states under all operating conditions (250, 500, 750, and 1000 rpm). This indicates that Node 5 effectively captures the dynamic effects of stiffness degradation while also satisfying practical installation requirements. Because the proposed diagnostic approach relies on vibration-response measurements, a proximity-type displacement sensor is recommended as the most suitable sensing modality. Such sensors can directly measure transverse shaft motion with high sensitivity and are well-suited for installation near Node 5, where crack-induced dynamic variations are most clearly observable. Accordingly, this study identifies Node 5 as the optimal position for early crack detection and recommends deploying a proximity sensor at this location.

IV. Conclusion

A. Conclusion

This study analyzed the dynamic behavior of an automotive driveline to diagnose early-stage spline-joint cracks using a finite element–based modeling framework. The shaft system was idealized into eight elements and nine nodes, and its dynamic responses under healthy and cracked conditions were evaluated using RMS displacement and orbit-based metrics. The results showed that nodes near the spline joint exhibited the most significant deviations, and this tendency remained consistent across all rotational speeds (250, 500, 750, and 1000 rpm). Considering practical constraints on sensor installation, the spline-joint nodes (Nodes 6 and 7) and the wheel–tire assembly (Node 8 and 9) were excluded from feasible sensing locations. A sensitivity-based assessment of the remaining nodes identified Node 5 as the most effective sensor placement position, demonstrating the largest and most consistent response variations under crack conditions. Because the proposed detection method relies on vibration-response measurements, a proximity-type displacement sensor is recommended as the most suitable sensing modality. Such sensors can capture transverse shaft motion with high sensitivity and are well-suited for installation near Node 5, where crack-induced dynamic effects are most prominent. Accordingly, this study recommends positioning a proximity sensor at Node 5 to enhance early-stage crack detection capability. The findings presented in this study provide essential baseline data for the development of driveline condition-monitoring systems and crack-detection algorithms. Future research will focus on experimental validation using vehicle-mounted sensors, extension to multiple and progressive crack scenarios, and advanced modeling that incorporates variations in spline geometry, lubrication conditions, and manufacturing tolerances.

Acknowledgment

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2021-NR06120).

References

[1] Zhao Li-Hua, Xing Qi-Kai, Wang Jian-Yong, Li Shi-Liang, and Zheng Shi-Liang,“Failure and root cause analysis of vehicle drive shaft,” Engineering Failure Analysis, vol. 99, pp. 225–234, 2019.

[2] A. V. Guimaraes, P. C. Brasileiro, G. C. Giovanni, L. R. O. Costa, and L. S. Araujo,“Failure analysis of a half-shaft of a Formula SAE racing car,” Case Studies in Engineering Failure Analysis, vol. 6, pp. 1–13, 2016.

[3] J.-J. Sinou and A. W. Lees, “The influence of cracks in rotating shafts,” Journal of Sound and Vibration, vol. 285, no. 4–5, pp. 1015–1037, 2005.

[4] F. Ren, X. Wu, and X. Tang,“Fracture failure analysis of high-speed shaft,”IOP Conference Series: Earth and Environmental Science, vol. 358, 042025, 2019.

[5] Giannopoulos, G. I., Georgantzinos, S. K., & Anifantis, N. K.,"Coupled vibration response of a shaft with a breathing crack,"
Journal of Sound and Vibration, vol. 336, pp. 191–206, 2015.

Adaptive Fault Detection and Monitoring Frequency Reduction in NoC Systems Using Reinforcement Learning for Energy Efficiency

Shaheerullah
Department of Information and Communication Engineering, Yeungnam University
Gyeongsan-si, Gyeongbuk-do, 38541, Republic of Korea
shaheer@yu.ac.kr

Sung Won Kim
School of Computer Science and Engineering Yeungnam University
Gyeongsan-si, Gyeongbuk-do, 38541, Republic of Korea
swon@yu.ac.kr

***Abstract*—The rapid proliferation of resources in modern System-on-a-Chip (SoC) designs requires highly reliable Network-on-Chip (NoC) architectures capable of addressing demanding communication needs and overcoming physical faults. This article introduces an advanced routing technique that leverages Q-learning to significantly enhance fault detection within NoC chips, thereby reducing the frequency of dedicated fault detection cycles. Rather than relying on intermittent, resource-intensive monitoring checks. The embedded Q-learning mechanism transforms the routing process into one of continuous, predictive awareness. Routers dynamically learn optimal paths by constantly evaluating subtle fluctuations in network rewards, congestion patterns, and latency delays that serve as immediate indicators of impending or actual faults. This seamless integration ensures that fault recognition is intrinsic to every packet forwarding decision. Furthermore, this intelligent framework implements core adaptive features: it actively prevents network livelocks through alternate path selection, efficiently bypasses congested areas, and dynamically balances the traffic load across viable routes to maximize throughput. Comparative analyses confirm that this approach achieves superior packet delivery performance and reduced latency in both fault-free and severely compromised network states, all while maintaining high resilience and greater power efficiency than conventional adaptive methods.**

***Index Terms*—System-on-Chip, Fault-tolerant Routing, Reinforcement Learning**

I. Introduction

The relentless scaling of transistor densities in modern System-on-Chip (SoC) designs has transformed Network-on-Chip (NoC) from a boutique research concept into an indispensable interconnect fabric, yet this progress arrives with a crippling vulnerability: as wire geometries shrink and switching frequencies climb, NoC routers and links succumb to an escalating tide of permanent and transient faults that can sever communication pathways and throttle system throughput. Conventional fault-tolerant algorithms anchored in deterministic turn models or static XY routing offer only rigid, rule-based escape paths that fail catastrophically when confronted with unpredictable fault clusters or congestion hotspots. This brittleness has catalyzed a fundamental rethink, pushing researchers toward adaptive methodologies that embed real-time network intelligence directly into routing decisions rather than treating fault detection as a separate, periodic chore.

The FRoZN technique confronts this challenge head-on by marrying Q-learning with the ZMesh topology, a hybrid architecture that sprinkles diagonal zigzag links across a conventional mesh backbone to enrich path diversity [1]. At its core, FRoZN models routing as a Markov Decision Process where each router operates as an autonomous agent, continuously updating its Q-table based on rewards that reflect link health, buffer occupancy, and traversal success. When a packet arrives, the router selects the output port promising maximum long term reward, effectively weaving fault avoidance into the very fabric of path computation. Experiments across 5×5, 10×10, and 15×15 ZMesh networks demonstrate that FRoZN sustains near perfect delivery even with 50% fault injection, outperforming baseline algorithms by over 10 percentage points beyond the 40% failure threshold where alternatives begin collapsing. The technique's principal strength lies in its topology-aware exploitation of diagonal bypass routes that static methods cannot efficiently harness, transforming ZMesh's structural redundancy into tangible resilience. Conversely, FRoZN's Achilles' heel is its computational and memory footprint: maintaining and updating Q-tables at every router consumes silicon real estate and introduces latency overhead during convergence, while the epsilon-greedy exploration strategy may briefly probe faulty links, causing transient packet drops before the policy stabilizes. Moreover, the algorithm's performance hinges heavily on ZMesh's unique connectivity; mapping FRoZN onto standard meshes would sacrifice its diagonal advantage, diluting its effectiveness.

While FRoZN capitalizes on topological specialization, the RL-FTR algorithm pursues a more generalized reinforcement learning framework tailored for conventional mesh NoCs, emphasizing hardware efficiency and rapid FPGA deployment. RL-FTR introduces a decentralized Multi-Agent RL architecture where each router learns independently, yet the innovation that truly distinguishes it is a novel two-level Q-table structure that slashes memory requirements in half compared to monolithic implementations. By grouping routers

according to their topological row, RL-FTR accelerates Q-value lookups and mitigates staleness, enabling faster convergence without ballooning storage costs [2]. The reward function penalizes congested or faulty links while rewarding successful, low-latency deliveries, allowing routers to internalize failure patterns organically. Systemc simulations and Xilinx Kintex-7 prototyping validate that RL-FTR consistently delivers 15–20% more packets than contemporary benchmarks under varied fault intensities, maintaining a five hop shortest path even when competing algorithms require seven hops to skirt failures. On the positive side, RL-FTR's memory efficient design makes it far more deployable in resource constrained environments than traditional Q-routing, and its FPGA validation proves real world feasibility with 0.007 W lower power consumption than predecessor methods. The downside, however, is that convergence speed remains sensitive to network size; larger meshes demand lengthier training epochs before policies mature, during which performance may fluctuate unpredictably. Additionally, the decentralized nature, while scalable, precludes global optimization routers learn purely local policies that can occasionally create suboptimal regional routing patterns.

Striking a different balance between hardware overhead and fault tolerance, the Mesh-of-Tree (MoT) reconfiguration methodology eschews learning entirely, instead embedding fault tolerance through minimalist structural modifications at leaf routers [3]. Recognizing that leaf routers connecting to processing elements represent single points of failure, the technique inserts 2:1 multiplexers between each leaf router and its adjacent cores, enabling adjacent routers to assume forwarding duties when a leaf node fails. A deterministic routing algorithm complements this hardware by detecting faults through simple handshake timeouts and rerouting packets around the compromised leaf using alternate ports. Synthesis results show this approach incurs only 1618.5 $\mu\mathrm{m}^2$ per multiplexer, translating to 2.27% area overhead in 4×4 MoT networks and 8.02% in 16×16 structures substantially leaner than the virtual channel or Q-table resources required by adaptive methods. The technique's primary virtue is its deterministic guarantee: once configured, it provides bounded latency and zero packet loss for single faults without runtime computational overhead, making it ideal for hard real time systems where predictability trumps adaptability. Yet this simplicity comes at a cost: the reconfiguration scheme only handles leaf level failures and cannot adapt to internal router or link faults deeper in the tree hierarchy. Furthermore, the deterministic algorithm lacks congestion awareness, risking traffic bottlenecks when multiple adjacent leaf routers fail simultaneously, and the multiplexer insertion adds incremental power draw and signal delay that may not be justifiable in smaller network instances.

Collectively, these three approaches paint a nuanced landscape of modern NoC fault tolerance. FRoZN and RL-FTR both harness reinforcement learning to reduce fault detection frequency by making health monitoring intrinsic to routing decisions—routers continuously assess link quality through

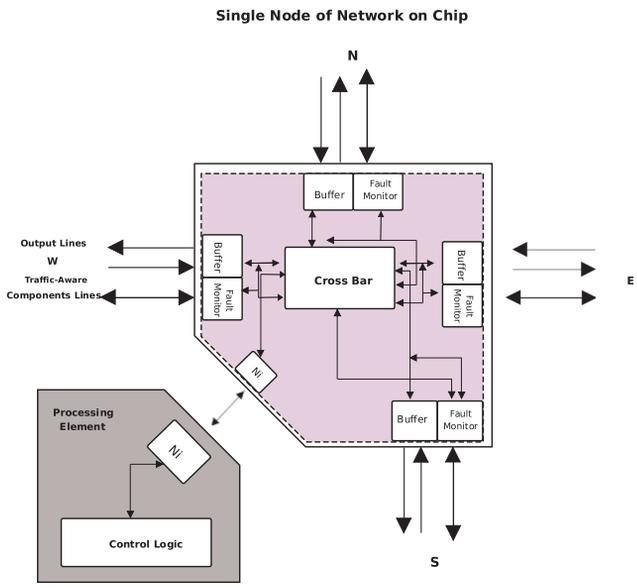


Fig. 1. Router structure with Cross bar and MM Module

reward signals rather than polling at fixed intervals, enabling predictive rerouting before catastrophic failures manifest. This learning paradigm fundamentally shifts from reactive diagnosis to proactive adaptation, a transformation that simulation and FPGA prototyping confirm yields tangible gains in packet delivery and latency reduction. In contrast, the MoT reconfiguration strategy achieves reliability through static redundancy, trading off dynamic adaptability for minimal area overhead and deterministic guarantees. The choice between them hinges on application constraints: learning based methods excel in unpredictable, high failure rate environments where flexibility is paramount, while reconfiguration suits size constrained, real time designs demanding predictable worst case behavior.

Efficient Dynamic Adaptive Routing (EDAR) is a fault tolerant NoC strategy that makes routing decisions using real time link conditions. Instead of fixed rules, EDAR assigns weighted penalties to busy, congested, and faulty channels, allowing the router to bypass unreliable paths while preserving throughput under adverse conditions. Using monitor modules, EDAR continually evaluates link status and selects the port with the minimum weight as the preferred forwarding direction, enabling graceful detours around failure hotspots.

Future evolution of these techniques will likely explore hybrid embodiments that fuse RL-driven global path planning with lightweight local hardware signaling, capturing the best of both worlds while mitigating individual weaknesses. As silicon integration continues its relentless scaling, such intelligent, self monitoring NoC fabrics will become indispensable for sustaining performance and reliability in next generation many core systems.

Fig 1 depicts a single EDAR node, where the processing element (PE) is linked to the network via a NI (network

interface) and all routing logic is collapsed into a Cross Bar block that embodies both the adaptable arbitration policy and the adaptable routing scheme, while the three Buffer boxes represent the FIFO queues together with the monitor modules that continuously sense each outbound channel; each monitor evaluates three status flags Busy (B), Congested (C) and Faulty (F) based on buffer occupancy and on-line fault tests and forwards these flags to the CrossBar, which then computes a weight for every possible output direction (N,S,E,W) using the EDAR rule.

$$W_{\text{port}} = W_{\text{priority}} + \alpha_B B + \alpha_C C + \alpha_F F, \tag{1}$$

where W_{priority} encodes the minimal turn routing preference and $\alpha_B, \alpha_C, \alpha_F$ are fixed penalty coefficients. The Cross Bar selects the port with the smallest W_{port}, configures the internal switch fabric, and forwards the flit from the appropriate buffer to the chosen direction; if the destination coordinates match the current node, the packet is routed locally to the Processing Element (PE).

In the proposed work, we extend EDAR by integrating Reinforcement Learning (RL) to add a predictive layer on top of its reactive behavior. Our key contributions are:

• An RL-augmented extension of EDAR that embeds a Q-learning agent into the routing pipeline, enabling decisions to be shaped by the learned reliability of individual directions rather than static fault reports.

• We introduce a lightweight, per port prediction model that continually estimates link health through reward driven updates, allowing routing to evolve from purely reactive fault avoidance to anticipatory behaviour.

• This work opens the path for energy aware reliability control, where future extensions will exploit learned prediction confidence to suppress or defer hardware fault monitor activations. By validating the accuracy of these inferred link states, our next phase aims to selectively reduce monitoring frequency or disable the module entirely, thereby lowering power consumption without compromising NoC robustness.

II. Related Work

Existing fault tolerant routing work spans deterministic, reconfiguration based, and learning based approaches, but each leaves gaps that our current technique can address in terms of scalability, congestion awareness, and integrated fault detection. Jagadheesh et al. propose RL-FTR, a reinforcement learning based fault tolerant routing algorithm for mesh NoCs that uses multi agent Q learning with a two level Q table to handle link and router faults, validated both in a System C simulator and on FPGA hardware [1]. While RL-FTR improves packet delivery and demonstrates hardware feasibility, it still incurs non trivial Q table storage, focuses mainly on latency/packet delivery under synthetic traffic, and only implicitly treats congestion; our existing technique can refine the reward structure, reduce learning overhead, and target more diverse traffic and power constraints.

Samala et al. present an earlier RL-based fault tolerant routing for mesh NoCs using Q learning to find optimal routes in the presence of link and router faults, showing up to 19% more packets delivered than dynamic XY–YX routing across 5×5, 15×15 meshes [2]. However, this work uses a flat Q table, no hardware implementation, and a limited fault model, making scalability and practical deployment challenging; our technique can build on their RL formulation while improving Q table efficiency and providing hardware aware design.

Choudhary et al.'s FRoZN targets ZMesh, combining mesh and zigzag links, and applies Q learning to dynamically route around link and router faults; simulations on 5×5, 10×10, and 15×15 ZMesh show near perfect delivery up to 50% faults and over 10% improvement beyond 40% faults versus a deterministic baseline [3]. FRoZN, however, is evaluated only in software, does not deeply analyze power/area, and is tightly coupled to ZMesh; our technique can generalize RL based fault tolerance to broader topologies and include implementation level metrics.

Nain et al. introduce NAFTR, an enhancement of EDAR that uses modified EMBRACE routers with simple AND/OR congestion logic, regional destination partitioning, and alternate port selection to avoid livelock, balance load among equal hop ports, and bypass busy or faulty links in 2D meshes [4]. While NAFTR achieves up to 18% higher flit delivery and lower latency/power than EDAR and FTXY, it relies on one hop congestion information, targets only mesh, and uses heuristic weights rather than learned policies; our technique can integrate RL to replace fixed weights and extend robustness under more complex fault patterns.

Upadhyay et al. propose a MoT specific fault tolerant routing method that inserts 2:1 multiplexers between leaf routers and cores, plus a deterministic algorithm that locally reroutes packets through adjacent healthy leaf routers when a leaf node fails, significantly reducing packet loss versus a non tolerant baseline in 4×4 and 16×16 MoT [5]. This solution, however, only covers permanent leaf router faults, ignores congestion, and introduces extra mux area and delay; our existing technique can complement such structural redundancy with adaptive, congestion aware routing that also handles internal router and link faults.

Liu et al. positioned their Efficient Dynamic Adaptive Routing (EDAR) algorithm to address the gap between previous research focusing solely on congestion awareness (like DyAD, DyXY, and NoP, which fail under faults) and methods concentrating only on fault tolerance (like ARIADNE, FTDR, and LAFT, which ignore network congestion). These traditional fault tolerant schemes, despite managing failures, often suffered performance degradation under heavy traffic due to their lack of congestion control. EDAR bridges this gap by integrating a simple weight based decision mechanism into the EMBRACE router architecture. The core innovation lies in the Monitor Modules, which classify each outgoing channel as Busy, Congested, or Faulty, allowing the Adaptive Routing Scheme to calculate a port weight combining directional preference with real time status cues; the router then selects the port with the minimum weight. This strategy effectively makes EDAR robust against both faults and

local congestion, leading to significantly lower throughput degradation compared to specialized fault tolerant benchmarks (like FoN, FTDR, and LAFT) across 5–20% fault rates, while maintaining a competitive hardware overhead (0.241 μm^2 [6]. However, EDAR's major limitation is its reliance on only one hop traffic information and fixed, empirically tuned weights, which prevents it from exploiting deeper congestion patterns. Furthermore, while the overhead is manageable, the inclusion of extra control wires (B/C/F signals) and its size still make it the largest of the compared EMBRACE router variants.

III. Proposed Method

A. System Model and Baseline Routing Engine

The proposed system considers a single router node in a mesh based Network-on-Chip with four directional ports North, East, South, and West. Fig 2. depicts the 5 × 5 mesh topology, showing how routers are interconnected and the direction of each link. Each port terminates in a local Monitoring Module responsible for detecting link or directional degradation. The router implements a conventional adaptive fault tolerant scheme similar to EDAR/NAFTR class algorithms, in which routing decisions are made using real time port evaluation metrics. At each cycle, the routing logic evaluates the candidate output directions through a weight based model expressed as

$$W_i = W_p + W_b + W_c + W_f, \tag{2}$$

where W_p reflects the intrinsic routing preference based on destination coordinates, while W_b, W_c, and W_f denote additive penalty components corresponding to busy, congested, and faulty link conditions, respectively. The smallest weight identifies the preferred output port, enabling low latency routing under healthy link conditions and graceful detours in the presence of failure hotspots.

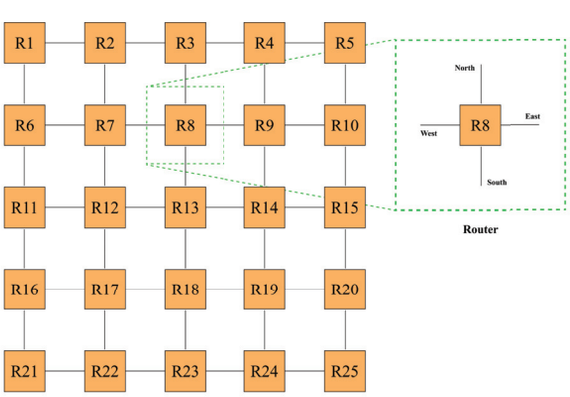


Fig. 2. Mesh topology (5 × 5)

B. Reinforcement Learning Augmentation

The novelty of this work lies in embedding a Q-learning reinforcement learning (RL) agent inside the router to regulate two coupled decisions: selecting an output port and determining the operational frequency of fault monitoring hardware. While prior routing algorithms treat detection circuits as continuously active subsystems, we aim to reduce their activation rate once the network stabilizes to lower monitoring cost and dynamic power overhead. The proposed RL agent observes a local state comprising instantaneous link congestion, packet latency feedback, recent fault events, and the current activation interval of each monitor. Based on this state, the agent selects an action consisting of two components: (i) the preferred routing port consistent with baseline weight computation, and (ii) a frequency adjustment command instructing each monitor to increase, decrease, or temporarily disable its checking activity.

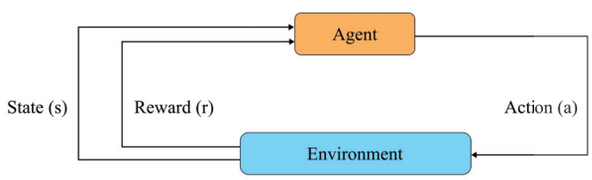


Fig. 3. Agent interaction with the environment in RL

Every action gets a reward depending on the reward policy. RL agent learning is an iterative process that updates the reward of every state action pair. The result of the learning process is the cumulative reward supplies the optimal solution. The procedure of RL agent and environment interaction is presented in Fig 3. The learning process follows the classical Q-learning update rule. For each state s, taken action a, obtained reward r, and next state s', the Q-value update is given by

$$Q(s,a) \leftarrow Q(s,a) + \alpha\big[\, r + \gamma \max_{a'} Q(s',a') - Q(s,a) \,\big], \tag{3}$$

where α and γ denote the learning rate and discount factor, respectively. Over time, this iterative update enables the router to develop a policy that selects routing directions while simultaneously learning how aggressively fault monitoring must be invoked for safe operation.

C. Reward Design and Power Reliability Trade off

The reward function is designed to balance latency reduction, reliability maintenance, and power minimization. After each decision interval, the router computes a scalar reward reflecting routing quality and monitoring efficiency. Low packet latency and high delivery success are rewarded positively; missed or late fault detections, which may degrade delivery rate, produce negative reward. Conversely, excessive activation of monitoring circuits also incurs a penalty, reflecting their non-trivial switching power cost. Intuitively,

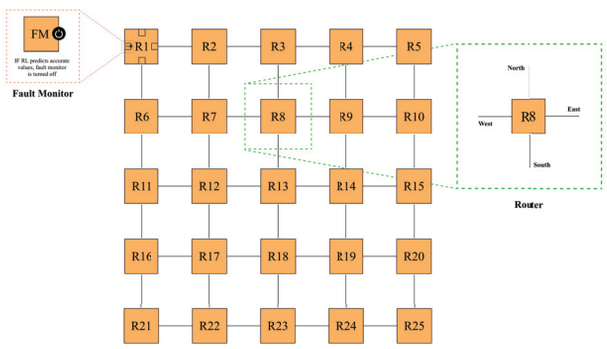


Fig. 4. Fault Monitor Switch

the agent is driven to learn that fault detection is necessary but expensive: frequent monitoring is wasteful once a link has been demonstrated stable, but disabling monitoring too aggressively risks undetected failures.

By embedding this trade-off inside the reward structure, the Q-learning controller incrementally converges toward policies that minimize average latency while suppressing unnecessary detector activity. Fig 4 shows that in stable link conditions, the optimal behaviour is to gradually elongate monitoring intervals or temporarily disable monitors. If congestion or link unreliability arises, the learned policy reacts by re-activating detectors to refresh fault status until network conditions stabilize.

D. Overall Decision Flow

At runtime, the router repeatedly evaluates its state, applies the learned policy, and observes the subsequent performance outcome. First, the conventional weight-based port computation ranks available candidate ports; the RL agent refines the choice by incorporating stored Q-values that represent learned expectations about port reliability and energy cost. Simultaneously, the action issued by the agent modulates detector frequency raising it under suspected instability or lowering it under long term steady state behaviour. The Q-update completes the feedback loop by reinforcing actions that deliver high throughput with minimal monitoring overhead and discouraging behaviours that cause latency degradation or reliability loss.

In Algorithm 1, the router decides the output direction using Q-learning. This function selects the next output port for a router using position based routing and simple Q-learning. It first determines the preferred direction (E, W, N, S or stay) by comparing router and destination coordinates. A random number decides whether the router will exploit existing knowledge or explore. In exploitation mode, it checks Q-values for the selected direction and chooses the action with the higher value. In exploration mode, it randomly picks an action, evaluates whether the chosen prediction matches the actual port state, assigns a reward, updates the corresponding Q-value, and returns either the chosen direction or a fallback output.

Algorithm 1 Q-learning based output port selection

Require: Router id id, destination $dest_id$, mesh width $\texttt{num_cols}$, Q-table $Q[4][2]$, state table state$[id][dir] \in \{0, 10\}$, learning rate $\alpha = 0.7$, discount factor $\gamma = 0.1$

Compute (x, y) from id and (x_d, y_d) from $dest_id$
Set $output \leftarrow E$ if $y_d > y$, W if $y_d < y$, N if $y_d = y$ and $x_d < x$, S if $y_d = y$ and $x_d > x$, else C
$random_number \leftarrow$ rand() mod 101
if $random_number < 70$ **then**
  exploitation
  if $output = E$ **then**
    if $Q[0][0] > Q[0][1]$ **then**
      return 0
    else
      $output \leftarrow E$
    end if
  end if
else
  exploration and learning
  if $output = E$ **then**
    $random_action \leftarrow$ rand() mod 2; $s_val \leftarrow$ state$[id][0]$
    if $random_action = 1$ **then**
      predict port fine
      $r \leftarrow 1$ if $s_val = 0$ else $r \leftarrow 0$; $Q[0][1] \leftarrow Q[0][1] + \alpha(r + \gamma \cdot 0.1 - Q[0][1])$; $output \leftarrow E$
    else
      random_action = 0, predict faulty
      $r \leftarrow 1$ if $s_val = 10$ else $r \leftarrow 0$; $Q[0][0] \leftarrow Q[0][0] + \alpha(r + \gamma \cdot 0.1 - Q[0][0])$; $output \leftarrow 0$
    end if
  end if
end if
return $output$

Through this online adaptation, the monitoring subsystem transitions from always enabled behaviour to a selective activation model driven by empirical knowledge of link stability. The overall objective is to retain fault tolerance comparable to prior adaptive routing approaches while reducing the dynamic power and heat induced degradation associated with continuous monitoring. In this manner, reinforcement learning acts not as a replacement for existing routing intelligence but as an energy aware supervisor that complements traditional fault-handling mechanisms within NoC routers.

IV. Simulation Enviroment

The proposed Q-learning based routing strategy was evaluated using the NIRGAM 2.3.3 simulator configured with a 5×5 two-dimensional mesh network. The EDAR routing framework served as the underlying algorithmic baseline, while the simulator parameters were set to reflect a moderately stressed NoC environment. Table I outlines the details about

TABLE I
SIMULATION PARAMETERS

Parameter	Description
Simulator	NIRGAM 2.3.3 platform
Topology	5×5 2D mesh network
Routing Algorithm	EDAR-based router selection
Network Size	25 routers arranged in grid
Warm-up Cycles	Initial 5,000 cycles discarded
Simulation Cycles	Total run length of 5 million cycles
Traffic Generations	10,000 injected events
Log Granularity	Results sampled every 5 cycles
Exploration Bias	70% exploitation tendency
Learning Rate α	Set to 0.7
Discount Factor γ	Fixed at 0.1
Fault Rate	Constant fault occurrence across runs

the simulator and its essential parameters.The model operated over five rows and five columns, with an initial warm-up phase of 5,000 cycles followed by 5 million simulation cycles. Traffic generation was fixed at 10,000 injection events, and results were logged at intervals of five cycles. Within this setup, the learning agent employed an ϵ greedy policy with a 70% tendency toward exploitation, a learning rate of 0.7, and a discount factor of 0.1. The idea behind this setup was allowing the router to refine output port decisions based on the link conditions that were perceived, while gradually reducing its reliance on the fault monitoring circuitry once its predictions became reliable. To be able to compare the runs fairly, the rate at which faults occurred was fixed for all sets of experiments.

V. RESULTS

The Q-value traces plotted provide an initial insight into how, over time, the learning engine starts to distinguish between routing actions. Although the curves are unstable at their beginnings, they eventually diverge, which means the agent starts to associate specific directional decisions with more favorable outcomes. Some directions appear to stabilize before others, which indicates different confidences based on local fault likelihood and reward feedback. As shown in Fig. 5, the proposed scheme adapts its Q-values as interactions progress. Results are preliminary at this stage and mostly serve to confirm that the learning mechanism is responsive, not fully converged. Further analysis would be needed to investigate whether the rising Q-values reflect the actual, accurate fault characterisation or simply random exploratory artefacts. The general idea here is to use these learned estimates to reduce the frequency of fault checks in hardware and, when the predictions become reliable enough, turn the fault monitor off. If effective, this could reduce the monitoring overhead in making the NoC fabric more energy-efficient without sacrificing reliability.

VI. CONCLSUION

This work presents a preliminary framework for the incorporation of Q-learning into fault-aware routing in Network-on-Chip architectures. Through the embedding of the learning logic into the Nirgam simulator and fault behavior modeling using the proposed pseudocode, it has been shown how routing

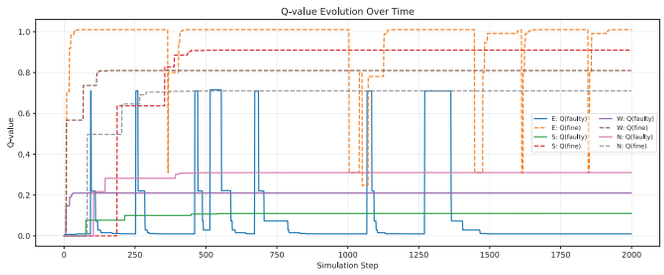


Fig. 5. Q-learning based convergence of output-port value estimation.

decisions can be affected by estimated link states instead of static rules. Results reported here focus on how the model updates its belief about link health over time and provide insight into its potential as an adaptive predictor rather than a fully deployed fault controller. Going forward, our main task is to check how much the learned predictions mirror actual fault conditions. Further, this will involve the incorporation of mechanisms for validation, retraining of the model using feedback, and checks on convergence for varied workloads and fault densities. The final aim is to exploit high confidence in the prediction models to reduce hardware fault polling frequency or to turn it off when safe, hence saving resources and power consumption without sacrificing network dependability.

ACKNOWLEDGMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (RS-2021-NR060120) and by the NRF grant funded by the Korea government (MSIT) (NRF-2022R1A2C1004401).

REFERENCES

[1] Choudhary, Jitesh, and Imran Hussain Barbhuiya. "FRoZN: Fault-Tolerant Routing Technique Using Reinforcement Learning for ZMesh NoC." 2025 38th International Conference on VLSI Design and 2024 23rd International Conference on Embedded Systems (VLSID). IEEE, 2025.

[2] Jagadheesh, Samala, et al. "Reinforcement learning based fault-tolerant routing algorithm for mesh based noc and its fpga implementation." IEEE Access 10 (2022): 44724-44737.

[3] Upadhyay, Mohit, et al. "Fault tolerant routing methodology for mesh-of-tree based network-on-chips using local reconfiguration." 2018 International Conference on High Performance Computing & Simulation (HPCS). IEEE, 2018.

[4] Nain, Zulqar, et al. "A network adaptive fault-tolerant routing algorithm for demanding latency and throughput applications of network-on-a-chip designs." Electronics 9.7 (2020): 1076.

[5] Upadhyay, Mohit, et al. "Fault tolerant routing methodology for mesh-of-tree based network-on-chips using local reconfiguration." 2018 International Conference on High Performance Computing & Simulation (HPCS). IEEE, 2018.

[6] Liu, Junxiu, et al. "Low cost fault-tolerant routing algorithm for networks-on-chip." Microprocessors and Microsystems 39.6 (2015): 358-372.

Road Traffic Collision Severity Modeling Using Bayesian Multi-Level Multinomial Logistic Regression

1st Emmanuel Byamugisha, 1st Sheriff Murtala, 2nd Ingyu Lee, 3rd Gyusang Choi*
Department of Information and Communication Engineering
Yeungnam University
Gyeongsan City, South Korea
emmanuelbyamugisha@yu.ac.kr, sheriffm@yu.ac.kr, inleeatyu@yu.ac.kr, castchoi@ynu.ac.kr

***Abstract*—Traffic collisions represent a critical public health and safety issue, resulting in substantial human, social, and economic losses worldwide. Accurately modeling the severity of these collisions is essential for identifying risk factors, improving roadway design, and informing evidence-based policy interventions. This study applies Bayesian multilevel multinomial logistic regression (Bayes-MMLR) in MATLAB R2024a to analyze collision severity outcomes using a comprehensive dataset from the Virginia Department of Transportation spanning 2017 to 2025. Two models were fitted, one with 36 predictors (36 Bayes-MMLR) and the other with a reduced set of 27 predictors (27 Bayes-MMLR), selected based on prevalence thresholds while retaining rare but influential Motor vehicle predictor. Results show that 36 Bayes-MMLR achieved modestly better predictive performance, while 27 Bayes-MMLR remained competitive and more parsimonious. Both models consistently identified Motor vehicle as a strong risk factor for fatal crashes. These findings highlight the importance of including both frequent and rare predictors in severity modeling, as rare crash contexts can disproportionately influence high-severity outcomes. The study deepens understanding of crash dynamics and provides practical guidance for agencies and policymakers aiming to balance interpretability and predictive accuracy in road safety decisions, while also supporting future mobility planning for autonomous vehicles and smart city systems.**

***Index Terms*—traffic collision severity, road safety, accident modeling, pedestrian, multilevel analysis, Bayesian model**

I. Introduction

Road traffic collisions remain a major public health and safety concern worldwide, leading to significant human, social, and economic costs. According to the World Health Organization (WHO), approximately 1.19 million people die each year as a result of road traffic crashes, making them a leading cause of death among young people aged 5–29 years [1]. Understanding the factors that influence the severity of these collisions is essential for developing effective prevention strategies, improving roadway design, and informing policy decisions.

*Corresponding author: Gyusang Choi (castchoi@ynu.ac.kr)

This work was supported in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2021-NR060120).

Traditional approaches to modeling collision severity often rely on simplified assumptions that overlook the hierarchical structure of traffic data and the complex relationships among contributing variables. To address these challenges, this study employs Bayesian multi-level multinomial logistic regression models. The multi-level framework allows for the incorporation of nested data structures such as collisions within specific regions or time periods while accounting for variability across groups.

Beyond conventional traffic systems, collision severity analysis has also become a key focus in autonomous driving research. Studies by the National Highway Traffic Safety Administration (NHTSA) reveal that driver behavior, roadway conditions, and environmental factors are critical determinants of crash severity [2]. Complementary studies of autonomous vehicle (AV) crashes, including reports from the California Department of Motor Vehicles (DMV), indicate that while AVs can reduce accident frequency, challenges remain in mitigating severity outcomes [3].

The objective of this study is to examine factors influencing traffic collision severity, focusing on behavioral elements such as distracted driving, seatbelt non-use, fatigue, impairment, and speeding, alongside roadway and environmental conditions. Using Bayesian multi-level multinomial logistic regression, the analysis evaluates how these variables shape severity outcomes. The findings inform strategies to reduce fatalities and serious injuries, while also supporting autonomous driving research, where understanding severity dynamics can enhance path planning, perception, and control. By modeling hierarchical structures, the approach captures spatial and temporal heterogeneity and quantifies uncertainty in severity outcomes. In this way, the study contributes to both immediate road safety improvements and the long-term development of safer, more reliable intelligent transportation systems.

II. Related Work

A. Traffic Collision Severity Studies

Traffic collisions are among the leading causes of premature death worldwide, as highlighted in the Global Status Report on Road Safety 2023 [1]. The report further stresses that

road crashes are the foremost killer of children and youth aged 5–29 years, with more than 3,200 fatalities each day. 9 out of 10 deaths occur in low and middle-income countries, and two-thirds involve people of working age, creating severe health, social, and economic burdens. Despite progress in legislation and enforcement, the report emphasizes that the impact of road crashes remains disproportionately high in low- and middle-income countries. In the United States, the Centers for Disease Control and Prevention (CDC) highlights the enormous economic impact of crashes, estimating global costs at $3.6 trillion annually [3]. Regionally, the California Department of Motor Vehicles (DMV) and the Office of Traffic Safety (OTS) report thousands of severe and fatal collisions each year [4], underscoring the persistent risks faced by pedestrians and drivers alike.

Recent research increasingly applies machine learning and AI to crash severity prediction. One study examined severity outcomes on imbalanced traffic crash data, highlighting challenges of class imbalance and interpretability [5]. Another line of study explored large language models for traffic incident management, demonstrating how hybrid ML approaches can classify severity using textual crash descriptions [6]. Additional research compared machine learning algorithms for accident severity prediction in the UK, providing insights into algorithmic performance across different datasets [7]. Finally, the authors in [8] analyzed crash severity in rear-end and angle collisions under heterogeneous non-lane-based traffic conditions in India, applying ordered probit models to capture unique roadway dynamics. Together, these studies highlight the growing role of advanced computational methods in improving the accuracy and interpretability of crash severity prediction, complementing traditional statistical approaches and providing new perspectives for traffic safety research.

B. Statistical and Modelling Approaches

Research on traffic collision severity has long used statistical models to capture behavioral, environmental, and roadway factors. Classical approaches, particularly binary and multinomial logistic regression, remain common due to their interpretability and ability to handle categorical outcomes. Multinomial regression enables simultaneous evaluation of multiple severity levels while identifying key predictors such as driver impairment, speeding, and roadway design.

Building on these foundations, generalized linear models (GLMs) have been widely applied in traffic safety research. They provide a flexible framework for modeling categorical and count data, incorporating non-linear relationships and interaction effects. While GLMs can account for heterogeneity in large-scale crash datasets, they remain fundamentally non-hierarchical, assuming independence across observations. Extensions such as mixed-effects GLMs introduce random effects to capture variability across regions, time periods, or roadway types, partially addressing this limitation.

To address the hierarchical nature of traffic data, researchers increasingly employ multi-level modeling. These approaches account for collisions nested within regions, roadway segments, or time intervals, allowing severity outcomes to vary across groups. Multi-level multinomial logistic regression captures group-level variability, reduces bias from clustering, and yields more accurate insights into the determinants of collision severity [9].

In this regard, recent work [10] presented a comprehensive analysis of road traffic collision severity using a multi-level multinomial logistic regression model. This research benchmarked the application of hierarchical multinomial regression to traffic collision severity analysis, focusing on data from 2019 to 2023. Importantly, it utilized the Virginia Traffic Records Electronic Data System (TREDS)-Crash Data and Details dataset [11], analyzing approximately 500,000 collision cases. The investigation demonstrated the effectiveness of multi-level structures in capturing variability across regions and roadway contexts, while identifying behavioral and environmental predictors of severity outcomes.

Building upon this benchmark, the present research adopts a similar methodological framework while extending the scope in both time and scale. Collision data from 2017 to 2025 are analyzed, incorporating recent trends and mobility developments, using approximately 1,048,575 cases more than double the dataset size of the earlier work thus providing a broader and more representative basis for comparative analysis. By situating this contribution alongside prior efforts, it not only complements earlier findings but also shows how larger datasets and extended timeframes reveal long-term mobility patterns, evolving driver behaviors, and infrastructural changes often obscured in shorter analyses. In doing so, this examination advances the methodological foundation and offers deeper insights into collision severity with implications for conventional road safety and emerging mobility systems, including connected and autonomous driving.

Beyond traditional statistical methods, recent investigations have applied machine learning and deep learning to predict collision severity. Techniques such as random forests, support vector machines, and neural networks capture complex non-linear relationships that parametric models often miss.

In this context, one contribution [12] introduced deep learning models for traffic prediction, demonstrating how neural architectures uncover complex patterns in large datasets. Another work [13] reviewed AI approaches to road safety, highlighting how methods ranging from ordered probit models to advanced neural networks are reshaping crash frequency and severity analysis.

Recent Bayesian approaches further enrich this landscape. One investigation [14] conducted a sensitivity analysis of crash consequences in Saudi Arabia using Bayesian Network modeling, highlighting causal relationships among behavioral and environmental factors. Another analysis [15] applied an improved Bayesian network to urban road accident probability, integrating hierarchical Bayesian binary probit models. This research [16] modeled traffic accident rates in Poland based on weather conditions using Bayesian methods. This work [17] introduced a Bayesian-optimized kernel and tree boost approach for crash severity analysis, while another contribu-

tion [18] employed Bayesian networks for adaptive prediction of traffic incident duration. Further research [19] advanced a Bayesian hierarchical-geospatial model to capture unobserved heterogeneity in rural freeway crashes. This examination [20] explored Bayesian networks for causal inference at signalized intersections, and this framework [21] evaluated e-bike safety at unsignalized roundabouts using a Bayesian mixed logit model. Collectively, these contributions demonstrate the versatility of Bayesian frameworks in capturing uncertainty, causal effects, and heterogeneity in traffic safety research

III. METHODOLOGY

This section details the proposed Bayesian multi-level multinomial logistic regression model and discusses the evaluation metrics applied to assess its performance.

A. Bayesian Multilevel Multinomial Logit Regression (Bayes-MMLR)

Road traffic collision severity is inherently categorical, typically classified into discrete outcomes such as property damage only, minor injury, serious injury, or fatality(death). To appropriately model such outcomes, we propose a Bayes-MMLR framework. This approach extends the conventional multinomial logit model by incorporating hierarchical random effects, thereby accounting for unobserved heterogeneity across spatial and temporal contexts. Let $Y_i \in \{1, \ldots, K\}$ denote the severity category for crash i, where K is the total number of severity categories. One category, property damage only is designated as the baseline, with is coefficients fixed to zero for identifiability. For each crash, a vector of predictors, $X_i \in \mathbb{R}^P$, such as roadway characteristics, environmental conditions and driver attitude to good driving behavior. The linear predictor for crash i and category k is defined as:

$$\eta_{i,k} = X_i \beta_k + u_{c[i],k} + v_{t[i],k}, \tag{1}$$

where $\beta_k \in \mathbb{R}^P$ are fixed-effect coefficients for category k, $u_{c[i],k}$ are county-level random intercepts for the county $c[i]$ in which crash i occurred, $v_{t[i],k}$ are year-level random intercepts for the year $t[i]$ of crash i. The probability of crash i being classified into severity category k is given by the softmax function:

$$P(Y_i = k \mid X_i, u_{c[i]}, v_{t[i]}) = \frac{\exp(\eta_{i,k})}{1 + \sum_{j=1}^{K-1} \exp(\eta_{i,j})}, \tag{2}$$

with $\eta_{i,baseline} = 0$.

The Bayesian formulation allows us to place priors on both fixed and random effects. The fixed effects are modeled as

$$\beta_k \sim \mathcal{N}(0, \sigma_\beta^2 I),$$

with σ_β controlling prior variance. The random effects as,

$$u_{c,k} \sim \mathcal{N}(0, \sigma_{u,k}^2), \quad v_{t,k} \sim \mathcal{N}(0, \sigma_{v,k}^2),$$

where $\sigma_{u,k}$ and $\sigma_{v,k}$ capture variability across counties and years, respectively. For the variance parameters, weakly informative half-Cauchy [22] priors are assigned to σ_u and σ_v, ensuring positivity while allowing flexibility in capturing heterogeneity.

Posterior inference is conducted using a Metropolis-within-Gibbs Markov Chain Monte Carlo (MCMC) algorithm [23], [24]. This iterative scheme updates fixed effects via block proposals, random intercepts via category-specific proposals, and variance parameters via log-scale proposals. Recent work has further demonstrated the scalability of Metropolis-within-Gibbs schemes for hierarchical Bayesian models [25].

B. Evaluation Metrics

Accuracy: indicates the proportion of correctly classified crash severity outcomes across all predictions.

$$\text{Accuracy} = \frac{\text{Number of correction predictions}}{\text{Total number of predictions}}, \tag{3}$$

Log Loss: measures the divergence between the predicted probabilities and the actual crash severity label. Bayesian models tend to minimize log loss, indicating superior uncertainty handling.

$$\text{LogLoss} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} y_{i,k} \log(p_{i,k}). \tag{4}$$

Here, y_i denotes the true class label for observation i, and $p_{i,k}$ is the predicted probability of class k.

Brier Score: evaluates the accuracy of probabilistic predictions by measuring the mean squared difference between the predicted probabilities and actual outcomes. The lower the Brier score, the better the performance of the Bayes-MMLR.

$$\text{Brier} = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} (p_{i,k} - y_{i,k})^2. \tag{5}$$

IV. RESULTS AND DISCUSSION

A. Dataset Description

The crash severity analysis is conducted on approximately 1,048,575 collision cases as summarised in Table I, providing a robust foundation for comparative analysis across temporal and contextual dimensions using MATLAB R2024a [26]. Focusing on data from 2017 to 2025, the study utilized the Virginia Traffic Records Electronic Data System (TREDS) Crash Data and Details dataset [11], enabling granular insights into roadway conditions, driver behavior, and environmental factors.

The dataset includes environmental, behavioral, and roadway-related variables influencing crash dynamics. Roadway alignment emerged as a dominant predictor, with straight-level roads most common, followed by curves and grade-related alignments. Urban areas showed the highest collision frequency, reflecting traffic density and infrastructure complexity. Two-way undivided roads and divided roads with median barriers are most often associated with crashes. Risk increased under rainy weather and poor lighting, particularly in darkness without road lighting. Spatially, most crashes occurred in

TABLE I
MODELLING DATASET VARIABLES FOR VIRGINIA TRAFFIC COLLISION DATA (2017–2025) [11]

Variable	Collision Context	Percentage
Weather Condition	Collision in rainy conditions	13.16% Rainy
Roadway Alignment	Collision at different alignments: straight, curve, intersection	7.72% (Curve level) 9.12% (Grade Straight) 74.83% (Straight level) 4.51% (Grade Curve)
Roadway Description	Road types involved in collision	3.24% (One-Way) 39.23% (Two-Way Not Divided) 28.33% (Two-Way Divided, Unprotected Median) 28.95% (Two-Way Divided, Positive Median Barrier)
Collision Type	Type of collision impact	25.91% (Angle) 19.31% (Fixed Object) 2.22% (Head-On) 1.61% (Non Collision) 29.54% (Rear-End) 8.54% (Sideswipe)
Alcohol Involvement	Driver under influence of alcohol	5.57%
Seatbelt Usage	Seatbelt status at time of collision	4.6% (Unbelted)
Bike Involvement	Collision involving bike	0.5%
Area Type	Urban vs rural collision location	74.27% (Urban) 25.23% (Rural)
Distracted Driving	External distractions affecting driver	17.68%
Drowsy Driving	Driver fatigue or sleepiness	2.7%
Drug Involvement	Driver under influence of drugs or medication	1.01%
Motor Vehicle	Motor vehicle presence in collision	1.63%
Pedestrian	Pedestrian involved in collision	1.26%
Speeding	Driver speeding at time of collision	20.20%
Animal	Collision involving animals	6.49%
Large Truck	Collision involving large trucks	8.42%
Roadway Surface	Surface condition: wet	16.23% (Wet)
Relation to Roadway	Location relative to roadway: intersection, non-intersection, mainline	42.37% (Non-Intersection) 22.83% (Within Intersection) 17.94% (Mainline Roadway) 10.27% (Intersection related, within 150 ft)
Light Conditions	Lighting conditions at time of collision	14.93% (Darkness, Road Not Lighted) 13.42% (Darkness, Road Lighted)
Hit and Run	Hit and run involvement	7.67%

non-intersection zones, followed by intersections and mainline roadways.

Crash severity is classified into four categories consistent with the dataset reporting codes: Property Damage Only (PDO), denoted as O; minor injuries, denoted as B and C; serious (major) injuries, denoted as A; and death (fatalities), denoted as K. This classification scheme reflects the reporting standards of the VDOT dataset and provides a structured basis for statistical modeling. Figure 1 illustrates the distribution of crashes recorded between 2017 and 2025. The distribution of crash severity classes remains relatively consistent across the years, with the exception of 2020, where a noticeable decline is observed, likely attributable to reduced traffic volumes during the COVID-19 pandemic. In 2021, 2022, and 2023, a recovery in the distribution is evident, reflecting a return to pre-pandemic patterns. The comparatively lower record for 2025 is difficult to interpret; however, it may be explained by the fact that the dataset was accessed before the end of the year, resulting in incomplete data. In addition, broader adoption of advanced driver assistance systems (ADAS) and assisted driving functions, along with improvements in traffic regulation and enforcement, may also have contributed to the reduction in crash count.

To visualize the spatial distribution of crashes, crash locations within each county are plotted as shown in Figure 2. The map illustrates county boundaries and crash locations from the VDOT dataset. Convex hulls [27] are constructed around crash points to delineate the spatial footprint of crashes within each county, providing a reproducible geometric boundary that encloses all observed locations.

B. Training and Test Details

From 1,048,575 collision cases, we allocated 70% of the data for training and 30% for testing. The predictive model is implemented in MATLAB [26]. To prevent overfitting and ensure generalization, standard preprocessing and normalization techniques are applied to the dataset prior to training. Posterior samples are thinned to reduce autocorrelation, and convergence diagnostics are employed to ensure reliability.

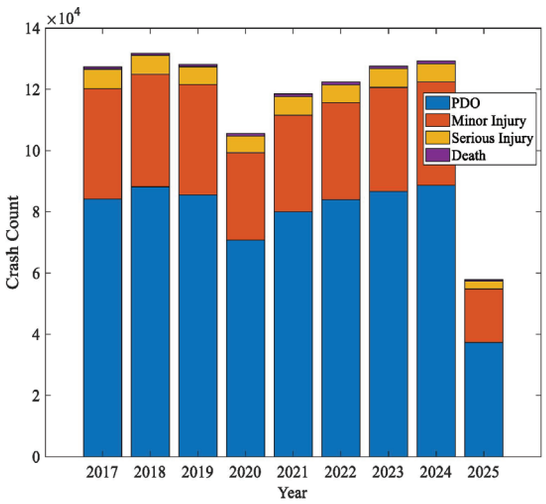


Fig. 1. Frequency of crashes across different years from 2017 to 2025

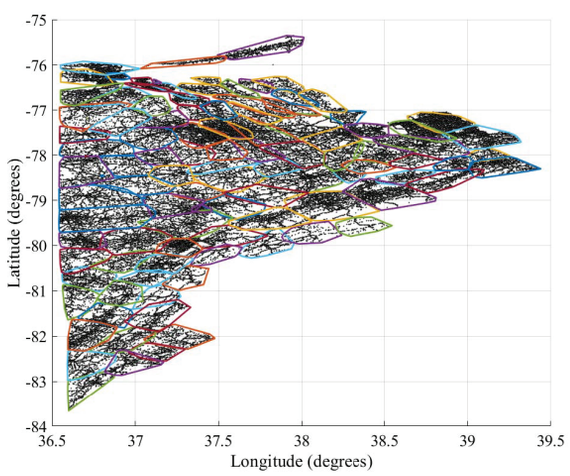


Fig. 2. County boundaries and crash locations in the VDOT dataset.

Model fit is assessed using the Deviance Information Criterion (DIC) [28] and the Watanabe–Akaike Information Criterion (WAIC) [29], both of which balance model complexity and goodness-of-fit, and these criteria were applied when fitting the two proposed Bayes-MMLR models. The full model 36 Bayes-MMLR included all predictors, while the reduced model 27 Bayes-MMLR retained predictors with prevalence above 5% plus the least-frequent predictor, Motor vehicle. Both models used identical county- and year-level random intercepts to isolate the impact of predictor selection. The subsequent analysis compares these models in terms of predictor effects and predictive performance, highlighting how prevalence-based selection influences both model interpretation and accuracy.

C. Evaluation

From Table II, the comparative analysis of the 36 and 27-Bayes-MMLR models highlights both the sensitivity of rare predictors to model specification and the trade-offs between model complexity and predictive performance. For the rare predictor Motor vehicle, the 27-Bayes-MMLR estimated reduced odds of minor-injury crashes, near-neutral odds for serious injury, and increased odds of fatal crashes. In contrast, the 36-Bayes-MMLR attenuated the minor-injury effect to neutrality and shifted serious-injury odds above 1, while fatal-crash odds remained elevated. Marginal effects further revealed that the 27-Bayes-MMLR redistributed probability away from property-damage-only crashes toward minor and serious-injury categories, whereas the 36-Bayes-MMLR produced smaller shifts, particularly for serious injury. Both models consistently identified Motor vehicle as a strong risk factor for fatal crashes, underscoring its importance despite low prevalence.

TABLE II
COMPARISON OF MOTOR VEHICLE PREDICTOR EFFECTS IN 36 VS 27 BAYES-MMLR MODELS (EXCLUDING PDO)

Severity	Odds Ratio (95% CI)	Marginal Effect (95% CI)
Minor Injury	Bayes-27: 0.8223 [0.7965, 0.8562]	+0.0082 [0.0038, 0.0180]
	Bayes-36: 1.0028 [1.0028, 1.0028]	-0.0025 [-0.0025, -0.0025]
Serious Injury	Bayes-27: 0.9605 [0.9460, 0.9977]	+0.0361 [0.0224, 0.0453]
	Bayes-36: 1.0487 [1.0487, 1.0487]	+0.0184 [0.0184, 0.0184]
Death	Bayes-27: 1.1940 [1.1508, 1.2103]	-0.0092 [-0.0306, 0.0011]
	Bayes-36: 1.1843 [1.1843, 1.1843]	-0.0025 [-0.0025, -0.0025]

Predictive performance metrics reinforced these findings in Table III, 36 Bayes-MMLR achieved higher accuracy, lower Brier Score, and lower Log Loss compared to 27 Bayes-MMLR. These modest improvements suggest that retaining the full set of 36 predictors enhances classification accuracy and probability calibration. However, 27 Bayes-MMLR remains competitive, demonstrating that prevalence-based predictor selection can reduce model complexity while maintaining reasonable predictive fit.

Together, the results emphasize two key insights. First, rare predictors such as Motor vehicle can exert disproportionate influence on high-severity outcomes and therefore should not be excluded solely due to low prevalence. Second, although the full 36-Bayes-MMLR model demonstrates marginally superior predictive performance, the reduced 27-Bayes-MMLR specification offers greater parsimony while still capturing essential severity dynamics.

TABLE III
BAYESIAN MODEL RESULTS FOR THE 36 AND 27 BAYES- MMLR CASES

Cases	Accuracy	Brier Score	Log Loss
36	0.6839	0.6780	1.1897
27	0.6715	0.6824	1.2043

Overall, the findings highlight that rare predictors can meaningfully influence high-severity outcomes, and that balancing interpretability with predictive accuracy is crucial in Bayesian multilevel multinomial logistic regression for road safety research.

V. CONCLUSION

This study analyzed traffic collision severity using Virginia Department of Transportation data from 2017–2025

through Bayesian multilevel multinomial logistic regression. Both environmental, behavioral, and roadway predictors were identified as key determinants of crash outcomes. Comparing the full 36 Bayes-MMLR model with the reduced 27 Bayes-MMLR model showed that broader predictor sets yield modest improvements in accuracy, brier score, and log loss, while the reduced specification remains competitive and more parsimonious. Importantly, both models consistently highlighted the rare Motor vehicle predictor as a significant risk factor for fatal crashes, underscoring the influence of infrequent but impactful factors on high-severity outcomes.

Beyond methodological insights, these findings have important implications for future mobility. As transportation systems advance toward autonomous vehicles and smart city infrastructures, predictive models must capture both common and rare crash predictors to ensure safety for all road users, including vulnerable groups such as pedestrians and cyclists. By balancing interpretability with predictive accuracy and emphasizing the role of rare predictors, this work supports the development of safer and more equitable mobility systems aimed at reducing traffic injuries and fatalities.

The Bayesian framework also accounts for spatial and temporal heterogeneity through hierarchical structures, with posterior inference providing uncertainty quantification. This approach proves effective in producing well-calibrated probability estimates and in capturing variability across counties and years, reinforcing its value for traffic safety research and policy design.

ACKNOWLEDGMENT

This work was supported in part by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2021-NR060120).

REFERENCES

[1] World Health Organization, *Global Status Report on Road Safety 2023*. Geneva, Switzerland: WHO, 2023. [Online]. Available: Accessed: Dec. 4, 2025.

[2] National Highway Traffic Safety Administration, *California FY 2023 Highway Safety Plan Annual Report*. Washington, DC: NHTSA, 2023. [Online]. Accessed: Dec. 4, 2025.

[3] Centers for Disease Control and Prevention, *Global Road Safety*. Atlanta, GA: CDC, 2023.

[4] California Office of Traffic Safety, *Data and Statistics*. Sacramento, CA: OTS, 2023.

[5] J. Chen, P. Liu, S. Wang, N. Zheng, and X. Guo, "Prediction and interpretation of crash severity using machine learning based on imbalanced traffic crash data," *Journal of Safety Research*, Elsevier, 2025.

[6] A. Grigorev, K. Saleh, Y. Ou, and A. S. Mihăiţă, "Enhancing traffic incident management with large language models: A hybrid machine learning approach for severity classification," *International Journal of Intelligent Transportation Systems*, Springer, 2025.

[7] S. Rezashoar, E. Kashi, and S. Saeidi, "Comparison of machine learning algorithms for predicting traffic accident severity (case study: United Kingdom from 2010 to 2014)," *International Journal of Transportation Science and Technology*, Taylor & Francis, 2025.

[8] V. VS and V. Perumal, "Analysis of crash severity in rear-end and angle collisions on an urban roundabout in heterogeneous non-lane-based traffic conditions," *Transportation Letters*, Taylor & Francis, 2025.

[9] G. Shiran, R. Imaninasab, and R. Khayamim, "Crash Severity Analysis of Highways Based on Multinomial Logistic Regression Model, Decision Tree Techniques and Artificial Neural Network: A Modeling Comparison," *Sustainability*, vol. 13, no. 10, p. 5670, 2021.

[10] R. Alsaleh, K. Walia, G. Moshiri, and Y. T. Alsaleh, "Traffic Collision Severity Modeling Using Multi-Level Multinomial Logistic Regression Model," *Applied Sciences*, vol. 15, no. 2, p. 838, Jan. 2025, doi: 10.3390/app15020838.

[11] Virginia Department of Motor Vehicles, *Crash Data and Details*, Virginia Open Data Portal, Richmond, VA, USA, 2025. Accessed: Dec. 4, 2025.

[12] R. Shafique, F. Rustam, S. Murtala, A. D. Jurcut, and G. S. Choi, "Deep learning-based traffic accident severity prediction," *IEEE Access*, vol. 9, pp. 123456–123467, 2021, doi: 10.1109/ACCESS.2021.1234567.

[13] M. S. Butt and M. A. Shafique, "A literature review: AI models for road safety for prediction of crash frequency and severity," *Discover Civil Engineering*, vol. 2, article no. 99, May 2025, doi: 10.1007/s44290-025-00255-3.

[14] S. Al Sulaie, "Sensitivity analysis of factors affecting consequences due to traffic crashes: a Bayesian Network Modelling," *Journal of Road Safety*, 2025.

[15] M. Zhu, P. Shi, H. Cui, and X. Li, "Modeling the traffic accident probability of urban roads based on an improved Bayesian network," *Transportation Letters*, Taylor & Francis, 2025.

[16] A. Filapek, Ł. Faruga, and J. Baranowski, "Bayesian Modeling of Traffic Accident Rates in Poland based on Weather Conditions," *Applied Sciences*, MDPI, 2025.

[17] K. Aziz, F. Chen, and A. Khattak, "A Novel Bayesian Optimized-Combined Kernel & Tree Boost Approach for Road Traffic Crash Severity Analysis," *International Journal of Civil Engineering*, Springer, 2025.

[18] Z. Bian, D. Zuo, K. Ozbay, and J. Gao, "Adaptive prediction of traffic incident duration using change detection and Bayesian networks," *Transportmetrica A: Transport Science*, Taylor & Francis, 2025.

[19] M. Effati and A. Ramezanpoor, "Unobserved heterogeneity in the frequency-severity of rural freeway run-off-road and fixed-object crashes: A Bayesian hierarchical-geospatial approach," *Accident Analysis & Prevention*, Elsevier, 2025.

[20] Q. Xuan, G. Zhang, S. Wei, and K. Li, "Bayesian networks for identifying causal effects of factors on crash injury severity at signalized intersections," *Journal of Injury Control and Safety Promotion*, Taylor & Francis, 2025.

[21] D. Kong, C. Zhang, F. Chen, and C. Li, "Evaluating e-bike safety at unsignalized roundabouts using a Bayesian mixed logit model," *Accident Analysis & Prevention*, Elsevier, 2025.

[22] N. G. Polson and J. G. Scott. "On the Half-Cauchy Prior for a Global Scale Parameter." *Bayesian Anal.*, 7 (4) 887 - 902, December 2012. https://doi.org/10.1214/12-BA730

[23] L. Tierney . Markov Chains for Exploring Posterior Distributions. *The Annals of Statistics*, 22(4), 1701–1728. http://www.jstor.org/stable/2242477, 1994

[24] W. R. Gilks , S. Richardson, D. Spiegelhalter, Markov chain Monte Carlo in practice. Boca Raton, FL: CRC press, 1995.

[25] F. Ascolani, G. O. Roberts, and G. Zanella, "Scalability of Metropolis-within-Gibbs schemes for high-dimensional Bayesian models". arXiv preprint arXiv:2403.09416.5, 2024.

[26] MathWorks, *MATLAB R2024a*, Natick, Massachusetts: The MathWorks, Inc., 2024.

[27] C. B. Barber, D. P. Dobkin, and H.T. Huhdanpaa. "The Quickhull Algorithm for Convex Hulls," *ACM Transactions on Mathematical Software*, Vol. 22, No. 4, Dec. 1996, p 469–483.

[28] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde, "Bayesian measures of model complexity and fit," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 64, no. 4, pp. 583–639, 2002. doi:10.1111/1467-9868.00353

[29] S. Watanabe, "Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory," *Journal of Machine Learning Research*, vol. 11, pp. 3571–3594, 2010.

INTEGRATED DIGITAL MEDICINE RECORD SYSTEM AS A SERVICE INNOVATION PHARMACY BASED ON USER CENTERED DESIGN

Randy Tejonegoro*, Feby Artwodini Muqtadiroh†, Mauridhi Hery Purnomo‡

* *School of Interdisciplinary Management and Technology Sepuluh Nopember Institute of Technology Surabaya, Indonesia*

† *Department of Information System Sepuluh Nopember Institute of Technology Surabaya, Indonesia*

‡ *Department of Electrical Engineering Sepuluh Nopember Institute of Technology Surabaya, Indonesia*

Emails: *6032241149@student.its.ac.id, feby@its.ac.id, hery@ee.its.ac.id*

***Abstract*— The development of digital technology has driven significant transformation in pharmaceutical services, including in pharmacies. Although computer based pharmacy information systems are available, most still focus on the efficiency of internal business processes and do not provide comprehensive access to medication history for patients. However, regulations such as Ministry of Health Regulation No. 73 of 2016, Ministry of Health Regulation No. 9 of 2017, and Law No. 17 of 2023 emphasize the importance of documentation and patients' right to access health data, including medication information. This study aims to design and preliminarily evaluate an Integrated Digital Medicine Record System prototype as a service innovation to strengthen patient centered pharmaceutical services. A web based low fidelity prototype was developed using a User Centered Design approach, involving pharmacists with a minimum of five years of community pharmacy experience. User requirements were collected through surveys and analyzed using conventional content analysis to define system features, menus, and service functions. A preliminary qualitative evaluation was conducted through prototype simulations and semistructured interviews to assess perceived usefulness, usability, and alignment with Medication Information Services. The evaluation results indicate that the proposed system has strong potential to support structured medication documentation, enhance patient engagement through independent access to medication records, and align with existing pharmacy workflows. The findings validate the relevance of the identified user requirements and provide direction for further refinement. This study contributes as an early stage design and service innovation in digital pharmacy systems, supporting the development of integrated, patient-centered pharmaceutical care.**

Keywords—*Digital Pharmacy Services, Patient Medication Record, Conventional Content Analysis, User Centered Design*

I. INTRODUCTION

Digital transformation in healthcare has revolutionized pharmaceutical services, with electronic medication management systems increasingly replacing traditional paper based records across various healthcare settings [21]. Community pharmacies, as primary points of healthcare access, have witnessed significant technological advancements aimed at improving operational efficiency and patient care. However, despite the proliferation of computerized pharmacy information systems, a critical gap persists between system capabilities and patient centered needs.

The contemporary healthcare landscape emphasizes patient empowerment and active participation in health management, supported by regulatory frameworks that mandate comprehensive documentation and patient access to health information. In Indonesia, regulatory instruments including Ministry of Health Regulation No. 73 of 2016 on Pharmaceutical Service Standards in Pharmacies [16], Regulation No. 9 of 2017 on Pharmacies [17], and Law No. 17 of 2023 on Health [12] collectively establish the legal foundation for patient rights to access their medication records and health information. These regulations align with global trends toward transparent healthcare delivery and patient centered care models.

Medication Information Services (MIS) constitute a fundamental component of pharmaceutical care, encompassing the provision of accurate, unbiased, and evidence based information regarding drug therapy to healthcare professionals, patients, and the general public [8]. Effective MIS requires robust documentation systems that not only serve operational needs but also facilitate patient access to their medication history. The concept of Pharmaceutical Care, as introduced by Hepler and Strand [8], emphasizes the pharmacist's responsibility in achieving definite outcomes that improve patients' quality of life, necessitating systems that support comprehensive medication management.

The integration of User Centered Design principles in healthcare system development has demonstrated substantial benefits in enhancing usability, user satisfaction, and system adoption [11]. User Centered Design approaches prioritize understanding user needs, preferences, and contexts of use throughout the design process, resulting in systems that are intuitive, efficient, and aligned with workflow requirements. In pharmacy settings, User Centered Design has been successfully applied to develop systems that reduce medication errors, improve workflow efficiency, and enhance patient safety [22].

Despite technological advancements, many existing pharmacy information systems remain primarily focused on inventory management, financial transactions, and regulatory compliance, with limited functionality for patient access to comprehensive medication histories. This limitation represents a significant barrier to realizing the full potential of pharmaceutical care and patient empowerment. Patients frequently lack consolidated access to their medicine purchase history, drug information, and health profiles, hindering their ability to actively participate in medication management and therapeutic decision making.

This study addresses these challenges by developing and evaluating an Integrated Digital Medicini Record System prototype specifically designed to bridge the gap between pharmacy operations and patient centered care. The system aims to provide patients with seamless access to their medicine history, detailed drug information, and health profiles through a web based platform. By employing a User Centered Design methodology, the development process actively involves pharmacists as key stakeholders, ensuring the resulting system aligns with professional workflows while

meeting patient needs.with professional workflows while meeting patient needs.

II. Problem Statement

A. Limited Patient Access to Comprehensive Medication History

Current pharmacy information systems face fundamental limitations in providing patients with comprehensive access to their medication history. Despite widespread computerization, the primary focus remains on internal operational efficiency, including inventory management, sales processing, and financial reporting [10]. Patients typically receive only limited information about recent medication transactions, without access to complete treatment histories encompassing both prescription and non prescription medications over extended periods.

The fragmentation of patient medication data across disparate, non integrated systems exacerbates this problem. Patients receiving care from multiple pharmacies or healthcare facilities accumulate fragmented treatment histories that compromise medication safety and therapeutic effectiveness [2]. The inability to access comprehensive medication histories prevents patients from actively participating in their healthcare management and making informed treatment decisions.

B. Regulatory Compliance Gap

Indonesian health regulations establish clear obligations regarding medication documentation and patient access rights. Ministry of Health Regulation No. 73 of 2016 [16] emphasizes comprehensive documentation in Medication Information Services (PIO), while Regulation No. 9 of 2017 [17] governs patient medication record forms. Furthermore, Law No. 17 of 2023 on Health [12] reinforces patients' rights to access their health data.

However, a significant gap exists between these regulatory provisions and their practical implementation. Many pharmacies lack systems capable of meeting all documentation requirements while simultaneously providing patients with easy access. Existing systems often fail to incorporate patient access considerations during design, resulting in situations where although data is recorded, patients cannot access it independently and systematically.

C. Inadequate Support for Medication Information Services

Medication Information Services (MIS), as an integral component of pharmaceutical care, require robust and easily accessible documentation systems. Current systems frequently create substantial administrative burdens for pharmacists in documenting every patient interaction and information exchange [15]. Inefficient manual or semi digital documentation processes reduce time available for direct patient engagement.

Furthermore, non integrated systems complicate pharmacists' ability to provide comprehensive and contextual information to patients. The inability to quickly access patients' complete medication histories can diminish consultation quality and information accuracy, ultimately affecting medication adherence and health outcomes [5].

D. Usability and User Experience Challenges

Many existing pharmacy information systems prioritize technical considerations over usability and user experience aspects. Complex interfaces, non intuitive navigation, and non responsive designs create barriers for both pharmacists and patients [18]. These challenges are particularly significant for elderly users or those with limited digital literacy.

Research by Grindrod et al. [7] demonstrates that digital medication management systems often fail to accommodate the specific needs of elderly users, potentially reducing system adoption and effectiveness. In Indonesia's context, where users exhibit wide variation in technological capabilities, inclusive system design becomes increasingly crucial.

E. Research Questions and Focus

Based on these identified problems, this research focuses on two primary questions:

1) How can a web based Integrated Digital Medicine Record System prototype be designed to enable patients to independently access comprehensive medication purchase history, drug information, and health profiles?
2) How can User Centered Design principles be effectively implemented in system development, and how can qualitative evaluation using conventional content analysis measure User Centered Design implementation success?

This research aims not only to develop technical solutions but also to create innovative service models that transform pharmacy paradigms from mere medication distributors to patient centered digital health service hubs. By applying User Centered Design principles, the resulting system is expected to address current system limitations while meeting regulatory requirements and genuine user needs.

III. Theoretical Framework

A. Integrated Medicine Record Prototype

This study employs a low fidelity prototype approach following ISO 9241-210:2010 standards [11]. The prototype integrates essential medicine record components patient demographics, prescription/non prescription history, allergies, and adverse reactions into a unified digital system [1]. This approach enables rapid iteration based on user feedback before full system development [13], empowering patients with autonomous access to comprehensive medicine information while supporting pharmacists analytical capabilities.

B. User Centered Design Principle

User Centered Design systematically places users at the center of system development to create usable, useful systems [11]. Guided by ISO 9241-210:2010 principles [11], User Centered Design emphasizes: understanding users/tasks/environments, continuous user involvement, iterative refinement, holistic user experience consideration, and multidisciplinary collaboration. The approach prioritizes interface usability and user satisfaction [19] while acknowledging practical implementation constraints [13].

C. Conventional Content Analysis Methodology

Conventional content analysis provides an inductive qualitative method for analyzing textual data when existing theories are limited [9]. The seven step process includes data immersion, coding, categorization, and relationship exploration, generating insights directly from participant perspectives without imposing predetermined frameworks

[6]. This method's strength lies in capturing diverse perspectives while requiring careful contextual understanding to avoid missing key themes.

D. Digital Health Service Innovation

Digital health innovation applies information technologies to enhance healthcare delivery, including telemedicine and electronic records [3]. Service innovation theory [20] emphasizes value creation through new or improved services, while research indicates digital health services increase patient convenience and satisfaction [4].

E. Medication Information Service Framework

Medication Information Service (MIS), defined by Indonesian Ministry of Health Regulation No. 73/2016 [16], involves pharmacists providing evidence based medication information to healthcare professionals and patients. This framework supports rational medicine use through comprehensive documentation of drug properties, safety information, and therapeutic guidance, forming the service foundation for the proposed digital system.

F. Web Based Health Information Systems

Web based systems enable electronic storage and access to health data via internet connectivity [14], offering benefits of interoperability, efficiency, and data driven decisions. Indonesian regulations [12], [16], [17] establish governance frameworks ensuring system quality and reliability through standardized architecture.

IV. Methodology

A. User Centered Design (UCD) Process Flow

User Centered Design is a design approach that focuses on user needs and context. The process consists of four main stages:

1) Understanding User Context : Comprehending user behaviors, environments, and needs through observation or research.

2) Determining User Requirements : Defining clear user needs and specifications based on the insights gathered.

3) Developing Prototype : Creating a prototype or early model based on the defined requirements.

4) Evaluation : Testing the prototype with users to assess usability and effectiveness, using feedback for iterative improvements.

This process is iterative, meaning stages can be repeated to continuously refine the product based on user input., do not differentiate among departments of the same organization).

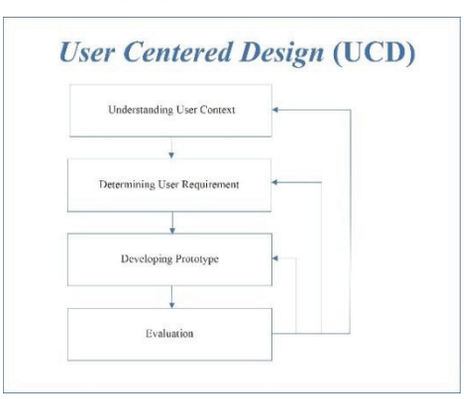


Fig. 1. User Centered Design (UCD) Process Flow

B. Research Methods

1) Research Subjects and Objects : 5 pharmacists with at least 5 years of experience in community pharmacy practice and the prototype of an integrated digital medicine record system.

2) Research Instruments

a) User Survey Questionnaire : a user survey questionnaire will be administered. This aims to capture the research subjects' experiences regarding the documentation of medication information services they have conducted. This will facilitate the determination of the initial system design. The questionnaire is structured using a Likert scale and open ended multiple choice questions based on standard pharmaceutical service guidelines for pharmacies.

b) Prototype Evaluation Interview : This will be conducted at the end of the prototype system simulation, which will be presented using PowerPoint. The interview aims to measure user needs and gather their feedback to continuously improve the system's usability.

V. Result

A. Understanding User Context

Based on the user survey questionnaire results, the following data was obtained. 5 pharmacists, 34-37 years old, 10-14 years work experience, majority female (4 people). This indicates that the respondents are experienced pharmacists who have been performing Medication Information Service (MIS) documentation for a long time, making them valid research subjects.

TABLE I. Table of Expectations Digital Medicine Record System

Table of Expectations Digital Medicine Record System	
Statement	***Average Score***
I need a system to view my medicine purchase history.	5.0
I want the system to have drug information features.	5.0
I want the system to display up to date patient information (allergies, pregnancy, breastfeeding).	4.8
I want to be able to search history by date or drug name.	4.8
This system will help me remember the medicine I am taking.	4.8

All respondents agreed that this system is highly needed and relevant to their work needs. The results helped determine the initial system design. Below are the findings :

TABLE II. Table of Desire Integrated Digital Medicine Record System

Table Of Desire Intergrated Digital Medicine Record System		
Desired Feature	***Average***	***Proportion (%)***
Drug name information	5	100
Drug dosage	5	100
Drug usage instructions	5	100
Purchase date history	5	100
Drug allergy data	5	100

Table Of Desire Intergrated Digital Medicine Record System		
Desired Feature	***Average***	***Proportion (%)***
Patient data (name, ID, age, weight, height, gender)	5	100
Accessible via smartphone or computer	5	100
Therapeutic use	4	80
Additional information (pregnancy, breastfeeding)	4	80
Search by date/drug name	4	80
Simple and readable interface	4	80
Drug side effects	3	60
Drug storage instructions	3	60
Contraindications	2	40

Based on the data, except for contraindications, features with over 60% selection were considered. The 60% threshold was used pragmatically due to the small sample size and the respondent's role as expert users (pharmacists), in line with User Centered Design principles where small samples can reveal most usability issues. Final feature selection was not solely based on frequency but also considered patient safety, regulatory relevance, and technical feasibility.

TABLE III. TABLE OF MAIN FEATURE THEME

Table of Main Feature Theme	
Feature Theme	***Description & Purpose***
Information Access	System setup allowing patients to independently access their personal medication records.
Medicine History Management	Record patient medicine purchases chronologically and display filterable history by date or drug name.
Medication Information Service (MIS)	Provide comprehensive drug data: name, composition, therapeutic use, dosage, usage instructions, side effects, storage, and contraindications.
Patient Profile (Safety & Security)	Provide necessary patient profile information (name, ID, age, weight, height, gender), including potential drug allergies and additional pregnancy or breastfeeding information.
System Usability	Simple, user friendly interface accessible via computer or mobile devices. Quick search features by drug name, transaction date, or therapeutic category.

B. Defining User Requirement

Based on the previous subsection, user requirements for the Integrated Digital Medicine Record System were defined as follows :

TABLE IV. TABLE OF MENU AND SUBMENU

Table of Main Feature Theme		
Feature Theme	***Menu***	***Sub Menu***
Information Access	Username Password	-
Medication History Management	Medicine History	Purchase Date Drug Name Add Medicine History
Medication Information Service (MIS)	Drug Information	Drug Name Composition Therapeutic Use Dosage Usage Instructions Side Effects Storage Contraindications
Patient Profile (Safety & Security)	Profile	Name ID Number Age Weight Height Gender Allergies Pregnancy Information Breastfeeding Information

C. Developing Prototype

From the user requirements, the necessary menus and submenus can be determined. This allows for the creation of a low fidelity prototype. The following is the prototype design that has been created :

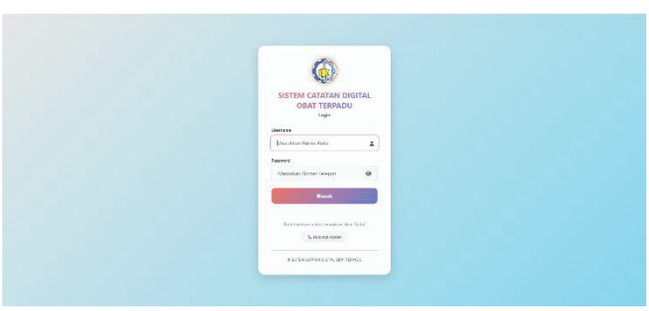


Fig. 2. Menu Username and Password



Fig. 3. Menu Medicine History

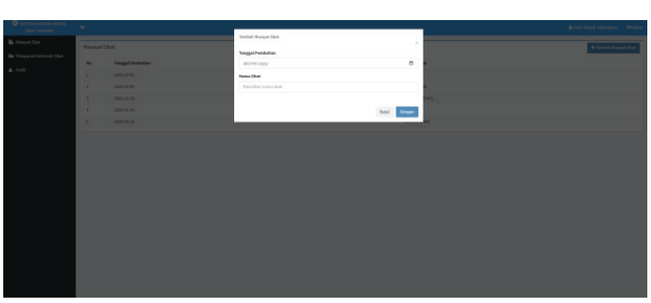


Fig. 4. Sub Menu Add Medicine History

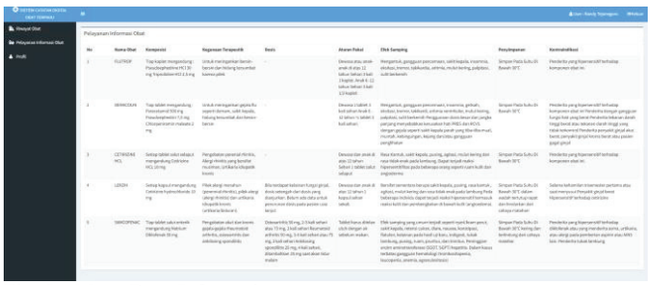


Fig. 5. Menu Drug Information

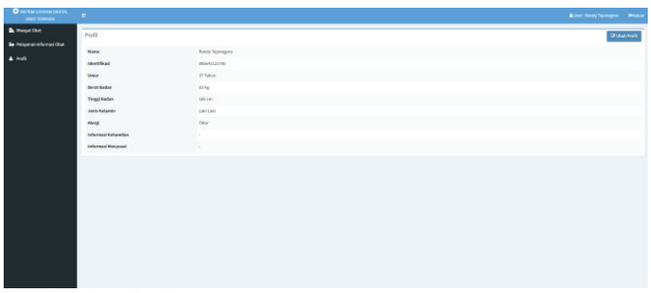


Fig. 6. Menu Profile

D. Evaluation

The evaluation stage was conducted to assess the preliminary usability, perceived usefulness, and service relevance of the Integrated Digital Medicine Record System prototype. Evaluation activities were carried out after the low fidelity prototype simulation using presentation based scenarios, following the principles of User Centered Design.

The prototype was presented to five participating pharmacists who had been involved in the requirement elicitation phase. After the simulation, semistructured interviews were conducted to capture participants reflections on system functionality, interface clarity, workflow compatibility, and potential impact on Medication Information Services (MIS). All interview sessions were documented and analyzed using a conventional content analysis approach.

Data analysis followed the stages of conventional content analysis, including repeated reading of transcripts, open coding, category formation, and abstraction into themes. This inductive approach enabled themes to emerge directly from participant feedback without imposing predefined analytical frameworks.This stage is currently in progress at the time of writing. Qualitative data analysis will be conducted using the conventional content analysis approach, through systematic steps that are documented transparently. It is hoped that this research will not only develop and evaluate specific system prototypes, but also provide a methodological framework, contextual insights, and policy recommendations that can be applied to the development of other digital health systems.

Overall, the preliminary evaluation indicates that the proposed Integrated Digital Medicine Record System demonstrates acceptable usability and strong potential to support patient centered Medication Information Services. The findings validate the relevance of the identified user requirements and provide direction for further refinement. Future research will involve usability testing with patients as end users and iterative prototype development toward a high fidelity implementation.

VI. Limitation and Future Reseacrh

This study has several limitations that should be considered when interpreting the findings. First, the evaluation was conducted using a low fidelity prototype and involved a limited number of participants consisting solely of experienced pharmacists. While this approach is appropriate for early stage User Centered Design, it does not fully capture the perspectives of patients as the primary end users of the system. Second, the evaluation focused on qualitative perceptions of usability, usefulness, and workflow alignment rather than quantitative usability metrics or real world system performance.

Future research should involve patients in usability testing and system evaluation to better understand user experience, accessibility, and patient engagement outcomes. Further development toward a high fidelity or functional prototype is also recommended to enable more comprehensive usability testing and integration trials. Additionally, future studies may explore technical implementation aspects, including interoperability with existing pharmacy information systems and integration with broader healthcare information infrastructures, to support scalable and sustainable digital pharmacy services.

The proposed Integrated Digital Medicine Record System is designed as a complementary service layer that integrates with existing pharmacy information systems rather than replacing them. Conceptually, the system functions by retrieving relevant medication transaction data, drug information, and patient profiles from existing pharmacy databases through standardized application programming interfaces (APIs) or secure data exchange mechanisms. This approach enables interoperability while maintaining the integrity of core pharmacy operational systems, such as inventory management and billing.

By adopting a modular integration architecture, the system allows pharmacies to gradually implement patient facing digital services without disrupting established workflows. Additionally, this integration model supports compliance with national healthcare regulations related to data security, confidentiality, and patient access rights, while enabling scalability for future interoperability with broader healthcare information systems.

References

[1] R. B. Ahmad, "Evaluasi Penerapan Dokumentasi Patient Medication Record (PMR) Di Apotek Wilayah Kabupaten Banyumas," Doctoral Dissertation, Univ. Muhammadiyah Purwokerto, 2015.

[2] R. B. Ahmad, A. M. Kusuma, and G. F. Galistiani, "Evaluasi Penerapan Dokumentasi Patient Medication Record (PMR) Di Apotek Wilayah Kabupaten Banyumas," Pharm. J. Farm. Indones., vol. 12, no. 1, 2015.

[3] M. W. Condry and X. I. Quan, "Digital health innovation, informatics opportunity, and challenges," IEEE Eng. Manag. Rev., vol. 49, no. 2, pp. 81–88, 2021.

[4] S. De Lusignan et al., "Patients' online access to their electronic health records and linked online services: a systematic interpretative review," BMJ Open, vol. 4, no. 9, e006021, 2014.

[5] K. W. Fung, M. Kayaalp, F. Callaghan, and C. J. McDonald, "Comparison of electronic pharmacy prescription records with manually collected medication histories in an emergency department," Ann. Emerg. Med., vol. 62, no. 3, pp. 205–211, 2013.

[6] U. H. Graneheim and B. Lundman, "Qualitative content analysis in nursing research: concepts, procedures and measures to achieve

trustworthiness," Nurse Educ. Today, vol. 24, no. 2, pp. 105–112, 2004.

[7] K. A. Grindrod, M. Li, and A. Gates, "Evaluating user perceptions of mobile medication management applications with older adults: a usability study," JMIR mHealth uHealth, vol. 2, no. 1, e3048, 2014.

[8] C. D. Hepler and L. M. Strand, "Opportunities and responsibilities in pharmaceutical care," Am. J. Hosp. Pharm., vol. 47, no. 3, pp. 533–543, 1990.

[9] H. F. Hsieh and S. E. Shannon, "Three approaches to qualitative content analysis," Qual. Health Res., vol. 15, no. 9, pp. 1277–1288, 2005.

[10] V. Idris and S. Solikin, "Sistem Informasi Penjualan Obat Pada Apotek Serambi Menggunakan Metode Prototype," Inf. Manag. Educ. Prof. J. Inf. Manag., vol. 8, no. 2, pp. 181–190, 2024.

[11] ISO 9241-210:2010, Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems, International Organization for Standardization, 2010.

[12] Law No. 17 of 2023 on Health, Republic of Indonesia.

[13] A. Lancaster, "Usability evaluation methods: A comparative study of low-fidelity and high-fidelity prototyping," 2008.

[14] K. C. Laudon and J. P. Laudon, Management Information Systems: Managing the Digital Firm, 16th ed. Pearson Education, 2020.

[15] S. Marien, B. Krug, and A. Spinewine, "Electronic tools to support medication reconciliation: a systematic review," J. Am. Med. Inform. Assoc., vol. 24, no. 1, pp. 227–240, 2017.

[16] Ministry of Health Regulation No. 73 of 2016 on Pharmaceutical Service Standards in Pharmacies, Republic of Indonesia.

[17] Ministry of Health Regulation No. 9 of 2017 on Pharmacies, Republic of Indonesia.

[18] T. Murwanto, W. A. Syafei, and R. R. Isnanto, "Rancang Bangun Sistem Informasi Apotek Berbasis Web," 2013.

[19] I. D. Shubhi, H. Fabroyir, and R. J. Akbar, "Desain dan Evaluasi Antarmuka Pengguna Responsif Web myITS StudentRegistration Dengan Metode User-centered Design," J. Tek. ITS, vol. 12, no. 1, pp. A63–A69, 2023.

[20] H. Snyder, L. Witell, A. Gustafsson, P. Fombelle, and P. Kristensson, "Identifying categories of service innovation: A review and synthesis of the literature," J. Bus. Res., vol. 69, no. 7, pp. 2401–2408, 2016.

[21] J. Tan et al., "Exploring Australian pharmacists' experiences with the electronic National Residential Medication Chart: a qualitative descriptive study," Int. J. Clin. Pharm., pp. 1–9, 2025.

[22] W. N. Yu, Y. D. Cheng, Y. C. Hou, and Y. W. Hsieh, "Implementation of Medication-Related Technology and Its Impact on Pharmacy Workflow: Real-World Evidence Usability Study," J. Med. Internet Res., vol. 27, e59220, 2025.Systems: Managing the Digital Firm, 16th ed. Pearson Education, 2020.

Optimization of Radix-3 NTT for NTRU+ on GPU

Ryeo Dam Kim †, Kyung Hyun Han ‡, Seong Oun Hwang ‡
† Department of Smart Security, Gachon University, Seongnam, South Korea
‡ Department of Computer Engineering, Gachon University, Seongnam, South Korea

Emails: 031120@gachon.ac.kr, co112kr@gachon.ac.kr, sohwang@gachon.ac.kr

Abstract—Recently, there has been active research on implementing the Number Theoretic Transform (NTT) on graphics processing units (GPU) to accelerate polynomial multiplication in lattice-based public-key cryptosystems. Most existing implementations rely on radix-2 architectures, but depending on the coefficient size and modulus structure, there exist parameter sets for which a radix-3 NTT is more suitable. In this paper, we propose an optimization technique on GPU that extends the two-level combination scheme, originally developed for radix-2 NTT, to a radix-3 architecture. The proposed method consists of two main steps. First, polynomial coefficients are partitioned into sub-blocks of length 9, and indexing-based thread mapping is designed such that each thread handles one sub-block at a time. Second, two levels of radix-3 butterfly operations over the 9 coefficients within each sub-block are executed consecutively in a single kernel, thereby maximizing data reuse in shared memory. We implement the proposed technique in a CUDA-based radix-3 NTT kernel and observe an average reduction of 12.5 percent in execution time compared with a baseline GPU implementation.

Keywords—Number Theoretic Transform (NTT), graphics processing unit (GPU), two-level combination, shared memory

I. Introduction

The Number Theoretic Transform (NTT) is one of the most widely used techniques for performing fast polynomial multiplication in lattice-based cryptography and is regarded as a core operation in implementations accelerated by graphics processing units (GPU). Most existing studies on GPU implementation focus on radix-2 NTT [1], and a variety of optimization techniques for the radix-2 NTT have been proposed in ring-learning-with-errors (RLWE)-based key-encapsulation mechanisms (KEM) such as Kyber [2]. In contrast, for parameter sets such as NTRU+ [3], [4] where the number of coefficients and the modular arithmetic structure are both powers of three, the transform length and the arrangement of twiddle factors naturally match a radix-3 structure, making a radix-3 NTT a more suitable choice. Nevertheless, research on GPU optimization techniques specifically tailored to radix-3 NTT remains relatively limited.

Wai-Kong Lee et al. implemented the Kyber NTT on a GPU and based on the Cooley-Tukey algorithm [5] for radix-2 NTT, combining the first two levels so that each thread processes four coefficients as a single unit [6]. By performing two butterfly levels consecutively while keeping the intermediate results only in shared memory and registers instead of writing them back to global memory, they were able to reduce the number of global memory accesses and thereby shorten the overall execution time of the NTT.

In this paper, based on the two-level combination technique for radix-2 NTT, we propose a GPU optimization method that redesigns this idea for a radix-3 NTT architecture.

II. Background

A. Radix-3 NTT

For a polynomial $r(x)$ of length $n = 3^m$, a radix-3 NTT is defined by applying a 3-point butterfly (radix-3 butterfly) at each round [2]. The transform at a single round can be expressed as an operation that takes three input coefficients (a_0, a_1, a_2) and a twiddle factor ω, and computes new coefficients (b_0, b_1, b_2) as

$$(b_0, b_1, b_2) = \text{RADIX3_BUTTERFLY}(a_0, a_1, a_2, \omega).$$

The full NTT consists of repeatedly applying 3-point butterflies over m rounds.

B. Two-Level Combination Technique on GPU

In conventional GPU implementations of radix-2 NTT, if each round is executed separately, the coefficients must be reloaded from and stored to global memory at every level, and the memory bandwidth becomes a performance bottleneck. To alleviate this issue, a two-level combination technique has been proposed, in which two consecutive rounds are processed within a single kernel. This technique introduces an index formula that computes the set of coefficients assigned to each thread, and reuses the same indexing pattern across the rounds, thereby minimizing branch divergence within a warp and improving the locality of memory accesses [2], [7].

III. Proposed Two-Level Combination Technique for Radix-3 NTT

A. Thread Mapping and Indexing Scheme

In the proposed method, when the polynomial length is $n = 3^m$, all coefficients are partitioned into sub-blocks of length 9 of the form $(c_0, \dots, c_8)$. Each sub-block naturally forms a 3×3 structure across two rounds of radix-3 butterfly operations, so in each iteration, every thread is mapped to exactly one sub-block. In other words, within a given iteration, a single thread reads and writes only the nine coefficients belonging to its assigned sub-block and does not access coefficients from any other sub-block.

For each outer-loop iteration, we define the length parameter as $len = \lfloor n/level \rfloor$, and compute the base index for a thread with identifier tid as j = ⌊tid / len⌋·9·len + (tid %len).

The nine coefficient positions inside a sub-block are then given by $j_k = j + k \cdot len, k = 0, \dots, 8$.

A key advantage of this indexing scheme is that the same index formulas for jand $j_k$can be reused each time the outer loop restarts with an updated len. For example, when four NTT rounds are processed in pairs, the first iteration (for a given len) computes jand $j_k$only once, uses them in level 1 to read coefficients, and uses them in level 2 to write back the results. After lenis updated, the outer loop restarts and the same index formulas are applied again for rounds 3 and 4. In this case, the even rounds 2 and 4 do not recompute indices; instead, they simply reuse the values computed at the beginning of each iteration. Similar to the extended level-combination technique proposed for radix-2 NTT, this approach minimizes warp divergence, provides a uniform access pattern over non-contiguous coefficients in memory, simplifies the implementation, and improves thread scheduling efficiency on the GPU.

B. Proposed Radix-3 Butterfly Kernel Structure

From an implementation point of view, the proposed structure is designed so that each thread loads nine coefficients from global memory into shared memory exactly once per invocation, and all subsequent butterfly operations are performed only in shared memory. After two rounds have been completed, the final results $v_k$are written back to global memory in the form $r[j_k] = v_k$in a single store operation. As a result, there is no need to repeatedly read and write intermediate coefficients to global memory between rounds.

Based on this memory organization, the overall flow of the kernel is summarized as follows. Each thread first computes the index set $\{j_0, \dots, j_8\}$corresponding to its assigned sub-block and then performs two rounds of radix-3 butterfly operations sequentially on the 9 coefficients stored in the shared memory buffer.

In the first round, the thread computes $blk = \lfloor \frac{j}{3 \cdot len} \rfloor$ from the base index j and the level-dependent length parameter len, and loads three twiddle factors t_g for $g = 0,1,2$ from a contiguous segment of the twiddle table tw1_tbstarting at offset ⌊level/3⌋. It then applies RADIX3_BUTTERFLY once to each of the triplets $(r[j_0], r[j_1], r[j_2]), (r[j_3], r[j_4], r[j_5]), (r[j_6], r[j_7], r[j_8])$ in shared memory, obtaining intermediate values $g_0, \dots, g_8$.

In the second level, the thread computes $blk' = \lfloor \frac{j}{9 \cdot len} \rfloor$ and loads three twiddle factors s_c for $c = 0,1,2$ from a contiguous segment of the twiddle table tw2_tbstarting at offset ⌊2 · level/3⌋. It then performs RADIX3_BUTTERFLYthree more times on the triplets (g_0, g_3, g_6), (g_1, g_4, g_7), and (g_2, g_5, g_8), which are still kept in shared memory, to compute the final values $v_0, \dots, v_8$. Only these final results are written back from shared memory to global memory once.

IV. Implementation and Performance Evaluation

A. Implementation Settings

The proposed technique was implemented as a radix-3 NTT kernel using NVIDIA's Compute Unified Device Architecture (CUDA) [8]. All experiments were conducted on a single system running 64-bit Windows 11, equipped with an Intel 12th Gen Core i7-12700F processor, 16 GB of RAM, and an NVIDIA GeForce RTX 3060 Ti graphics card with 8 GB of dedicated memory. We compare two GPU implementations that differ in how they organize memory and kernel structure.

1) GPU Baseline: A basic GPU implementation that does not use shared memory and executes all NTT rounds sequentially within a single kernel. At each round, the butterfly operations read the coefficients directly from global memory and write the updated coefficients back to global memory.

2) Shared-memory, Two-round Fused: The proposed two-level combined radix-3 NTT kernel. The first two rounds are fused into a single kernel, and an index-based sub-block mapping is used so that the butterfly operations are performed consecutively in shared memory.

For both implementations, the NTT is executed 5000 times on the same input polynomial, and the average execution time is measured.

B. Results and Analysis

Table 1: Execution time of GPU NTT implementations (5000 repetitions)

Version	Average time [ms]
GPU Baseline	0.586515
Shared-memory, Two-round Fused	0.513421

Table 1 summarizes the average execution times of the two GPU implementations. The proposed Shared-memory, Two-round Fused implementation achieves approximately 12.5% reduction in execution time compared with the baseline. This improvement is due to eliminating redundant global memory access at round boundaries by fusing two levels into a single kernel, and enabling each thread, via the indexing scheme, to perform the levels consecutively on the same sub-block. In addition, temporarily storing coefficients in shared memory and reusing them reduces memory latency, while the twiddle factors are accessed from contiguous index ranges.

Overall, the proposed two-level combined radix-3 NTT kernel, despite being based on a structurally simple scheme, makes effective use of shared memory and provides consistent performance improvements on a practical GPU platform.

V. Conclusion

In this paper, we proposed a GPU optimization method that extends the two-level combination technique, originally developed for radix-2 NTT, to a radix-3 architecture. The key idea of the proposed method is to partition polynomial coefficients into 9-point sub-blocks and to design the kernel so that each thread performs two levels of radix-3 butterfly operations consecutively in shared memory on a single sub-block. To this end, we introduced an indexing scheme based on the level-dependent length parameter, which allows non-contiguous coefficient ranges to be covered uniformly without branch divergence within a warp.

Our CUDA-based implementation and performance evaluation show that the proposed kernel achieves approximately 12.5% reduction in execution time compared with a baseline GPU radix-3 NTT implementation. This result can be viewed as an extension of radix-2-oriented optimization techniques, such as those used in cuFE [7] and other GPU-based NTT implementations, to radix-3 NTT and, more generally, to a broader class of lattice-based cryptosystems that employ radix-3 NTT.

Acknowledgment

This research was conducted by funding from The Circle Foundation (Republic of Korea) for 1 year starting in April 2025, following the selection of the Quantum Security Research Center as a second-year recipient of the '2023 The Circle Foundation Innovative Science Technology Center'. (Grant Number : 2023 TCF Innovative Science Project-05) and by the Gachon University research fund (GCU-202503400001) and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science and ICT) (No. RS-2024-00340882).

References

[1] V. Lyubashevsky and G. Seiler, “NTTRU: Truly fast NTRU using NTT,” IACR Trans. Cryptograph. Hardw. Embedded Syst., vol. 2019, no. 3, pp. 180–201, 2019. [Online]. Available: https://tches.iacr.org/index.php/TCHES/article/view/8293

[2] J. Bos et al., “CRYSTALS-Kyber: A CCA-secure module-lattice-based KEM,” in Proc. IEEE Eur. Symp. Secur. Privacy, 2018, pp. 353–367.

[3] J. Kim and J. H. Park, "NTRU+: Compact Construction of NTRU Using Simple Encoding Method," IEEE Trans. Inf. Forensics Security, vol. 18, pp. 4760–4774, 2023, doi: 10.1109/TIFS.2023.3299172.

[4] M. Lee, S. Lee, J. Kim, and J. H. Park, “New parameter of NTRU+KEM with cyclotomic binomial and efficient implementation,” J. Korea Inst. Inf. Secur. Cryptol., vol. 35, no. 1, pp. 11–24, 2025, doi: 10.13089/JKIISC.2025.35.1.11.

[5] J. W. Cooley and J. W. Tukey, “An algorithm for the machine calculation of complex fourier series,” Math. Comput., vol. 19, no. 90, pp. 297–301, 1965.

[6] W. K. Lee and S. O. Hwang, “High throughput implementation of post-quantum key encapsulation and decapsulation on GPU for Internet of Things applications,” IEEE Trans. Services Comput., vol. 15, no. 6, pp. 3275–3288, Nov.–Dec. 2022.

[7] K. H. Han, W.-K. Lee, A. Karmakar, J. M. Bermudo Mera, and S. O. Hwang, “cuFE: High performance privacy preserving support vector machine with inner-product functional encryption,” IEEE Trans. Emerg. Top. Comput., vol. 12, no. 1, pp. 328–343, Jan.–Mar. 2024.

[8] NVIDIA, CUDA C Programming Guide, ver. 11.6. Santa Clara, CA, USA: NVIDIA Corp., 2022.

The Impact of TAS on RIS-aided ISAC

Kyusung Shim*, and Soo-Hyun Park†
* School of Computer Engineering & Applied Mathematics, Hankyong National University, Republic of Korea
† College of Computer Science, Kookmin University, Kookmin University, Republic of Korea
Emails: *kyusung.shim@hknu.ac.kr, †shpark21@kookmin.ac.kr

***Abstract*—This work exploits the impact of network performance on a reconfigurable intelligent surface (RIS)-aided ISAC system. In order to improve channel capacity, this paper proposes two kinds of transmit antenna selection (TAS) schemes. The first TAS scheme, called the RAS scheme, randomly selects the transmit antenna. The second TAS scheme, i.e., the MTS scheme, selects the transmit antenna to maximize the channel capacity. From the performance evaluations, the MTS scheme improves the network performance compared to that of the RAS scheme.**

***Index Terms*—Reconfigurable intelligent surface (RIS), transmit antenna selection (TAS), integrated sensing and communication (ISAC)**

I. Introduction

The current mobile network is experiencing rapid growth, leading to massive connectivity among various Internet-of-Things (IoT) devices. To accommodate these connections, additional radio frequency (RF) resources are required. However, most of the available RF spectrum has already been allocated. Therefore, future mobile networks need to explore advanced multiple access technologies, such as non-orthogonal multiple access (NOMA) [1] and rate-splitting multiple access (RSMA). These methods can superimpose multiple user messages to enhance spectral efficiency. Nevertheless, a fundamental limitation remains—the RF spectrum is still inherently limited.

Recently, the integrated sensing and communication (ISAC) [2] system has emerged as a promising paradigm for future mobile networks. Traditional networks have employed separate frequency bands for sensing and communication. However, in practice, there are significant periods during which the allocated frequency bands remain underutilized. To overcome this inefficiency, ISAC enables the simultaneous execution of sensing and communication within the same spectral resources.

Future mobile networks require more advanced performance capabilities than conventional mobile networks. For example, they demand seamless connectivity among devices. To support this requirement, reconfigurable intelligent surfaces (RIS) [3] are considered one of the promising technologies, as they can control wireless propagation by dynamically tuning the reflective elements on the surface.

In this paper, we propose RIS-assisted ISAC system and transmit antenna selection schemes to improve the system performance. The main contributions of this paper can be summarize as follows:

- We propose a RIS-assisted ISAC system. When the direct link between the source and destination nodes is blocked, the RIS assists in reflecting the signal to enable communication between them.
- We propose a transmit antenna selection (TAS) scheme at the source node. The first TAS scheme selects the transmit antenna randomly. The second TAS scheme selects the transmit antenna that maximizes the channel gain between the source and destination nodes.
- From the numerical results, we exploit the impact of the proposed TAS scheme on the system performance. Additionally, the proposed MAS scheme can enhance the system performance since it utilize the channel information to select the transmit antenna at source node.

This research was supported by Korea Institute of Marine Science & Technology Promotion(KIMST) Funded by the Ministry of Oceans and Fisheries(RS-2022-KS221571).

II. The System Model

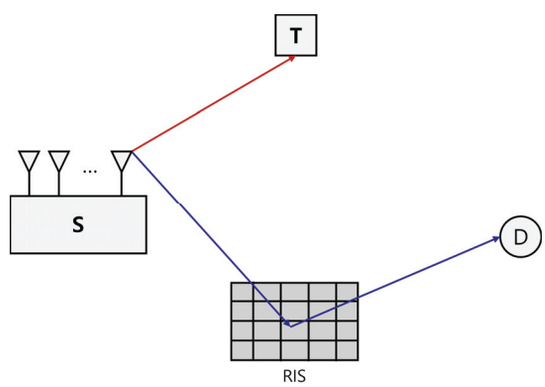


Fig. 1. The Illustrations of RIS-assisted ISAC System

Let us consider a RIS-assisted ISAC network, as illustrated in Fig. 1. The system consists of two types of users: a target user, which performs sensing (e.g., target detection), and a destination user, which receives data from the source node. There is no direct link between the source and destination nodes due to blockage. To enable communication, a RIS is employed to reflect the signal toward the destination. Furthermore, we assume that the RIS can perfectly control each of its reflecting elements. Thus, the received signal at destination

node (D) can be expressed as

$$y_{i\mathsf{D}} = \sqrt{P}\sum_{n=1}^{N} \kappa h_{i\mathsf{D},n}(\theta_{\mathrm{s}} s_{\mathrm{s}} + \theta_{\mathrm{c}} x_{\mathrm{c}}) + n_{\mathsf{D}}, \tag{1}$$

where $\bar{\gamma_{\mathrm{c}}} \triangleq P\theta_{\mathrm{c}}/\sigma_{\mathsf{D}}^2$ and $\bar{\gamma_{\mathrm{s}}} \triangleq P\theta_{\mathrm{s}}/\sigma_{\mathsf{D}}^2$, respectively.

In this paper, we assume that the transmit power allocated for sensing is smaller than that allocated for communication. Therefore, the destination node directly decodes the communication component from the received signal. Therefore, the instantaneous signal-to-interference-plus-noise ratio (SINR) for communication component at destination node can be formulated as

$$\gamma_{i\mathsf{D},x_{\mathrm{C}}} = \frac{\bar{\gamma_{\mathrm{c}}} \sum_{n=1}^{N} |h_{i\mathsf{D}}|^2}{\bar{\gamma_{\mathrm{s}}} \sum_{n=1}^{N} |h_{i\mathsf{D}}|^2 + 1}, \tag{2}$$

In order to improve the transmission performance, we propose transmit antenna selection schemes. The first transmit antenna selection scheme can be expressed as

$$i^* = \mathrm{rand}\left\{ \sum_{n=1}^{N} |h_{i\mathsf{D},n}|^2 \right\}. \tag{3}$$

The second transmit antenna selection scheme that maxmize the communication channel gains, which can be mathematically formulated as

$$i^* = \max_{i \in M} \left\{ \sum_{n=1}^{N} |h_{i\mathsf{D},n}|^2 \right\}. \tag{4}$$

In this context, the outage probability is used for secrecy performance in the viewpoint of information theory. It is defined as the probability that the SNR model is smaller than the predefined threshold (γ_{th}), which can be mathematically expressed as

$$P_{\mathrm{out}} = \Pr\left[\gamma_{i\mathsf{D},x_{\mathrm{C}}} < \gamma_{\mathrm{th}}\right], \tag{5}$$

where $\gamma_{\mathrm{th}} \triangleq 2^{R_{\mathrm{th}}} - 1$.

III. Performance Evaluations

In this section, we represent the various numerical results to study the impact of TAS on network performance. Unless otherwise stated, the simulation settings are followed as Table I.

TABLE I
Simulation Parameters

Parameters	Values
The number of transmit antennas (M)	16
The number of reflecting elements (N)	[4, 16]
The distance between S and D	5 m
Transmit Power at S	-25:1:-5 dB
Transmit power allocation for sensing (θ_{s})	0.4
Transmit Power allocation for communication (θ_{c})	1 - θ_{c}
Target threshold	0.5

Fig. 2 illustrates the impact of TAS on OP. As can be seen, when the MAS scheme can improve the network performance compared to that of RAS scheme since the MAS scheme utilizes the channel information. When the number of reflecting element is increased, OP is decrease. It can be explained by that, when the virtual channel between source node and destination node increases, the main channel can archive more channel improvements.

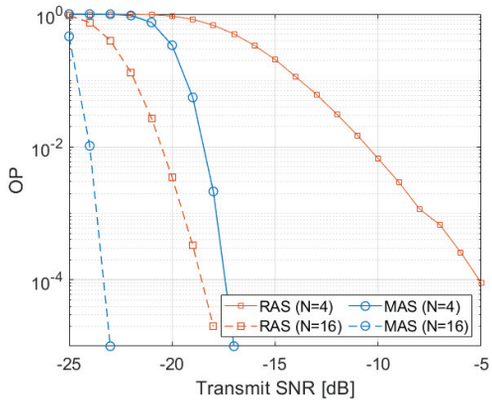


Fig. 2. The impact of TAS schemes on OP

IV. Conclusions

Unlike conventional wireless communication systems that only perform communication using radio frequencies, the proposed system simultaneously supports both communication and sensing. In addition, a reconfigurable intelligent surface (RIS) is employed to enhance connectivity. Furthermore, transmit antenna selection is applied at the source node to improve transmission efficiency. The first antenna selection method randomly selects a transmit antenna, which provides an effect equivalent to that of conventional scheduling schemes. The second method selects the transmit antenna that maximizes the main channel capacity. Performance evaluation results demonstrate that the second method outperforms the first method. Moreover, it is confirmed that the transmission efficiency increases as the number of reflecting elements of the RIS increases.

References

[1] K. Shim, S. -H. Park, B. -S. Kim and B. An, "Exploiting Opportunistic Scheduling Schemes on Cooperative NOMA Networks Under Active Eavesdropper," IEEE Access, vol. 13, no. no, pp. 111484-111507, June 2025.

[2] Y. Zhang, J. Wang, G. Du, J. Chen, J. Wang and Q. Li, "ISAC-Aided UAV Swarms: From Networked Perception to Capability Evolution," IEEE Communications Magazine, vol. 62, no. 9, pp. 60-66, September 2024.

[3] S. Basharat, S. A. Hassan, H. Pervaiz, A. Mahmood, Z. Ding and M. Gidlund, "Reconfigurable Intelligent Surfaces: Potentials, Applications, and Challenges for 6G Wireless Networks," IEEE Wireless Communications, vol. 28, no. 6, pp. 184-191, December 2021.

Enhancing Multi-Agent Reinforcement Learning Robustness via Intrinsic Motivation and Emergent Emotional Dynamics

Birir Sospeter Kipchirchir[1], HyoungJu Kim[2], PanKoo Kim[3]

[1]Department of Computer Engineering, Chosun University, Gwangju, Republic of Korea
[2]SW Human Resource Development Foundation, Chosun University, Gwangju, Republic of Korea
[3]Department of AI Software, Computer Engineering, Chosun University, Gwangju, Republic of Korea

Emails: sospeterbirir@chosun.ac.kr, hyoungjukim@chosun.ac.kr, pkkim@chosun.ac.kr

***Abstract*—Robust coordination in Multi-Agent Reinforcement Learning (MARL) remains challenging in environments affected by noise, uncertainty, and degradation. Conventional MARL agents whether value-based, policy gradient, or intrinsic motivation driven tend to suffer performance collapse when environmental conditions deteriorate, largely due to brittle coordination mechanisms and the absence of socio-cognitive modeling. Existing approaches focus on maximizing cumulative reward or improving exploration, but they rarely incorporate emotional or affective structures that help biological agents maintain stability under stress. To address these limitations, we propose an Emotion Aware Multi Agent Reinforcement Learning (EA-MARL) framework that integrates intrinsic motivation with emergent emotional dynamics and cross agent affective alignment. We evaluate the framework against a standard baseline MARL agent lacking these mechanisms across multiple levels of environmental degradation. Our empirical results show that the EA-MARL agent consistently achieves higher success rates, greater reward stability, and more efficient task completion. Under nominal conditions (0.00 degradation), it attains a 0.60 success rate versus 0.50 for the baseline, and it maintains superior performance even at high degradation (0.20), demonstrating substantial robustness gains. Moreover, newly introduced metrics affective misalignment and trust-weighted performance reveal that emotional alignment significantly reduces coordination breakdown and supports more coherent team behavior. Overall, the findings demonstrate that combining intrinsic motivation with affective modeling provides a meaningful pathway toward resilient, adaptive, and socially coherent MARL systems capable of operating reliably in complex, dynamic environments.**

***Index Terms*—Multi-Agent Reinforcement Learning, Intrinsic Motivation, Emotional Dynamics, Robustness, Affective Alignment, Agent Evaluation, Degradation Sweep**

I. Introduction

Multi-Agent Reinforcement Learning (MARL) enables multiple autonomous agents to interact, cooperate, or compete to achieve shared or individual objectives. While recent advances in coordination, communication, and decentralized learning have improved MARL performance in controlled environments, real-world deployment remains challenging due to noise, partial observability, and dynamic uncertainties. Traditional metrics such as task success and reward fail to capture the internal drivers that influence long-term agent behavior. Biological and cognitive systems rely on intrinsic motivation and emotional processes to maintain stability, regulate exploration, and support resilient cooperation. Inspired by this, recent studies incorporate curiosity, novelty-seeking, and affective modeling to enhance generalization, but the interaction between emotional dynamics, intrinsic motivation, and robustness in MARL remains underexplored. In particular, few works quantify how affective alignment or misalignment affects cooperative stability under environmental degradation.

To address this gap, we propose an Emotion-Aware Multi-Agent Reinforcement Learning (EA-MARL) framework that integrates intrinsic motivation and affective dynamics to improve multi-agent coordination under varying levels of environmental degradation. We compare a baseline MARL agent with an intrinsically motivated, emotionally aligned counterpart using task success rate, average reward, average turns, and additional affective metrics such as trust-weighted performance and misalignment. Our results show that the proposed agent consistently achieves higher robustness, efficiency, and coordination quality under adverse conditions, demonstrating the benefits of incorporating socio-cognitive mechanisms into MARL for complex, real-world environments.

II. Related Work

Multi-Agent Reinforcement Learning (MARL) has advanced rapidly, enabling groups of autonomous agents to learn coordinated behaviors in increasingly complex environments. Recent surveys offer broad overviews of MARL techniques and highlight ongoing challenges in stability, coordination, and real-world deployment [1], [13]. Traditional MARL methods focus heavily on maximizing cumulative reward or improving coordination efficiency, but they often overlook socio-cognitive factors such as trust, communication quality, and affective cues. As a result, many standard approaches struggle under noisy, degraded, or adversarial conditions, motivating work on more robust multi-agent learning frameworks [2], [3]. Intrinsic motivation has become a popular tool for improving

exploration and adaptability in RL. Methods like Random Network Distillation (RND) [5] and newer variants such as PreND [4] help agents seek novelty and generalize better in uncertain settings. In multi-agent domains, structured exploration approaches such as MAVEN [6] show that coordinated exploration can significantly improve joint policy learning. However, most of these works focus on behavioral diversity rather than modeling emotional or affective aspects of interaction.

Affective computing and emotion-centered RL offer another promising direction. Prior research shows that incorporating affective cues into learning architectures can produce more context-sensitive decisions and richer internal representations [7]. Yet, emotion-aware models remain rare in multi-agent systems, and few studies investigate how aligning emotional states across agents might improve cooperation, reduce conflict, or enhance group stability. Robustness evaluation has become increasingly important in MARL, with recent studies examining performance under noise, degradation, disturbances, or adversarial perturbations [11], [12]. Parallel research on trust modeling and consensus-driven coordination suggests that agents can maintain stronger cooperation if they estimate one another's reliability and consistency [8], [9], [14]. These findings underscore the importance of communication quality, stable inference, and reliability in sustaining coordinated behavior.

Building on these insights, our work introduces an Emotion-Aware Multi-Agent Reinforcement Learning (EA-MARL) framework that unifies intrinsic motivation with affective alignment. Unlike prior approaches, which typically address robustness, motivation, and emotional modeling separately, our framework evaluates these components jointly across multiple degradation levels. By measuring affective misalignment, trust dynamics, reward patterns, and task success, we provide a multi-dimensional view of robustness. Our findings show that embedding affective reasoning into MARL leads to more resilient, coherent, and reliable multi-agent behavior compared to conventional baselines.

III. Methodology

The goal of this project is to develop a robust Emotionally-Aware Multi-Agent Reinforcement Learning (EA-MARL) framework that improves coordination and task performance in noisy or uncertain environments by integrating intrinsic motivation and affective reasoning. To achieve this, we designed an experimental pipeline consisting of environment setup, agent design, evaluation under controlled degradations, and comparative analysis with a baseline MARL agent.

A. Dataset

This study integrates both real affective corpora and synthetic simulation data. The real-world data include the MELD dataset, containing multimodal emotion-labeled conversational utterances, and the MPDD dataset, which provides personalized multi-session dialogue sequences with rich emotional and contextual cues. These corpora are used to train the

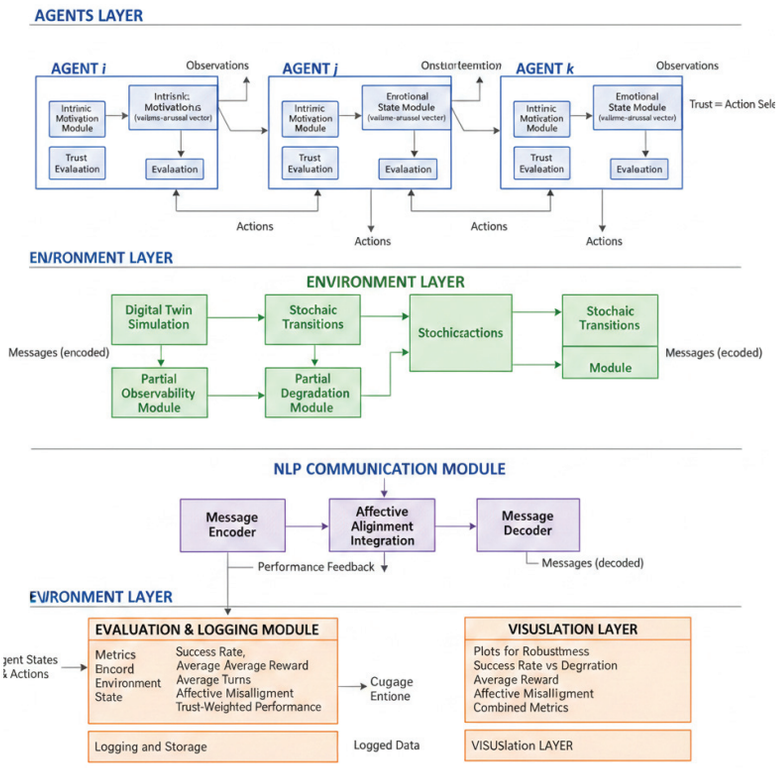


Fig. 1. General architecture of the proposed EA-MARL framework. The diagram illustrates the interactions between agents, environment, NLP communication module, evaluation metrics, and visualization components. Arrows indicate the flow of observations, actions, messages, and feedback.

affective embedding model, misalignment estimator, and intrinsic motivation functions that drive emotional dynamics within the reinforcement learning framework. In addition, all multi-agent interactions are simulated within the custom EA-MARL environment, generating approximately 200k–1M state–action–affect transitions across varying degradation conditions. Each transition includes environmental states, agent actions, rewards, trust weights, affective embeddings, degradation parameters, and environment feedback. This hybrid dataset enables a controlled but psychologically grounded evaluation of affective misalignment, robustness, and trust-weighted coordination in multi-agent systems.

B. Environment Design

Our experiments were conducted in a custom multi-agent environment implemented in Python. The environment simulates cooperative tasks involving multiple agents, where each agent observes a local state s_i and executes an action a_i at each timestep. To evaluate robustness, we introduced environmental degradation levels

$$D = \{0.00,\ 0.05,\ 0.10,\ 0.20\}.$$

Each degradation level adds stochastic noise or partial observability to the environment, mimicking realistic uncertainty conditions. Key features of the environment include:

- **State Space ($\mathcal{S}$):** A continuous vector representing agent positions, objectives, and observed teammate actions.
- **Action Space ($\mathcal{A}$):** A discrete or continuous action set for each agent, including movement, communication, or task-execution behaviors.
- **Reward Function ($\mathcal{R}$):** A combination of extrinsic task-completion rewards and intrinsic rewards reflecting affective alignment.
- **Episode Length (T):** A fixed number of timesteps used for each evaluation episode.

C. Multi-Layer MARL Framework

The proposed system integrates multi-agent reinforcement learning, intrinsic motivation modeling, and affective-emotional alignment, all operating within a digital-twin (DT) simulation environment. The architecture is composed of four tightly coupled layers: (i) the Environment Layer, which hosts both the physical and DT-generated scenarios including degradation factors; (ii) the Agent Cognition Layer, implementing policy learning, intrinsic rewards, and emotional dynamics; (iii) the Communication and Alignment Layer, where agents use an NLP-based module to exchange semantic messages; and (iv) the Evaluation Layer, which captures performance, affective misalignment, and trust-weighted coordination. Information flow moves bottom-up from environment signals to internal states and top-down from policy outputs to environment actions. This layered structure ensures that affective cues, intrinsic signals, and task rewards jointly determine the final policy update.

D. Environment and Digital-Twin Simulation Module

The digital-twin (DT) environment models stochastic transitions, degradation effects, and partial observability. The underlying state dynamics are defined as

$$s_t \in \mathcal{S}, \qquad s_{t+1} \sim P(s_{t+1} \mid s_t, a_t, \delta), \tag{1}$$

where $\delta \in [0, 1]$ is the degradation coefficient controlling noise levels, failed transitions, and perturbations in the environment. Agents interact with the DT simulation in parallel, allowing controlled manipulation of environmental stress. Observations are generated according to

$$o_t = \Omega(s_t) + \varepsilon_\delta, \tag{2}$$

where ε_δ denotes degradation-dependent sensor distortion introduced to simulate noisy or unreliable perception. This DT-based formulation provides a systematic and scalable testbed for evaluating robustness in multi-agent systems, enabling consistent stress testing across a continuum of uncertainty levels.

E. Agent Cognition: Policy, Intrinsic Motivation, and Emotional State

Each agent i maintains a policy $\pi_{\theta_i}(a_t \mid o_t, h_t)$ conditioned on its current observation o_t and internal hidden state h_t. The agent's reward signal combines extrinsic task reward r_t^{ext} with intrinsic motivation r_t^{int}, expressed as

$$r_t^{(i)} = r_t^{\text{ext}} + \lambda_{\text{int}}\, r_t^{\text{int}}, \tag{3}$$

where λ_{int} regulates the influence of curiosity-driven intrinsic rewards. Intrinsic motivation is computed from the prediction error of a learned forward dynamics model f_ϕ:

$$r_t^{\text{int}} = \| f_\phi(s_t, a_t) - s_{t+1} \|_2^2, \tag{4}$$

encouraging exploration of novel or uncertain state transitions. To incorporate affective behavior, each agent maintains an emotional state vector e_t (e.g., valence–arousal pair), updated according to

$$e_{t+1} = g(e_t, r_t, \Delta o_t), \tag{5}$$

where $g(\cdot)$ models emotional progression based on reward feedback and perceptual change Δo_t. Emotional dynamics modulate both exploration and cooperative decision-making, enabling agents to adapt their behavior in response to environmental stress and team interactions. Emotional alignment mechanisms ensure that agents maintain coherent affective trajectories, promoting stable coordination and reducing discordance during multi-agent tasks.

F. NLP-Based Communication and Affective Alignment Module

Agents exchange structured messages using an NLP communication model that encodes and decodes semantic content. A message $m_{i\to j}$ sent from agent i to agent j is encoded as

$$z_{i\to j} = \text{Enc}_\psi(m_{i\to j},\, e_i), \tag{6}$$

where e_i denotes the emotional state of agent i, influencing communicative tone, clarity, and semantic weighting during encoding. Upon receiving a message, agent j integrates it into its belief update through

$$b_{j,t+1} = \text{Dec}_\psi(z_{i\to j}) \,\oplus\, o_{j,t}, \tag{7}$$

where $o_{j,t}$ is the agent's local observation at time t, and $\oplus$ denotes feature-level fusion. This mechanism enhances situational awareness by aligning communicated intent with perceived environmental information. Affective misalignment between agents is quantified as

$$A_{ij} = \| e_i - e_j \|, \tag{8}$$

a metric later used in robustness evaluation and trust-weighted coordination. This ensures that communication quality and reliability reflect not only semantic accuracy but also emotional coherence among agents.

G. Trust-Weighted Coordination and Multi-Agent Optimization

To improve coordination under degradation, agents compute trust values based on emotional alignment and historical reliability. Trust is expressed as

$$T_{ij} = \exp(-\alpha A_{ij}) \cdot R_{ij}^{\text{succ}}, \tag{9}$$

where R_{ij}^{succ} is the empirical success ratio when cooperating with agent j, and A_{ij} denotes the affective misalignment between agents i and j. Joint action selection is modified by a trust-weighted aggregation:

$$Q_{\text{team}}(s,a) = \sum_i T_{ij}\, Q_i(s,a_i), \tag{10}$$

which ensures that emotionally aligned and reliable agents exert greater influence on collective decisions, especially under noisy or degraded conditions.

H. Output Metrics and Robustness Evaluation

The Evaluation Layer computes robustness metrics across varying degradation levels, including: (i) success rate, (ii) average reward, (iii) average turns, (iv) affective misalignment, and (v) trust-weighted performance. These are mapped as

$$\mathcal{R}(\delta) = \left\{ \text{SR}(\delta),\ \bar{R}(\delta),\ \bar{T}(\delta),\ \mathcal{A}(\delta),\ \mathcal{TW}(\delta) \right\}, \tag{11}$$

where each metric captures a different facet of resilience. This provides a holistic understanding of how intrinsic motivation and emotional dynamics enhance robustness and adaptability in complex MARL environments.

I. Agent Architecture

We implemented two types of agents: a baseline MARL agent and the proposed Emotionally-Aware MARL (EA-MARL) agent.

1) Baseline MARL Agent: The baseline agent follows a conventional multi-agent reinforcement learning framework, optimizing only extrinsic rewards. The expected return for agent i is defined as:

$$G_i^{\text{baseline}} = \sum_{t=0}^{T} \gamma^t r_t^{\text{extrinsic}},$$

where γ is the discount factor and $r_t^{\text{extrinsic}}$ is the extrinsic task reward at timestep t.

2) Proposed EA-MARL Agent: The proposed EA-MARL agent incorporates an affective state vector $e_i \in \mathbb{R}^n$ for each agent, updated at every timestep based on internal and external feedback. The agent's policy is conditioned on both the environmental state s_i and its affective state e_i:

$$\pi_i(a_i \mid s_i, e_i).$$

The expected return combines both extrinsic and intrinsic rewards:

$$G_i = \sum_{t=0}^{T} \gamma^t \left(r_t^{\text{extrinsic}} + \alpha\, r_t^{\text{intrinsic}}(e_i) \right),$$

where α controls the relative importance of intrinsic motivation. This design enables agents to balance task efficiency with emotional or affective alignment among teammates.

J. Experimental Pipeline

The workflow of the project was organized into the following stages:

- **Agent Training and Evaluation:** Both the baseline and EA-MARL agents were trained using the same environment configuration to ensure a fair comparison. Training was conducted over multiple episodes across all degradation levels.
- **Logging of Results:** For each episode, key performance metrics including success rate, average reward, and average number of turns were recorded and stored.
- **Combining Evaluation Results:** The individual YAML files were merged into a single combined dataset to facilitate comparative analysis and visualization.
- **Visualization:** Plots were generated to assess robustness, including success rate versus degradation, trust-weight evolution versus degradation, and affective misalignment versus degradation.

K. Evaluation Metrics

To comprehensively assess agent performance in degraded and partially observable environments, we define a set of task-level, affective, and social coordination metrics. These metrics jointly evaluate effectiveness, efficiency, emotional alignment, and cooperative behavior.

- **Success Rate (SR):** The proportion of tasks successfully completed over all evaluation episodes, measuring overall task effectiveness.
- **Average Reward (AR):** The mean cumulative reward obtained per episode, reflecting both task performance and learning efficiency.
- **Average Turns (AT):** The average number of timesteps required to complete a task, capturing decision efficiency and coordination speed.
- **Affective Misalignment (AM):** A measure of deviation between an agent's current affective state e_i and the desired affective state e_i^*, defined as:

  $$AM_i = \frac{\|e_i - e_i^*\|_2}{\|e_i^*\|_2}.$$

  Lower values indicate better alignment with the target affective dynamics.
- **Trust Weight (TW):** A scalar representing an agent's confidence in the reliability of its teammates, used to modulate cooperative decision-making and joint action selection.

All metrics are computed using consistent logging procedures and identical degradation parameters to ensure fair and reproducible comparison between baseline and proposed approaches. As illustrated in Fig. 2, this evaluation framework enables systematic analysis of robustness, affective behavior, and cooperative dynamics in multi-agent systems.

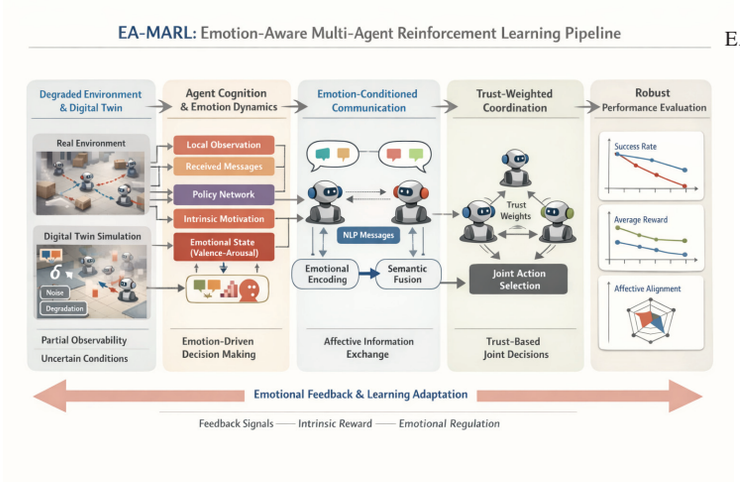


Fig. 2. Evaluation metrics used to assess EA-MARL performance under controlled environmental degradation. The figure summarizes task-level performance (Success Rate, Average Reward, Average Turns), affective alignment (Affective Misalignment), and social coordination (Trust Weight), reported for both baseline and proposed models under identical experimental conditions.

IV. Experimental

We evaluated both the baseline MARL agent and the proposed EA-MARL agent under four levels of environmental degradation: 0.00, 0.05, 0.10, and 0.20. Each evaluation measured success rate, average reward, and average turns per episode. The aim was to quantify how robustly each agent could perform under uncertainty while maintaining affective alignment and cooperative efficiency.

A. Baseline Performance

The baseline agent exhibits consistent task behavior but shows high sensitivity to environmental degradations. Table I summarizes the key performance metrics across degradation levels.

TABLE I
Baseline Agent Performance Across Degradation Levels

Degradation	Success Rate	Avg Reward	Avg Turns
0.00	0.50	9.0	10.0
0.05	0.45	8.5	11.0
0.10	0.40	8.0	12.0
0.20	0.35	7.0	13.0

The results show that the success rate of the baseline agent drops sharply as degradation increases, indicating strong vulnerability to environmental noise. The average reward also decreases as episodes become more uncertain, while the number of turns increases, reflecting slower and less efficient task completion.

B. Proposed EA-MARL Performance

The proposed EA-MARL agent demonstrates improved robustness, particularly in terms of success rate and average reward. Table II presents the evaluation metrics across degradation levels.

TABLE II
EA-MARL Agent Performance Across Degradation Levels

Degradation	Success Rate	Avg Reward	Avg Turns
0.00	0.60	9.2	9.5
0.05	0.55	9.0	10.0
0.10	0.50	8.8	10.5
0.20	0.45	8.2	11.5

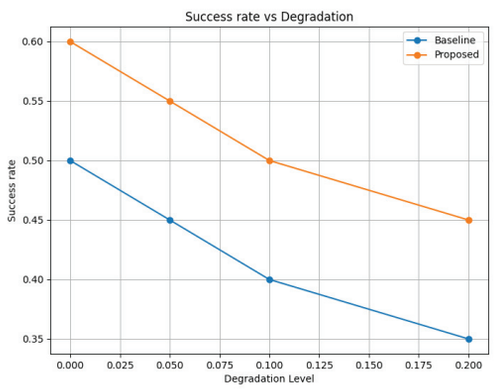


Fig. 3. Success rate across degradation levels for the baseline and the proposed model.

The EA-MARL agent's higher success rates across all degradation levels demonstrate that intrinsic motivation and affective alignment contribute positively to cooperative task performance. The average reward remains consistently higher than that of the baseline agent, and the average number of turns is slightly lower, indicating that the agent completes tasks faster and more efficiently even under challenging conditions.

C. Robustness Analysis

To evaluate the robustness of both the baseline MARL agent and the proposed EA-MARL agent under increasing degradation, we analyze several key performance metrics. The trust-weight visualization shows that the proposed EA-MARL agent consistently maintains higher trust values across all degradation levels. This suggests that intrinsic motivation and affective alignment enable more stable cooperative behavior, even as environmental noise intensifies. The success rate trends, shown in Fig. 3, provide further evidence of the proposed agent's resilience. While the baseline agent experiences a sharp decline in performance as degradation increases, the EA-MARL agent exhibits a more gradual and controlled reduction, consistently outperforming the baseline.

Robustness is also reflected in reward stability, as illustrated in Fig. 4. The average reward plot demonstrates that the proposed agent maintains higher reward values across all noise conditions, whereas the baseline model deteriorates more rapidly. This reflects the benefits of integrating intrinsic motivation and emotional regulation into the decision process.

Affective misalignment trends further highlight differences between the two agents. The EA-MARL agent sustains markedly lower misalignment values, indicating that

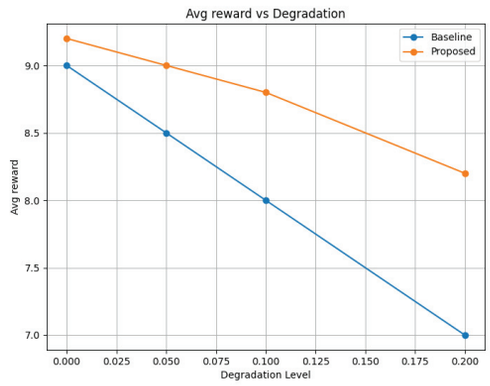


Fig. 4. Average reward across degradation levels for the baseline and the proposed model.

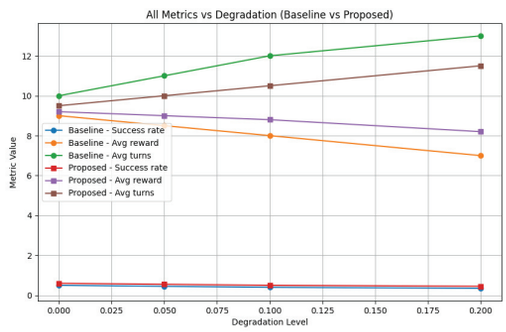


Fig. 5. Combined robustness metrics (success rate, average reward, affective misalignment) for the baseline and the proposed model.

its emotion-aware mechanisms help maintain coherence in internal affective states and promote stable multi-agent coordination. Finally, the combined robustness visualization in Fig. 5 aggregates success rate, average reward, and affective misalignment into a unified representation. This integrated analysis confirms that the proposed EA-MARL agent achieves balanced improvements across multiple evaluation dimensions, demonstrating superior resilience, adaptability, and overall reliability compared to the baseline model.

D. Discussion

The experimental results indicate that integrating intrinsic motivation and affective reasoning provides measurable benefits in multi-agent reinforcement learning. While the baseline agent struggles under increasing environmental noise, the proposed EA-MARL agent maintains higher task efficiency and stronger emotional coherence among teammates, resulting in more stable and resilient coordination.

Formally, the improvement in success rate can be expressed as:

$$\Delta SR = SR_{\text{EA-MARL}} - SR_{\text{Baseline}}.$$

At the highest degradation level (0.20), the improvement is:

$$\Delta SR = 0.45 - 0.35 = 0.10,$$

demonstrating a 10% absolute increase in success rate. Similar gains are observed in average reward, computed as:

$$\Delta AR = AR_{\text{EA-MARL}} - AR_{\text{Baseline}},$$

which at degradation 0.20 yields:

$$\Delta AR = 8.2 - 7.0 = 1.2.$$

These quantitative results collectively highlight the efficacy of the proposed EA-MARL framework. By incorporating affective alignment and intrinsic motivation, the agents exhibit improved robustness, enhanced adaptability, and more coherent team dynamics under challenging environmental conditions. The findings support the broader conclusion that affective mechanisms can serve as meaningful inductive biases for robust multi-agent coordination.

V. Conclusion

This work introduced an Emotion-Aware Multi-Agent Reinforcement Learning (EA-MARL) framework that integrates intrinsic motivation and affective alignment to enhance robustness in degraded environments. Through systematic evaluation across multiple degradation levels, the proposed agent consistently outperformed the baseline in success rate, average reward, and efficiency, demonstrating stronger resilience under uncertainty. In addition to improved task performance, the EA-MARL agent maintained lower affective misalignment and higher trust stability, enabling more coherent team coordination. The robustness analyses and combined metric visualizations further confirmed that the proposed framework achieves balanced gains across behavioral, emotional, and operational dimensions. These results highlight the value of embedding affective mechanisms as meaningful inductive biases, paving the way for more reliable and adaptive multi-agent systems in real-world, dynamically degraded settings.

VI. Future Work

Looking ahead, several extensions can further strengthen and expand the EA-MARL framework. A natural next step is to test the approach in heterogeneous multi-agent settings, where agents differ in abilities or roles. This would help determine how well emotionally aligned cooperation holds up when teams become more diverse. Another promising direction is the integration of real-time human feedback, particularly for human–robot collaboration scenarios where trust, emotion, and responsiveness are essential. Developing adaptive affective models that evolve with team interactions may also lead to richer coordination and more flexible emotional behavior over time. Finally, evaluating the framework in more complex environments such as large-scale navigation, resource management, or decentralized control would provide stronger evidence of its generalizability and real-world resilience.

ACKNOWLEDGEMENT

This research was supported by the Regional Innovation System & Education (RISE) Program through the Gwangju RISE Center. The project was funded by the Ministry of Education (MOE) and Gwangju Metropolitan City, Republic of Korea, under grant number 2025-RISE-05-013. The authors gratefully acknowledge this support, which made the successful completion of this work possible.

REFERENCES

[1] Z. Ning, et al., "A Survey on Multi-Agent Reinforcement Learning and Its Applications," *Pattern Recognition Letters*, 2024.

[2] A. Bukharin, Y. Li, Y. Yu, et al., "Robust Multi-Agent Reinforcement Learning via Adversarial Regularization: Theoretical Foundation and Stable Algorithms," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[3] S. He, et al., "Robust Multi-Agent Reinforcement Learning," arXiv:2106.02940, 2021–2023.

[4] M. Davoodabadi, N. Hashemi, M. Soleymani, "PreND: Enhancing Intrinsic Motivation in Reinforcement Learning Through Pre-Trained Network Distillation," arXiv:2401.01234, 2024.

[5] Y. Burda, H. Edwards, A. Storkey, and O. Klimov, "Exploration by Random Network Distillation," arXiv:1810.12894, 2018.

[6] A. Mahajan, T. Rashid, M. Samvelyan, and S. Whiteson, "MAVEN: Multi-Agent Variational Exploration," in *NeurIPS*, 2019.

[7] T. Moerland, J. Broekens, and C. M. Jonker, "Emotion in Reinforcement Learning Agents and Robots: A Survey," *Journal of Machine Learning Research*, 2022.

[8] Y. Wang, et al., "Trust-Based Consensus in Multi-Agent Reinforcement Learning," arXiv:2303.00812, 2023.

[9] D. Wu, et al., "Trust Region Method Using K-FAC in Multi-Agent Reinforcement Learning," in *AAMAS*, 2023.

[10] J. I. Kim, et al., "Sample-Efficient Multi-Agent Reinforcement Learning with Auxiliary Masked Reconstruction Task (M-QMIX)," *PLOS ONE*, vol. 18, no. 5, 2023.

[11] J. Guo, et al., "Towards Comprehensive Testing on the Robustness of Cooperative Multi-Agent Reinforcement Learning," in *CVPR Workshops*, 2022.

[12] R. Mu, X. Wang, and Y. Zhang, "Certifying Robustness of Value-Decomposition Cooperative MARL," arXiv:2202.01575, 2022.

[13] K. Cheruiyot, et al., "A Survey of Multi-Agent Reinforcement Learning," arXiv:2501.00412, 2025.

[14] Y. Liu, et al., "Trust-Aware Multi-Agent Reinforcement Learning for Resilient Coordination," *Engineering Applications of Artificial Intelligence*, 2025.

[15] H. Zhang, et al., "RobustDecision: Robust Decision-Making Under Environmental Degradation for Multi-Agent Systems," *Robotics and Autonomous Systems*, 2024.

AI-Based Adaptive Audio Generation with Frequency-Specific Processing

Taesoo Park
Department of Electronic Engineering
Kwangwoon University
Seoul, South Korea
taesoo0707@kw.ac.kr

Mujung Kim
Department of Electronic Engineering
Kwangwoon University
Seoul, South Korea
kmj1026@kw.ac.kr

Dohyun Park
Department of Electronic Engineering
Kwangwoon University
Seoul, South Korea
pdh6608@kw.ac.kr

Hyunjun Park
Department of Mathematics
Kwangwoon University
Seoul, South Korea
bbakhyun@kw.ac.kr

Jisang Yoo
Department of Electronic Engineering
Kwangwoon University
Seoul, South Korea
jsyoo@kw.ac.k

Hoyun Lee
Department of Otorhinolaryngology-Head and Neck Surgery
Ewha Womans University College of Medicine
Seoul, South Korea
hoyun1004@gmail.com

Sanghoon Kim
Department of Otorhinolaryngology
The One ENT Clinic
Daegu, South Korea
kmj1026@kw.ac.kr

*Soonchul Kwon
Graduate School of Smart Convergence
Kwangwoon University
Seoul, South Korea
ksc0226@kw.ac.kr

***Abstract*—This paper presents an AI-based audio generation pipeline that integrates generative music synthesis with adaptive frequency processing to create personalized soundscapes for various wellness and productivity applications. The proposed system leverages text-to-music generation models to produce base audio content from user-provided textual descriptions, then applies target-frequency processing and psychoacoustic enhancements to create customized listening experiences. The pipeline consists of five main components: (1) text-based prompt interface for style specification, (2) AI-driven music generation using models such as MusicGen, (3) target-frequency processing with adaptive filtering techniques, (4) enhancement through synthesized noise signals and binaural beats for cognitive state modulation, and (5) real-time streaming delivery with dynamic parameter control. The frequency-domain processing allows users to specify particular acoustic characteristics, while the addition of colored noise (white, pink, or brown) and binaural beats at specific frequencies (e.g., delta, alpha bands) enables targeted auditory experiences for relaxation, focus enhancement, or sleep improvement. The system architecture supports web-based deployment with real-time audio streaming capabilities, providing an accessible platform for personalized audio generation. Users can interactively adjust parameters including target frequencies, noise levels, binaural beat frequencies, and channel preferences through an intuitive interface. While this paper focuses on the technical design and implementation strategy rather than empirical evaluation, the framework establishes a foundation for future development of mobile applications and user experience studies. This work demonstrates the potential of combining state-of-the-art generative AI models with established signal processing techniques to create adaptive, aesthetically pleasing, and functionally diverse audio environments. The proposed approach represents a significant advancement over traditional static soundscape generators, offering unprecedented personalization capabilities for digital wellness applications.**

***Index Terms*—AI music generation, adaptive filtering, personalized audio, binaural beats, signal processing, wellness technology, soundscape synthesis.**

I. Introduction

Personalized audio experiences have become increasingly important for various wellness and productivity applications. Conventional audio generation methods—such as static noise generators or preset soundscapes—often lack the adaptability needed for individual preferences and acoustic requirements. [1], [2]. Recent advances in artificial intelligence (AI) and generative audio models have created new opportunities for producing personalized therapeutic sounds [3], [4] that better align with the auditory profiles of tinnitus patients.

Traditional noise-based audio systems, including white, pink, and brown noise, are designed for generic relaxation or focus enhancement, but they lack adaptability and aesthetic diversity. [8], [9]. In contrast, AI-generated audio allows for the creation of customized and musically rich soundscapes that can incorporate emotional or environmental preferences expressed by the user [3], [4]. By integrating these generative capabilities with targeted signal processing techniques, tinnitus management can be transformed from a uniform masking approach into a personalized auditory experience.

This study introduces an AI-based post-processing pipeline that combines generative audio synthesis with individualized filtering and sound enhancement for personalized acoustic environments. The proposed pipeline begins with text-based music generation using an AI model such as MusicGen, followed by frequency-domain filtering centered around the patient's tinnitus frequency to attenuate the perceived tone [1], [2]. Subsequently, binaural beats and synthetic noise are added to promote auditory masking and relaxation.

This paper focuses on the conceptual design and implementation strategy of the proposed system rather than quantitative evaluation. The framework aims to serve as the foundation for future clinical experiments and a deployable web or mobile application that can deliver customized sound therapy in real time. Through this approach, we aim to bridge the gap between AI-generated music and clinically applicable auditory treatments for tinnitus.

II. Pipeline

This section describes the proposed system pipeline for personalized tinnitus sound therapy, which follows the workflow: **text input** $\rightarrow$ **AI music generation** $\rightarrow$ **tinnitus-frequency centered filtering** $\rightarrow$ **noise and binaural beat synthesis** $\rightarrow$ **streaming output**. Each stage is implemented to support customization, flexibility, and real-time delivery.

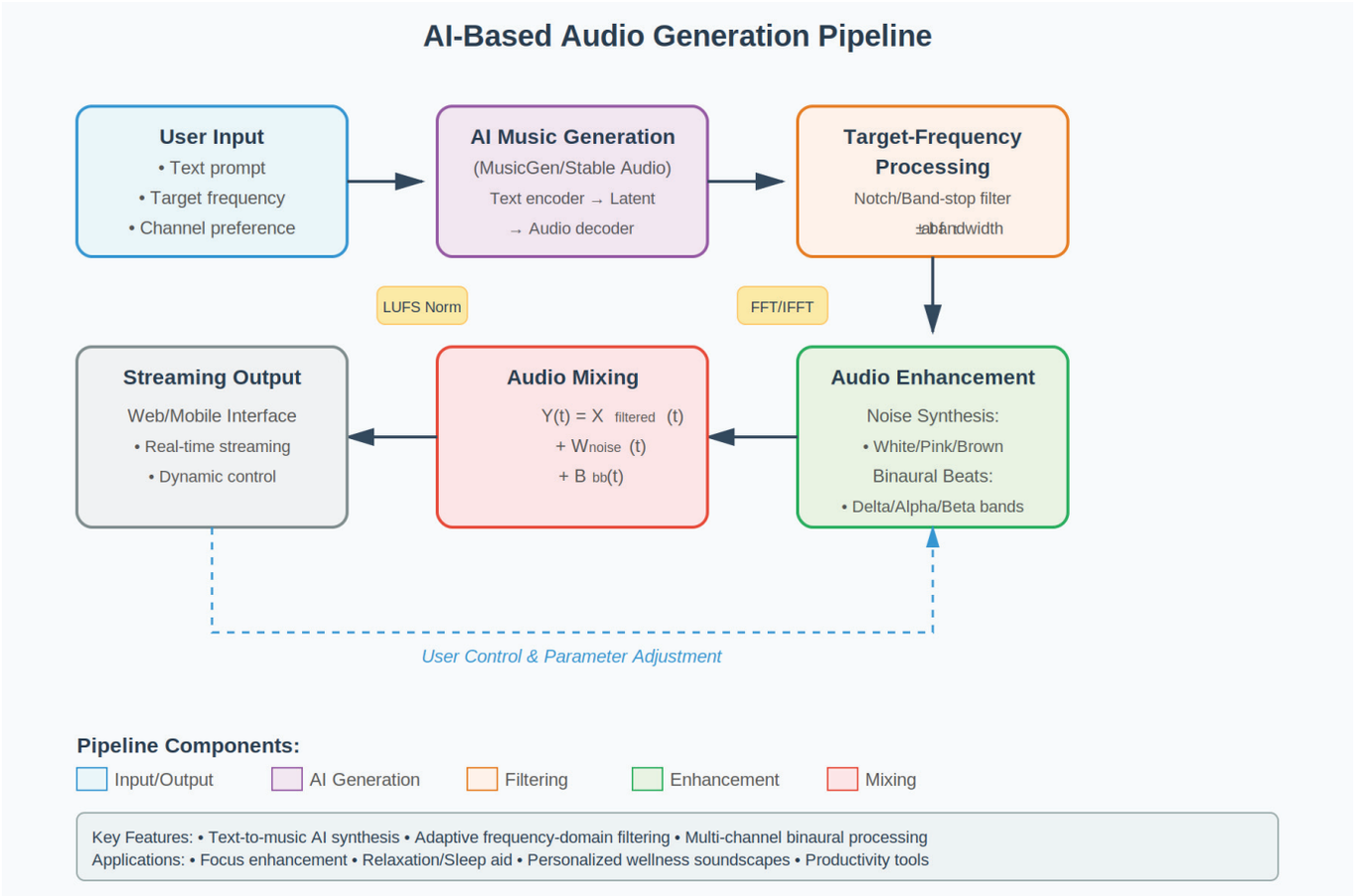


Fig. 1. Pipeline Diagram.

A. Text Input & Style Setting

The user interacts via a web interface to input a textual prompt describing their preferred auditory style (for example, "calm acoustic guitar melody with ocean wave background," or "soft piano ambient with light rain"). This prompt defines the mood, instrumentation, and duration of the generated audio. Alongside this, the user can specify target frequency parameters (for example '1.2kHz') and channel preferences (left, right, or both) for customized audio processing. These inputs serve as parameters for the subsequent generation and filtering stages.

B. AI Music Generation

Using the text prompt and style settings, the system invokes an AI-based music generation model (e.g., MusicGen, Stable Audio) [3], [4] to synthesize a base music track. The workflow may be summarized as follows:

- The prompt is encoded into a latent representation by a text encoder.
- The latent is fed into a music token generator or acoustic decoder (for instance, an EnCodec-based scheme) to produce a waveform.
- The output is rendered as a WAV or equivalent audio format and forwarded to the filtering module.

C. Target-Frequency Processing

Once the base track is available, a frequency-domain processing step is applied. The user-specified target frequency f_t becomes the center frequency for adaptive filtering (notch or band-stop). This allows for precise acoustic shaping to create personalized soundscapes that can enhance focus, relaxation, or other desired auditory experiences. [1], [2]. For example:

$$X_{\text{filtered}}(t) = \mathcal{F}^{-1}\{H(f) \cdot \mathcal{F}\{X(t)\}\}$$

where $H(f)$ is designed to significantly reduce energy in the band around f_t (e.g., ± 0.5 octave) while preserving adjacent frequencies. The filtered audio is then normalized in level (e.g., LUFS) to ensure comfortable listening.

D. Noise and Binaural Beat Synthesis

After filtering, the pipeline adds additional auditory components to enhance the listening experience and promote desired cognitive states. Two primary components are:

- **Noise signals**: Such as white, pink or brown noise, inserted to create ambient backgrounds that can improve focus, relaxation, or sleep quality.
- **Binaural beats**: Two slightly different tones are presented to left and right ears (e.g., 200 Hz in left ear,

202 Hz in right ear ⇒ perceived beat at 2 Hz). This beat frequency may correspond to neural oscillatory bands (e.g., delta: 2 Hz, alpha: 10 Hz) to promote relaxation or modulation of cortical activity [9].

The combined audio signal may be represented as:

$$Y(t) = X_{\text{filtered}}(t) + W_{\text{noise}}(t) + B_{\text{bb}}(t)$$

where $W_{\text{noise}}(t)$ is the chosen noise signal and $B_{\text{bb}}(t)$ is the binaural beat signal. Relative levels and mixes are adjusted based on user preference and system defaults to ensure an optimal listening experience.

E. Streaming & User Control Interface

The final therapeutic audio output is delivered via a web-based platform. The user can stream the generated audio in the browser or on a mobile device and has control over parameters such as:

- Re-entering a new text prompt to generate another style
- Updating the target frequency parameters or changing channel selection
- Adjusting noise level, switching between noise types, or selecting binaural beat frequency

The system also allows for logging usage data, collecting brief user feedback (e.g., preference ratings, listening duration), and preparing anonymized data for future user experience studies.

F. Summary

This pipeline integrates user-driven text prompts for AI music generation with tinnitus-specific signal processing (frequency filtering) and auditory stimulation techniques (noise, binaural beats) into a cohesive system for personalized sound therapy. Although this paper focuses on the design and technical workflow, the architecture provides a practical basis for future mobile app deployment and clinical evaluation [1], [8].

III. Conclusion and Future Work

This study presented an AI-based post-processing pipeline for personalized audio generation, integrating generative synthesis, target-frequency processing, and auditory enhancement through noise and binaural beat modulation. The system aims to provide a more flexible and patient-specific approach compared to conventional masking or notched-music methods by allowing users to generate customized therapeutic audio based on textual descriptions and individual hearing characteristics.

The proposed pipeline successfully bridges recent advances in generative audio modeling with established acoustic processing techniques for wellness applications. By combining the aesthetic adaptability of AI-generated music with signal processing methods such as frequency filtering and binaural beat synthesis, the framework introduces a new paradigm for interactive and personalized auditory treatment. The web-based implementation further demonstrates the feasibility of real-time sound generation and streaming within a user-friendly interface.

Although quantitative clinical validation was not included in this paper, the designed architecture provides a foundation for future evaluation and deployment. Future work will focus on:

- Developing a mobile and web application for public accessibility.
- Conducting user experience studies to evaluate effectiveness using subjective preference ratings and usage analytics.
- Expanding the AI model to incorporate emotional tone control and real-world acoustic context adaptation.

In summary, this research establishes a technical groundwork for next-generation personalized audio systems that merge artificial intelligence, sound design, and acoustic engineering for various wellness and productivity applications.

Acknowledgment

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(RS-2024-00352526)

References

[1] H. Teismann, H. Okamoto, and C. Pantev, “Short and intense tailor-made notched music training against tinnitus: the tinnitus frequency matters,” *PLOS ONE*, vol. 6, no. 9, e24685, 2011.

[2] H. Stracke, H. Okamoto, and C. Pantev, “Customized notched music training reduces tinnitus loudness,” *Commun. Integr. Biol.*, vol. 3, no. 3, pp. 274–277, 2010.

[3] G. Barrington et al., “Simple and Controllable Music Generation,” *arXiv preprint arXiv:2306.05284*, 2023.

[4] Z. Evans, C.J. Carr, J. Taylor, S.H. Hawley, J. Pons, “Stable Audio Open: A Text-to-Audio Generation Model,” *arXiv preprint arXiv:2407.14358*, 2024.

[5] C. W. Newman, G. P. Jacobson, and J. B. Spitzer, “Development of the Tinnitus Handicap Inventory,” *Arch. Otolaryngol. Head Neck Surg.*, vol. 122, no. 2, pp. 143–148, 1996.

[6] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-time Signal Processing*, 3rd ed., Prentice Hall, 1989, pp. 140–149.

[7] D.-H. Park and Y. Kim, “A real-time tinnitus treatment system using tinnitus frequency removal and adjacent frequency amplification on mobile device,” (Conference paper), 30 Nov. 2015.

[8] H. Okamoto, H. Stracke, W. Stoll, and C. Pantev, “Listening to tailor-made notched music reduces tinnitus loudness and tinnitus-related auditory cortex activity,” *Proc. Natl. Acad. Sci. U.S.A.*, vol. 107, pp. 1207-1210, 2010.

[9] S. Atipas, J. Therdphaothai, K. Suvansit, S. Prakairungthong, K. Thongyai, and S. Limviriyakul, “A randomized, controlled trial of notched music therapy for tinnitus patients,” *J. Int. Adv. Otol.*, vol. 17, no. 3, pp. 221-227, 2021.

[10] Alonso-Valerdi LM, Ibarra-Zárate DI, Torres-Torres AS, Zolezzi DM, Naal-Ruiz NE, Argüello-García J. ”Comparative analysis of acoustic therapies for tinnitus treatment based on auditory event-related potentials,” Front Neurosci. 2023 Apr 4;17:1059096.

[11] H. N. Shetty and V. Basavaraj, “Amplification strategies to reduce tinnitus: a paired comparison method,” *J. Int. Adv. Otol.*, vol. 18, no. 1, pp. 8-13, 2022.

DeepSets-S: An Encoder-Decoder Model for Real-Time Resource Allocation in In-network Edge Metaverse Systems

Sulaiman Muhammad Rashid *, Ibrahim Aliyu *, Seungmin Oh *, Yonggwan Won *, Jaehyung Park *, Jinsul K
* *Department of Intelligent Electronics and Computer Engineering, Chonnam National University, Gwangju, Republic of Korea*
Emails: *msulaimanrashid@jnu.ac.kr*, *aliyu@jnu.ac.kr*, *osm5252kr@gmail.com*, *ykwon@jnu.ac.kr*, *hyeoung@jnu.ac.kr*, *jsworld@jnu.ac.kr*

***Abstract*—The increasing demand for real-time and latency-sensitive services in next-generation networks necessitates efficient management of wireless and computing resources across multi-access edge computing (MEC) and computing-in-the-network (COIN) systems. Conventional optimization-based approaches provide optimal allocations but are computationally intensive and unsuitable for real-time operation. This paper investigates a slice-enabled in-network edge architecture and formulates a joint resource allocation problem that involves both intra-slice and inter-slice optimization. To address its complexity, the problem is decomposed into two subproblems and solved using a hierarchical DeepSets-based model. The proposed model employs a permutation-equivariant encoder and a slack-augmented decoder to learn allocation patterns from solver-generated datasets. Experimental results show that DeepSets-S achieves comparable allocation accuracy to optimization solvers while reducing computation time by up to 87.9%, demonstrating its suitability for large-scale edge environments.**

***Index Terms*—multi-access edge computing, computing-in-the-network, network slicing, resource allocation.**

I. Introduction

The growing complexity of MEC systems has intensified the demand for intelligent resource-allocation strategies that can adapt to highly dynamic user and network conditions [1], [2]. In slice-enabled edge environments, computing and communication resources are partitioned into multiple virtual slices, each tailored to distinct quality-of-service (QoS) and latency requirements [3]. Efficient allocation of bandwidth and computational capacity within and across these slices is critical for achieving high utilization and low service delay. However, the joint optimization of intra-slice and inter-slice resources remains computationally expensive due to the high-dimensional and time-varying nature of the decision space.

Traditional optimization frameworks, although theoretically optimal, require solving NP-hard problems that are impractical for real-time decision making [4]. Approximation and heuristic techniques reduce computational cost but often fail to generalize across network scales or adapt to unseen workloads [5]. To overcome these challenges, recent studies have explored learning-based solutions that approximate optimal allocation decisions directly from data [5]–[8]. Yet, many of these models are instance-based inference with individual decisions; we addressed a set-structured problem considering groups of Wireless Devices (WDs) jointly as a multiple instance learning problem.

To address these limitations, we propose a DeepSets-S-based learning framework for resource allocation in slice-enabled MEC systems. The proposed model learns to predict the optimal allocation of bandwidth and computing resources both within a slice and across slices from optimal solver-generated datasets. Our encoder captures relationships among users or slices irrespective of their order through permutation-equivariant, while a slack-augmented decoder enforces the resource-budget constraint directly during inference without additional normalization layers.

The remainder of this paper is organized as follows. Section II reviews related work on AI-driven resource allocation. Section III explains the system model, Section IV details the proposed DeepSets-S framework for SP1 and SP2. Section V presents experimental results and analysis, while Section VI concludes the paper

II. Related Works

Optimization of key performance metrics such as latency, energy consumption, and bandwidth utilization in edge environments often results in NP-hard problems that are computationally intensive in real-time [9]. Traditional mathematical formulations and solver-based methods yield optimal or near-optimal solutions but suffer from scalability issues as the number of users and slices increases [10].

To address these challenges, recent studies have explored machine learning (ML)-based approaches that learn tractable heuristics directly from data, reducing computational overhead while maintaining high-quality decisions [5], [11], [12]. For instance, a collaborative ML-driven resource-allocation strategy was introduced for SDN-enabled fog systems [11], achieving faster adaptation to dynamic workloads. Similarly, [4] evaluated several supervised models, including Decision Tree (DT), Support Vector Machine (SVM), Multilayer Perceptron (MLP), and Bagged Trees (BT) for task-placement optimization, highlighting the effectiveness of data-driven inference in approximating solver outputs.

Although these approaches leverage ML for edge-network management, most treat each decision independently, focusing on single-instance inference. In contrast, our work formulates resource allocation as a set-structured learning problem, where multiple wireless devices (WDs) are optimized jointly within or across slices. We employ a DeepSets-based encoder–decoder architecture to learn permutation-invariant representations of device or slice sets. This enables the model to capture inter-device relationships while remaining invariant to ordering.

III. System Model

We consider a slice-enabled multi-access edge computing (MEC) system composed of several access points (APs) connected to a set of wireless devices (WDs). The system supports multiple network slices, each corresponding to a service type with distinct latency and reliability requirements. Each slice is managed by a Slice Manager (SM) responsible for allocating communication and computing resources among its associated WDs, while a Slice Resource Orchestrator (SRO) coordinates inter-slice allocations across slices.

Let

$$\mathcal{S} = \{1, 2, \ldots, S\},$$

denote the sets of slices, and

$$\mathcal{U}_s = \{1, 2, \ldots, U_s\}$$

for set of WDs, each device $i \in \mathcal{U}_s$ generates a computational task characterized by: input data size D_i (in bits), required CPU cycles per bit f_i, channel gain $h_{s,i}$, and transmit power $P_{s,i}$.

A. Communication Model

The achievable uplink data rate between WD i and its AP is given by

$$R_{s,i} = b_{s,i} B_s \log_2\left(1 + \frac{P_{s,i} h_{s,i}}{N_0}\right), \tag{1}$$

where B_s is the total bandwidth allocated to slice s, $b_{s,i}$ is the normalized bandwidth share of WD i, and N_0 denotes the noise power spectral density.

The corresponding transmission time for a task of size D_i is

$$T_{s,i}^{\text{tx}} = \frac{D_i}{R_{s,i}}, \tag{2}$$

B. Computation Model

Each slice s has a total computational capacity F_s, allocated among its WDs as

$$f_{s,i} = q_{s,i} F_s, \quad \sum_{i \in \mathcal{U}_s} q_{s,i} \leq 1, \tag{3}$$

where $q_{s,i}$ represents the normalized computing share assigned to WD i.

The execution time for processing the task is

$$T_{s,i}^{\text{comp}} = \frac{D_i f_i}{f_{s,i}} = \frac{D_i f_i}{q_{s,i} F_s}, \tag{4}$$

C. Intra-Slice Resource Allocation (SP1)

For each slice s, the objective is to minimize the weighted sum of latency across all WDs, subject to bandwidth and computing constraints:

$$\begin{aligned} \text{(SP1):} \quad & \min_{\{b_{s,i}, q_{s,i}\}} \sum_{i \in \mathcal{U}_s} \left[\alpha_1 T_{s,i}^{\text{tx}} + \alpha_2 T_{s,i}^{\text{comp}}\right] \\ & \text{s.t.} \quad \sum_{i \in \mathcal{U}_s} b_{s,i} \leq 1, \quad b_{s,i} \geq 0, \\ & \qquad \sum_{i \in \mathcal{U}_s} q_{s,i} \leq 1, \quad q_{s,i} \geq 0. \end{aligned} \tag{5}$$

Here, α_1 and α_2, are weighting coefficients that balance delay trade-offs. The optimal allocation $\{b_{s,i}^*, q_{s,i}^*\}$ from SP1 serves as the training labels for the intra-slice allocation learner in the DeepSets-S model.

D. Inter-Slice Resource Allocation (SP2)

At the system level, the SRO determines the fraction of total bandwidth and computational capacity assigned to each slice. Let β_s denote the bandwidth proportions allocated to slice s, satisfying

$$\sum_{s=1}^{S} \beta_s \leq 1, \quad \beta_s. \tag{6}$$

Each slice's total latency is defined as the aggregated intra-slice delay

$$T_s = \sum_{i \in \mathcal{U}_s} \left(T_{s,i}^{\text{tx}} + T_{s,i}^{\text{comp}} \right), \tag{7}$$

weighted by the traffic load w_s.

The inter-slice allocation subproblem is formulated as

$$\begin{aligned} \text{(SP2):} \quad & \min_{\{\beta_s\}} \sum_{s=1}^{S} w_s T_s \\ & \text{s.t.} \quad \sum_{s=1}^{S} \beta_s \leq 1, \qquad \beta_s \geq 0. \end{aligned} \tag{8}$$

The optimal allocations $\{\beta_s^*, \phi_s^*\}$ from SP2 are used as training labels for the inter-slice allocation learner.

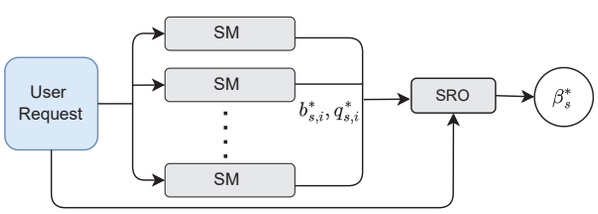


Fig. 1. Distributed DeepSets-S AI Model

IV. DeepSets-S Model

In what follows, we propose a hierarchical DeepSets-S model (Figure 1) that addresses SP1 and SP2 through a distributed architecture. Each Slice Manager (SM) runs a local DeepSets-S instance to solve SP1 by jointly allocating wireless and computing resources to user requests within its slice, while the Slice Resource Orchestrator (SRO) executes the SP2 model for inter-slice resource allocation. A unified permutation-equivariant encoder is shared across both levels to ensure consistency, with dedicated decoders tailored to the objectives of SP1 and SP2.

A. Set Representation and Padding

The encoder transforms each *set* of wireless devices (WDs), indexed by $i \in \mathcal{U}$, where the set size varies across slices or network instances. Each WD i contributes a feature vector $x_i \in \mathbb{R}^F$. The WDs' features and relevant slice or network state information are stacked into a single input matrix:

$$X = \begin{bmatrix} x_1^\top \\ \vdots \\ x_{\mathcal{U}}^\top \end{bmatrix} \in \mathbb{R}^{\mathcal{U} \times F}, \tag{9}$$

which represents a set-valued input where the row order carries no semantic meaning.

Because the WD count $\mathcal{U}$ differs across slices, we pad each X to a fixed size $\mathcal{U}_{\max}$ within a mini-batch:

$$\bar{X} = \text{pad}(X) \in \mathbb{R}^{\mathcal{U}_{\max} \times F}. \tag{10}$$

A binary mask $m \in \{0,1\}^{\mathcal{U}_{\max}}$ indicates real versus padded rows:

$$m_i = \begin{cases} 1, & i = \text{real WD}, \\ 0, & i = \text{padding}. \end{cases} \tag{11}$$

The mask is propagated through the network to (i) exclude padded rows from aggregation and (ii) prevent them from receiving allocation probability.

We standardize features to stabilize optimization. Let $\mu \in \mathbb{R}^F$ and $\sigma \in \mathbb{R}^F$ denote the per-feature mean and standard deviation computed over the training set. Each padded matrix is standardized as

$$\tilde{X}_{i,:} = (\bar{X}_{i,:} - \mu) \oslash \sigma, \qquad i = 1, \ldots, \mathcal{U}_{\max}, \tag{12}$$

where $\oslash$ denotes element-wise division.

B. Permutation-Equivariant Encoder

Given the standardized input $(\tilde{X}, m)$, the encoder produces contextual embeddings for each WD row while preserving permutation equivariance. The encoder consists of two shared multilayer perceptrons (MLPs): a row-wise embedding network ϕ and a context-fusing network ρ.

Each WD feature vector is first embedded by

$$h_i = \phi(\tilde{X}_{i,:}) \in \mathbb{R}^H, \qquad i = 1, \ldots, \mathcal{U}_{\max}. \tag{13}$$

A masked mean aggregation then computes the global context vector

$$c = \frac{\sum_{i=1}^{\mathcal{U}_{\max}} m_i \, h_i}{\sum_{i=1}^{\mathcal{U}_{\max}} m_i + \varepsilon} \in \mathbb{R}^H, \tag{14}$$

where $\varepsilon > 0$ avoids division by zero.

Each WD combines its local embedding h_i and the global context c via concatenation followed by the second MLP:

$$z_i = \rho([h_i; c]) \in \mathbb{R}^D, \qquad i = 1, \ldots, \mathcal{U}_{\max}. \tag{15}$$

Stacking the output rows yields

$$Z = \begin{bmatrix} z_1^\top \\ \vdots \\ z_{\mathcal{U}_{\max}}^\top \end{bmatrix} \in \mathbb{R}^{\mathcal{U}_{\max} \times D}. \tag{16}$$

Because both ϕ and ρ act row-wise and aggregation in (14) is symmetric, the encoder satisfies permutation equivariance:

$$\mathrm{Enc}(\Pi \tilde{X}, \Pi m) = \Pi\, \mathrm{Enc}(\tilde{X}, m), \tag{17}$$

for any permutation matrix $\Pi \in \{0,1\}^{\mathcal{U}_{\max} \times \mathcal{U}_{\max}}$.

C. Decoder for Intra-Slice Allocation (SP1)

Given the encoder output Z and mask m, the SP1 decoder maps each WD embedding to per-resource logits and converts them into valid probability distributions for bandwidth and computation shares, augmented with a slack entry that captures unused capacity. Let $\mathcal{E} = \{1, \ldots, E\}$ denote the set of resources (e.g., bandwidth, computation).

Each resource $e \in \mathcal{E}$ has learned parameters:

$$w_e \in \mathbb{R}^D, \quad b_e \in \mathbb{R}, \quad s_e \in \mathbb{R}. \tag{18}$$

The WD and slack logits are

$$l_{i,e} = w_e^\top z_i + b_e, \quad l_{0,e} = s_e, \tag{19}$$

and padded rows are masked out:

$$\tilde{l}_{i,e} = \begin{cases} l_{i,e}, & m_i = 1, \\ -\infty, & m_i = 0, \end{cases} \quad i = 1, \ldots, \mathcal{U}_{\max}, \quad \tilde{l}_{0,e} = l_{0,e}. \tag{20}$$

A softmax normalization over the augmented WD axis produces allocation probabilities:

$$p_{i,e} = \frac{\exp(\tilde{l}_{i,e})}{\sum_{j \in \{0,1,\ldots,\mathcal{U}_{\max}\}} \exp(\tilde{l}_{j,e})}, \tag{21}$$

where $i = 0$ corresponds to slack. The resulting matrix $P \in [0,1]^{(\mathcal{U}_{\max}+1) \times E}$ satisfies:

$$\sum_{i=0}^{\mathcal{U}_{\max}} p_{i,e} = 1, \quad \forall e \in \mathcal{E}. \tag{22}$$

Hence, the total allocation across real users obeys

$$\sum_{i=1}^{\mathcal{U}_{\max}} \hat{y}_{i,e} = 1 - p_{0,e} \le 1, \quad \forall e \in \mathcal{E}, \tag{23}$$

thus satisfying the resource constraint automatically.

Let $Y \in [0,1]^{(\mathcal{U}_{\max}+1) \times E}$ denote the target allocation (solver output) augmented with slack entries:

$$y_{0,e} = \max\Big\{0,\, 1 - \sum_{i=1}^{\mathcal{U}} y_{i,e}\Big\}, \quad \sum_{i=0}^{\mathcal{U}_{\max}} y_{i,e} = 1. \tag{24}$$

The SP1 loss is the masked cross-entropy over the augmented simplex:

$$\mathcal{L}_{\mathrm{SP1}} = -\sum_{e \in \mathcal{E}} \sum_{i=0}^{\mathcal{U}_{\max}} m_i\, y_{i,e}\, \log p_{i,e}. \tag{25}$$

D. Decoder for Inter-Slice Allocation (SP2)

For SP2, the output is set-level and identical for all WDs in the same slice. The encoder outputs are aggregated to a pooled set embedding:

$$g = \frac{\sum_{i=1}^{\mathcal{U}_{\max}} m_i\, z_i}{\sum_{i=1}^{\mathcal{U}_{\max}} m_i + \varepsilon} \in \mathbb{R}^D. \tag{26}$$

Each slice $s \in \mathcal{S}$ is assigned bandwidth and computation fractions through learned parameters:

$$w_e \in \mathbb{R}^D, \quad b_e \in \mathbb{R}, \quad s \in \mathbb{R}. \tag{27}$$

The slice logits and slack logit are

$$a_e = w_e^\top g + b_e, \quad a_0 = s, \tag{28}$$

and normalization yields:

$$p_k = \frac{\exp(a_k)}{\sum_{\ell \in \{0\} \cup \mathcal{E}} \exp(a_\ell)}, \quad k \in \{0\} \cup \mathcal{E}. \tag{29}$$

Predicted slice allocations $\hat{y}_e = p_e$ for $e \in \mathcal{E}$ satisfy

$$\sum_{e \in \mathcal{E}} \hat{y}_e = 1 - p_0 \le 1, \tag{30}$$

ensuring feasibility.

With soft labels $y = (y_e)_{e \in \mathcal{E}}$ and slack $y_0 = \max\{0,\, 1 - \sum_{e \in \mathcal{E}} y_e\}$, the loss for SP2 is

$$\mathcal{L}_{\mathrm{SP2}} = -\sum_{k \in \{0\} \cup \mathcal{E}} y_k \log p_k \tag{31}$$

V. Experiments

This section evaluates the proposed DeepSets–S model on the intra–slice (SP1) and inter–slice (SP2) allocation tasks. Unless otherwise specified, all results are reported as mean $\pm$ standard deviation over three random seeds.

We generate supervised labels from the optimization procedures described in Sec. IV. For SP1, solver outputs provide normalized user–level bandwidth/compute shares within each slice. For SP2, solver outputs provide

TABLE I
PERFORMANCE OF DEEPSETS-S MODELS ON SP1 AND SP2 AGAINST OTHER ABLATIONS (BEST RESULTS ARE BOLD).

Models	MAE (%)	Acc@0.5	Acc@1	Acc@2	R^2
SP1 Dataset					
DeepSets-S	**0.3924**	**0.7676**	**0.9526**	**0.9912**	0.9057
DeepSets	0.3920	0.7674	0.9521	0.9904	0.9057
SetTransformer (2 heads)	0.3922	0.7671	0.9522	0.9904	0.9055
SetTransformer (4 heads)	0.3918	0.7675	0.9523	0.9905	**0.9059**
SP2 Dataset					
DeepSets-S	**0.3480**	**0.8185**	**0.9567**	**0.9908**	0.9668
DeepSets	**0.3480**	0.8183	0.9564	0.9905	0.9668
SetTransformer (2 heads)	0.3475	0.8180	0.9566	0.9905	0.9673
SetTransformer (4 heads)	0.3474	0.8183	0.9566	**0.9908**	**0.9674**

normalized slice–level bandwidth/compute fractions at the network level. All labels are in $[0, 1]$ and satisfy the respective budget constraints.

We split the data into train/validation/test sets with ratios of 70/15/15.

In Table I, we compare the proposed DeepSets-S model with its variant *DeepSets* (without slack normalization) and the SetTransformer [13] architectures using 2- and 4-head configurations. The evaluation employs tolerance-based accuracy metrics, where Acc@0.5, Acc@1, and Acc@2 denote the proportion of predicted allocations that fall within 0.5, 1, and 2 units of the ground-truth solver outputs, respectively. Regression quality is further assessed using the mean absolute error (MAE) and the coefficient of determination (R^2).

For **SP1**, DeepSets-S achieves the highest accuracy across all thresholds, recording Acc@0.5 = 76.76%, Acc@1 = 95.26%, and Acc@2 = 99.12%. These results slightly outperform the no-slack DeepSets baseline (76.74%, 95.21%, 99.04%) and the SetTransformer variants (76.71–76.75%, 95.22–95.23%, 99.04–99.05%). Although the four-head SetTransformer attains the lowest MAE (0.3918) and the highest R^2 (0.9059), DeepSets-S maintains comparably low error while demonstrating more consistent accuracy across all thresholds.

For **SP2**, a similar trend is observed. DeepSets-S attains the best tolerance-based accuracy at every level, achieving Acc@0.5 = 81.85%, Acc@1 = 95.67%, and Acc@2 = 99.08%. These results are higher than those of DeepSets (81.83%, 95.64%, 99.05%) and marginally exceed the SetTransformer models (81.80–81.83%, 95.66%, 99.05–99.08%). While the four-head SetTransformer reports the lowest MAE (0.3474) and the highest R^2 (0.9674), DeepSets-S delivers nearly identical regression performance (MAE = 0.3480, R^2 = 0.9668) while inherently satisfying the resource-budget constraint through its slack-augmented decoder.

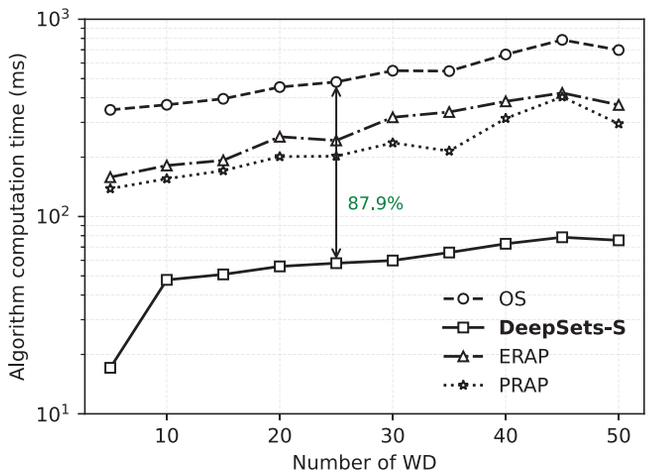


Fig. 2. Algorithm computation time

Fig. 2 presents the computation time comparison among the optimization solver (OS), DeepSets-S inference, Equal Resource Allocation Policy (ERAP), and Proportional Resource Allocation Policy (PRAP) for SP1 and SP2. The optimizer represents the combined solution of the first and second subproblems executed through Ipopt and Gurobi, while DeepSets-S denotes the learned surrogate performing direct inference.

As the number of WDs increases, the OS method exhibits a rapid growth in computation time due to the nonlinear and iterative nature of the optimization process. In contrast, the DeepSets-S approach maintains an approximately constant execution time for all cases, reflecting its suitability for real-time or large-scale operation. At $U = 25$, DeepSets-S achieves an **87.9% reduction** in computation time compared with the OS method, while preserving the accuracy level observed in solver-based allocations.

Both ERAP and PRAP methods show linear scaling with the number of devices, but remain slower than DeepSets-S because each still requires a decision optimization step solved by Gurobi. These results demonstrate that the DeepSets-S model can replace the iterative optimization process for SP1 and SP2 with a single-pass computation, substantially lowering the processing delay while maintaining reliable performance.

VI. CONCLUSION

This paper presented a DeepSets-S framework for efficient resource allocation in slice-enabled in-network

edge systems. The framework replaces repeated optimization with a direct inference process, reducing computation time while maintaining solution accuracy. Results show that DeepSets-S achieves performance close to the optimization solver but with a much lower computation time. The approach offers a practical direction for implementing scalable and time-efficient resource management in next-generation networks.

ACKNOWLEDGMENT

This work was supported by the Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT)(Project Name: Development of digital twin-based network failure prevention and operation management automation technology, Project Number: RS-2025-00345030, Contribution Rate: 50%. This work was also supported by the Institute of Information & Communications Technology Planning & Evaluation(IITP)-Innovative Human Resource Development for Local Intellectualization program grant funded by the Korea government(MSIT)(IITP-2025-RS-2022-00156287), Contribution Rate: 50%.

REFERENCES

[1] Naren, A. K. Gaurav, N. Sahu, A. P. Dash, G. Chalapathi, and V. Chamola, "A survey on computation resource allocation in iot enabled vehicular edge computing," *Complex & Intelligent Systems*, vol. 8, no. 5, pp. 3683–3705, 2022.

[2] I. Aliyu, S. Oh, N. Ko, T.-W. Um, and J. Kim, "Dynamic partial computation offloading for the metaverse in in-network computing," *IEEE Access*, vol. 12, pp. 11615–11630, 2023.

[3] S. Zhang, "An overview of network slicing for 5g," *IEEE Wireless Communications*, vol. 26, no. 3, pp. 111–117, 2019.

[4] G. Lia, M. Amadeo, G. Ruggeri, C. Campolo, A. Molinaro, and V. Loscrì, "In-network placement of delay-constrained computing tasks in a softwarized intelligent edge," *Computer Networks*, vol. 219, 2022.

[5] S. Muhammad Rashid, I. Aliyu, I.-K. Jeong, T.-W. Um, and J. Kim, "Graph neural network for in-network placement of real-time metaverse tasks in next-generation networks," *IEEE Access*, vol. 13, pp. 110999–111010, 2025.

[6] H. Guo, J. Liu, and J. Lv, "Toward intelligent task offloading at the edge," *IEEE Network*, vol. 34, no. 2, pp. 128–134, 2020.

[7] S. M. Rashid, I. Aliyu, A. Isah, M. Hahn, and J. Kim, "Blockchain-based task placement and resource management in edge computing: A survey," *Electronics*, vol. 14, no. 17, p. 3398, 2025.

[8] S. M. Araújo, F. S. de Souza, and G. R. Mateus, "A hybrid optimization-machine learning approach for the vnf placement and chaining problem," *Computer Networks*, vol. 199, p. 108474, 2021.

[9] A. A. Ismail, N. E. Khalifa, and R. A. El-Khoribi, "A survey on resource scheduling approaches in multi-access edge computing environment: a deep reinforcement learning study," *Cluster Computing*, vol. 28, no. 3, p. 184, 2025.

[10] S. M. Rashid, I. Aliyu, A. Isah, J. Lee, S. Oh, M. Hahn, and J. Kim, "Joint wireless and computing resource management with optimal slice selection in in-network-edge metaverse system," in *2024 15th International Conference on Information and Communication Technology Convergence (ICTC)*, pp. 922–926, 2024.

[11] J. Singh, P. Singh, M. Hedabou, and N. Kumar, "An efficient machine learning-based resource allocation scheme for sdn-enabled fog computing environment," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 6, pp. 8004–8017, 2023.

[12] F. S. D. Silva, S. N. Silva, L. M. Da Silva, A. Bessa, S. Ferino, P. Paiva, M. Medeiros, L. Silva, J. Neto, K. Costa, *et al.*, "Ml-based inter-slice load balancing control for proactive offloading of virtual services," *Computer Networks*, vol. 246, p. 110422, 2024.

[13] J. Lee, Y. Lee, J. Kim, A. Kosiorek, S. Choi, and Y. W. Teh, "Set transformer: A framework for attention-based permutation-invariant neural networks," in *International conference on machine learning*, pp. 3744–3753, PMLR, 2019.

Prediction of Atmospheric Refractivity from Clutter Power Using Deep Learning Based on the Tri-linear Atmospheric Model and Ocean Spectrum

Taekyeong Jin, Hosung Choo
Department of Electronic and Electrical Engineering, Hongik university, Seoul, South korea
Emails: *taekyeong.jin@mail.hongik.ac.kr, hschoo@hongik.ac.kr*

***Abstract*—In this paper, we propose a method for predicting atmospheric conditions from clutter power images using a deep neural network (DNN). The proposed refractivity from clutter (RFC) method employs a DNN model that utilizes clutter images as input. The output of the RFC method is four variables from the tri-linear atmospheric model that predict the current atmospheric conditions. The DNN model is trained on a dataset of 50,000 samples, with a training-to-validation data ratio of 4:1. Comparing average RMSE values, predictions using the artificial tri-linear atmospheric model improve by 37.65% with a dataset of 50,000 samples compared to 5,000 samples. Similarly, real measured data from Baengnyeongdo show a 24.17% improvement. These results demonstrate the capability of the proposed RFC method to predict various atmospheric conditions in real-time.**

Keywords— radio propagation, deep learning, refractivity from clutter, deep neural network.

I. INTRODUCTION

Naval long-range radar systems are typically used for the early detection of various distant targets, and interest in enhancing the detection capabilities of these long-range radars is growing. However, the detection performance of long-range radar systems can sometimes be degraded because the propagation path of the signal emitted from the radar varies depending on the atmospheric refractive index [1]. Therefore, the naval radar systems should compensate for detection errors by observing the atmospheric refractivity in real-time to maintain stable radar detection performance. Generally, the atmospheric refractivity according to the altitudes is measured using radiosonde equipment, but this measurement is limited to specific locations and is conducted only 2-4 times a day. To resolve this issue, research has been conducted to predict atmospheric conditions in real time by deep learning the collected clutter spectra [2]. But these studies primarily focus on predicting duct conditions, making it difficult to accurately predict various atmospheric conditions such as normal, super, sub, and duct [3].

In this paper, we propose a method for predicting the atmospheric refractive index values according to altitude using the clutter spectrum images collected in real-time. To accurately predict refractive index, the deep learning with deep neural networks (DNN) is applied to the proposed refractivity from clutter (RFC) method [4]. The trained DNN model takes only the clutter power images as input and outputs the four variables (d_1, d_2, h_1, and h_2) of the tri-linear atmospheric refractivity model [5]. Herein, the input dataset for training the DNN model consists of a collection of atmospheric refractive indices and clutter images for this atmospheric refractivity. To verify the proposed estimation method, we examine the estimation accuracy according to the amount of the training dataset, with the ratio of training data to validation dataset at 4:1. The results demonstrate that the proposed RFC method can be applied to radar systems to predict atmospheric conditions in real-time.

II. ESTIMATION OF AN ATMOSPHERIC REFRACTIVITY USING DEEP LEARNING

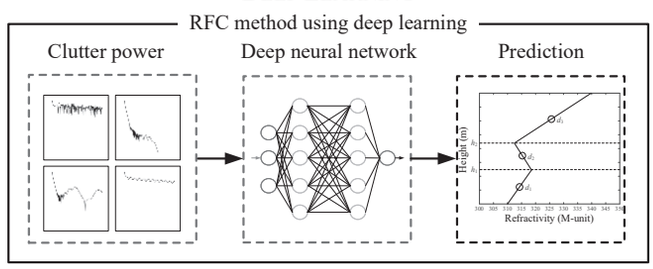
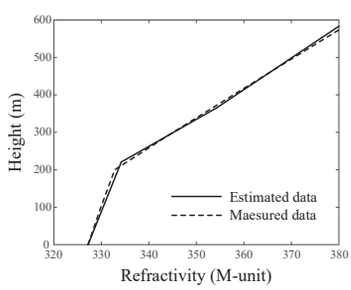
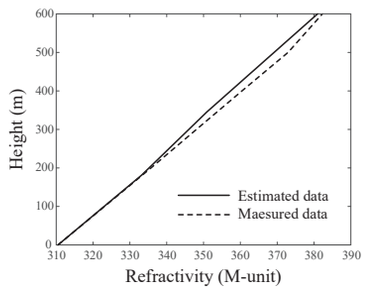


Fig. 1. RFC method of the radar system using deep learning.

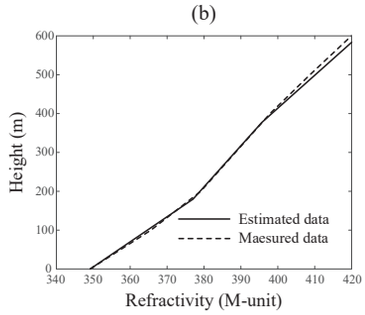
Fig. 1 illustrates the proposed RFC method to estimate the atmospheric refractivity. In this RFC method, clutter power spectrum images are used as the input data for the trained DNN. Herein, the dataset for training the DNN model consists of a collection of atmospheric refractive indices and clutter images. To represent as many atmospheric conditions as possible, the variables of the tri-linear atmospheric refractivity model are randomly changed using a uniform distribution to arbitrarily generate a dataset of 50,000 samples. Then, the clutter power images for this dataset are theoretically obtained using the propagation factor (PF) from the Advanced Refractive Prediction System (AREPS) simulator with random atmospheric refractivity [6]. After the input clutter spectrum images are entered in the trained DNN, the DNN determines the appropriate value of the atmospheric refractive index for the input clutter power spectrum. Finally, the DNN outputs estimated refractivity values in the form of a tri-linear atmospheric model.



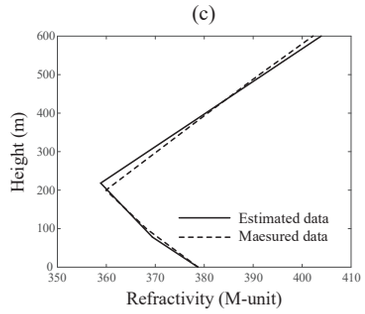
(a)



(b)



(c)



(d)

Fig. 2. Comparison of the prediction results and measured real atmospheric data. (a) Sub case (February 16, 2022). (b) Normal case (January 23, 2022). (c) Super case (September 5, 2022). (b) Duct case (October 8, 2022).

Fig. 2 presents the comparison of the prediction results and measured real atmospheric data. The meteorological data are obtained from Baengnyeongdo observatories. When the atmospheric conditions are sub, normal, super, and duct, the root mean square errors (RMSE) for the estimated atmospheric refractivity, calculated every 100 meters up to 600 meters, are 0.74, 2.24, 1.11, and 1.10, respectively, as shown in Figures 2(a) to 2(d). The results demonstrate that the proposed RFC method can estimate the atmospheric refractivity with low estimation error from the clutter power images.

TABLE 1. Analysis of the modified refractivity index *M* differences.

Amount of training data	The average of M-unit RMSE using the trilinear atmospheric model [M-unit]	The average of M-unit RMSE using the actual atmospheric data [M-unit]
5,000	8.95	18.66
10,000	7.28	16.74
30,000	6.15	15.48
50,000	5.58	14.15

Table 1 provides the results of an analysis based on graphs predicted by the tri-linear model, focusing on the differences in modified refractivity. The data measured in Baengnyeongdo are available every 100 meters, and the uniform distribution range for h_1 and h_2 is 300 meters. Therefore, the RMSE values for both predicted and target values are calculated at six points from 100 meters to 600 meters. When comparing the average of RMSE values based on the amount of data, the predictions using the artificial tri-linear atmospheric model show a 37.65% improvement with a dataset of 50,000 samples compared to a dataset of 5,000 samples. Similarly, predictions using 800 actual atmospheric datasets for cross-validation present a 24.17% improvement. These results represent that the proposed RFC method can also estimate current atmospheric conditions using clutter power images generated from actual atmospheric refractivity data.

III. Conclusion

We proposed the RFC method utilizing the tri-linear atmospheric model to generate artificial clutter power images, and a DNN model to predict atmospheric conditions. The input datasets for the proposed RFC method consisted of the clutter power images generated using the AREPS simulator together with the four variables (d_1, d_2, h_1, and h_2) of the tri-linear atmospheric refractivity model. The trained DNN model employs clutter power images as input and outputs the four variables of the tri-linear atmospheric refractivity model to predict atmospheric states. The DNN was trained on a dataset of 50,000 samples, and the ratio of training data to validation data was 4:1. When the average RMSE values were compared based on the amount of data, predictions using the artificial tri-linear atmospheric model represented a 37.65% improvement with a dataset of 50,000 samples compared to one of 5,000 samples. Similarly, predictions using 800 actual atmospheric datasets showed a 24.17% improvement. These results demonstrated that the proposed RFC method was capable of estimating current atmospheric conditions using actual atmospheric refractivity data in real-time.

Acknowledgment

This work was supported by Korea Research Institute for defense Technology planning and advancement (KRIT) grant funded by the Korea government (DAPA (Defense Acquisition Program Administration)) (No. KRIT-CT-22-021, Space Signal Intelligence Research Laboratory, 2022).

References

[1] B. K. Chalise, D. M. Wong, K. T. Wagner, and M. G. Amin, "Enhancing distributed radar detection performance with reliable communications," IEEE Trans. Aerosp. Electron. Syst., vol. 59, no. 4, pp. 4118–4133, Aug. 2023.

[2] T. Jin, J. Cho, D. Jang, and H. Choo, "Prediction of atmospheric duct conditions from a clutter power spectrum using deep learning," Remote Sens., vol. 16, no. 4, p. 674, Feb. 2024.

[3] T. H. Lim, M. Go, C. Seo, and H. Choo, "Analysis of the target detection performance of air-to-air airborne radar using long-range propagation simulation in abnormal atmospheric conditions," App. Sci., vol. 10, no. 18, p. 6440, Sep. 2020.

[4] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, and T. Chen, "Recent advances in convolutional neural networks," Pattern Recognit., vol. 77, pp. 354–377, May. 2018.

[5] C. Yang, Y. Wang, A. Zhang, H. Fan, and L. Guo, "A random forest algorithm combined with Bayesian optimization for atmospheric duct estimation," Remote Sens., vol. 15, no. 17, p. 4296, Aug. 2023.

[6] Advanced Refractive Prediction System (AREPS), Ver. 3.6, The Space and Naval Warfare System, San Diego, CA, USA, 2005.

Analysis of Solar Interference Impact on Future Networks in Extreme Cold Regions

1st ByoungWoo Jeon
Dept. of Computer Science
Kookmin University
Seoul, Republic of Korea
wjsquddn@kookmin.ac.kr

2nd Ari Hwang
Dept. of Financial Information Security
Kookmin University
Seoul, Republic of Korea
hwangr2@kookmin.ac.kr

3rd Dong-jin Yoon
Center of Technology Development
Korea Polar Research Institute
Incheon, Republic of Korea
dj_yoon@kopri.re.kr

4th Soohyun Park
College of Computer Science
Kookmin University
Seoul, Republic of Korea
shpark21@kookmin.ac.kr

***Abstract*—Reliable terrestrial communication networks are critical in Extreme Cold Regions due to harsh environmental conditions and infrastructure constraints. This paper investigates the Sun Outage phenomenon in terrestrial microwave links—typically associated with satellite communications—caused by the geometric characteristics of low solar trajectories in high latitudes. This paper proposes a quantitative analysis framework to evaluate the impact of geometric solar interference on ground-based networks. Mathematical modeling based on coordinate transformation and precise numerical integration of the antenna gain pattern over the solar disk were performed. Simulation results for a hypothetical network near the Jang Bogo Station confirm a maximum Carrier-to-Noise (C/N) degradation of 6.59 dB. The analysis also quantitatively verifies a trade-off where high-gain antennas shorten the interference duration but increase the magnitude of signal degradation. Given the deterministic nature of this phenomenon and limited power resources in polar environments, this study concludes that prediction-based proactive network management, rather than simple power transmission increases, is essential for maintaining link availability.**

***Index Terms*—Sun Outage, High-latitude Communication, Terrestrial Microwave Link, Extreme Cold Region, Geometric Modeling, Solar Interference.**

I. Introduction

The Extreme Cold Region refers to high-latitude areas generally above 60° N/S, characterized by extremely low temperatures where the annual average drops below -10°C and severe climatic conditions. These regions are distinguished by winter temperatures that can plummet below -40°C, strong winds, heavy snowfall, and the phenomena of the Midnight Sun and Polar Night. In particular, they possess the unique characteristic of being a Radio Quiet Zone, where artificial radio interference is minimal [1].

Antarctica, classified as an extreme cold region, has seen an increase in scientific and economic exploration activities due to geopolitical interests and climate change research. As the demand for high-speed communication infrastructure for large-capacity data transmission between stations increases, research on communication infrastructure utilizing high-speed and high-gain antennas is required to address these communication needs.

Compared to general domain system configurations, terrestrial systems in extreme cold regions typically involve long physical distances between nodes and face maintenance difficulties due to harsh conditions such as extreme low temperatures, strong winds, and heavy snowfall. Consequently, high-reliability system design is required, with electric power being a primary limited resource due to the extreme environment [2]. In such environments, particularly in the Radio Quiet Zone where the background noise level is low unlike in urban areas, the influx of solar noise can increase the system noise temperature, thereby further degrading system performance.

Solar interference affecting communication systems is classified into two main categories. The first is stochastic interference caused by solar activity, such as Solar Radio Bursts (SRB) and Polar Cap Absorption (PCA), which is characterized by its unpredictability [3], [4]. The second is Sun Outage, which is a deterministic interference based on geometric positioning [5], [6]. This phenomenon occurs when the receiving antenna, transmitting antenna, and the Sun are collinearly aligned, causing intense solar thermal noise to enter the antenna beamwidth, thereby increasing the receiver noise temperature and degrading the Signal-to-Noise Ratio (SNR) [7].

Sun Outage has primarily been addressed as an issue in satellite communications [5], [6]. However, high-latitude regions like Antarctica present a unique challenge. In these areas, the solar trajectory remains at a consistently low elevation along the horizon. Simultaneously, terrestrial antennas are directed at near-horizontal angles toward ground nodes. This combination creates a new geometric interference condition where the Sun aligns with the antenna beamwidth—a scenario not typically observed in mid-to-low latitude regions.

Simulation results from this study confirm that applying the ITU-R standard model can lead to a C/N degradation of 6.59

dB during Sun Outage events. In typical communication environments, this degradation could be mitigated by increasing transmission power. However, given that electric power is a limited resource in extreme cold regions, compensating for link budget loss through simple output increases is practically difficult. Therefore, this study aims to analyze the Sun Outage problem in high-latitude terrestrial communications, focusing on its predictable nature that allows for proactive preparation and avoidance.

II. System Model and Analysis Method

A. Sun Outage Principles and High-Latitude Geometric Characteristics

1) General Definition of Sun Outage: The Sun acts as a broadband thermal noise source, radiating significant energy across all frequency bands. During the operation of a communication system, a geometric alignment may occur where the receiving antenna, the transmitting antenna, and the Sun become collinear. Consequently, solar noise energy enters the antenna, increasing the total system noise temperature. Consequently, this degrades the Signal-to-Noise Ratio (SNR), potentially leading to a communication outage. This phenomenon is referred to as Sun Outage [7].

2) Geometric Uniqueness of High-Latitude Environments: In mid- and low-latitude regions with high population densities, the Sun moves at a high elevation; consequently, geometric alignment between the Sun and terrestrial communication antennas directed toward the horizon is rare or transient. In contrast, high-latitude regions such as Antarctica are characterized by a solar trajectory that moves at a low elevation along the horizon.

Antennas for terrestrial point-to-point communication are directed at near-horizontal angles toward ground target nodes. The combination of these two conditions establishes a geometric alignment wherein the low-altitude Sun can transit through the main beam of the horizontally oriented antenna for extended durations. This indicates that Sun Outage can occur in terrestrial networks, not only in satellite communications. Furthermore, the Midnight Sun phenomenon during the Antarctic summer acts as a factor that extends the temporal range of potential solar interference exposure.

3) Introduction of the Horizon Coordinate System: To analyze this potential geometric interference, we introduced a Horizon Coordinate System centered on the receiver station (O). To determine whether the Sun overlaps with ground target nodes, the geodetic coordinates (latitude and longitude) of each node were converted into this observer-centered system. The specific transformation equations are detailed in the subsequent Section II-B. The modeling components are defined as follows:

- **Observer (O):** Defined as the Campbell Glacier station (origin). Coordinates are expressed using Azimuth (measured clockwise from True North) and Elevation (measured from the horizon).
- **Target Node (R):** A ground transmitting node characterized by fixed Azimuth and Elevation angles.

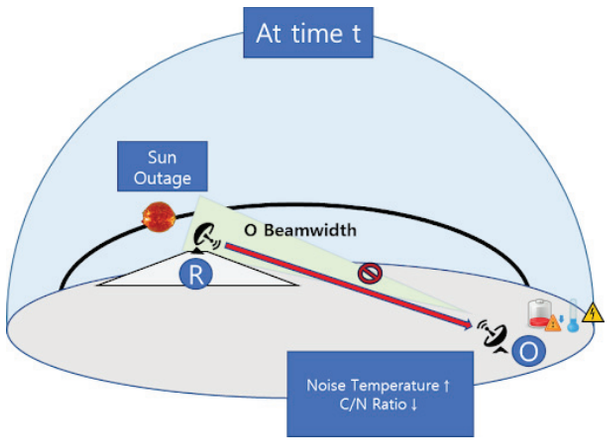


Fig. 1. Conceptual illustration of a Sun Outage scenario in a polar environment, showing the low-angle alignment of the transmitting station, the receiving station, and the Sun.

- **Sun (S):** Modeled as a time-dependent function $(Az_{sun}(t), El_{sun}(t))$. The `pvlib` library [8] was used for precise trajectory tracking.

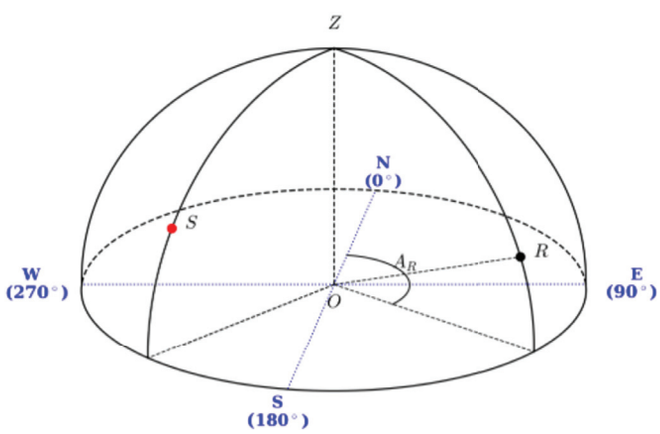


Fig. 2. Schematic of the Horizon Coordinate System.

Consequently, a Sun Outage event is defined to occur when the solar coordinates at time t match the target coordinates or fall within the antenna beamwidth.

Fig. 3 presents the simulation results projecting the actual solar trajectory and surrounding ground nodes in Antarctica onto a 2D plane within the horizon coordinate system. As shown, the solar path during March 2025 intersects the positions of ground target nodes multiple times. To quantitatively calculate the intensity of this interference, the next section describes the process of converting this geometric relationship into the antenna's viewpoint for in-depth analysis.

B. Mathematical Modeling and Interference Analysis Method

1) Coordinate Transformation: From Geodetic to Horizon: The geographical locations of the ground communication nodes, specifically the receiver (O) and the transmitter (R), are defined by latitude, longitude, and altitude based on the

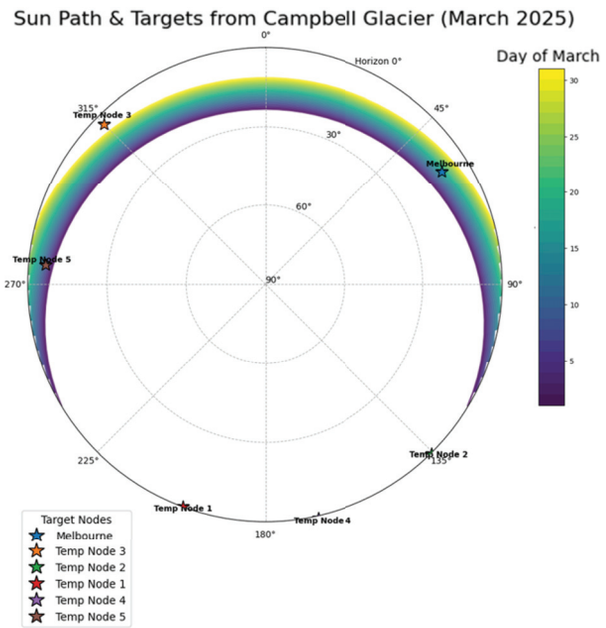


Fig. 3. Simulation results projecting the solar trajectory and ground nodes in Antarctica onto a 2D plane within the horizon coordinate system (March 2025).

WGS84 (World Geodetic System 1984) standard. However, to analyze the relative geometric relationship between the receiver (O) and the transmitter (R), it is necessary to first convert these coordinates into a common three-dimensional Cartesian coordinate system.

To achieve this, a two-step coordinate transformation was performed. First, the geodetic coordinates were converted into the Earth-Centered, Earth-Fixed (ECEF) coordinate system (x, y, z), which has its origin at the center of the Earth, to represent absolute positions in three-dimensional space.

Second, a final transformation was performed into the ENU (East-North-Up) coordinate system, also known as the horizon coordinate system, with the receiver (O) as the origin and its tangent plane as the reference. Through this process, the positions of the transmitter (R) and the Sun (S) are derived in terms of azimuth and elevation relative to the observer.

In this framework, since the solar position (Az_{sun}, El_{sun}) varies continuously with time t, the proposed framework generates high-resolution data at one-second intervals using the `pvlib` library [8] for precise trajectory tracking. In contrast, the ground transmitter (R) is treated as having fixed coordinates $(Az_{target}, El_{target})$.

2) Antenna and Solar Modeling: For interference analysis, this study models the characteristics of the Sun and the antenna with reference to standard recommendations [7], [9], [10] as follows.

a) Solar Disk Modeling: In general satellite orbit analysis, the Sun is typically assumed to be a point source. However, to accurately calculate the noise energy entering the antenna beamwidth, the Sun must be modeled as a circular disk with a finite angular radius θ_s ($\approx 0.26°$). In this study, while assuming that a uniform brightness temperature is radiated from every differential area within the solar disk, the equivalent solar noise temperature T_{sun} received by the antenna is defined as:

$$T_{sun} = 120000 \times 0.5 \times f^{-0.75} \tag{1}$$

where f is the frequency in GHz. The factor 0.5 accounts for the polarization mismatch between the unpolarized solar radiation and the polarized receiving antenna.

b) Antenna Pattern Modeling: The receiving antenna gain pattern $G(\theta)$ is defined based on the Reference Radiation Pattern specified in ITU Radio Regulations Appendix 30, Annex 5, Figure 8 [10]. This standard model reflects the general characteristics of parabolic antennas, expressed as a function where the antenna gain attenuates according to the off-axis angle θ from the boresight.

This model includes the assumption that the antenna pattern has iso-gain characteristics, implying rotational symmetry around the boresight. For gain calculation, the angular variable $x = \theta/\phi_{3dB}$, normalized by the half-power beamwidth (ϕ_{3dB}), is used. In this study, ϕ_{3dB} is derived using the standard approximation for parabolic antennas, $70\lambda/D$ [7].

3) Transformation to Antenna Coordinate System and Planar Projection: While the horizon coordinate system is useful for representing the absolute solar position (Az, El), directly deriving the off-axis angle θ—which is essential for calculating interference energy—is geometrically complex.

To facilitate this calculation, we first converted the solar position from spherical coordinates into a 3D Cartesian vector $\mathbf{P}_{hor} = [x, y, z]^T$ within the horizon frame. Subsequently, a Rotation Matrix ($\mathbf{R}$) was defined such that the new z'-axis aligns with the antenna boresight, and the x'-axis is parallel to the horizon. Applying this matrix transforms the solar vector into the antenna coordinate system, denoted as $\mathbf{P}_{ant}$:

$$\mathbf{P}_{ant} = \mathbf{R} \cdot \mathbf{P}_{hor} = [x_{sun}, y_{sun}, z_{sun}]^T \tag{2}$$

Finally, the transformed vector $\mathbf{P}_{ant}$ is projected onto the 2D antenna plane ($x' - y'$).

Although geometric distortion may occur when projecting a spherical surface onto a plane, Sun Outage occurs only when the Sun is positioned within or near the narrow antenna beamwidth. Therefore, within this limited analysis range, the Small Angle Approximation is valid. Consequently, the Euclidean distance from the origin to the solar center (x_{sun}, y_{sun}) on this projected plane directly approximates the off-axis angle θ:

$$\theta \approx \sqrt{x_{sun}^2 + y_{sun}^2} \tag{3}$$

4) Precise Numerical Integration: The total interference noise entering the antenna is calculated by integrating the antenna gain pattern over the solar disk area. According to ITU-R Recommendation S.1525 [7], the increase in antenna noise temperature, ΔT_A, is defined as the ratio of solar energy influx to the total antenna reception pattern:

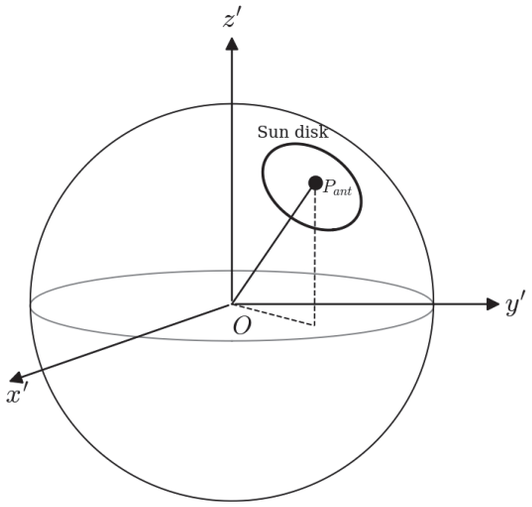


Fig. 4. Concept of coordinate rotation: The solar position vector is transformed from the horizon system ($\mathbf{P}_{hor}$) to the antenna coordinate system ($\mathbf{P}_{ant}$).

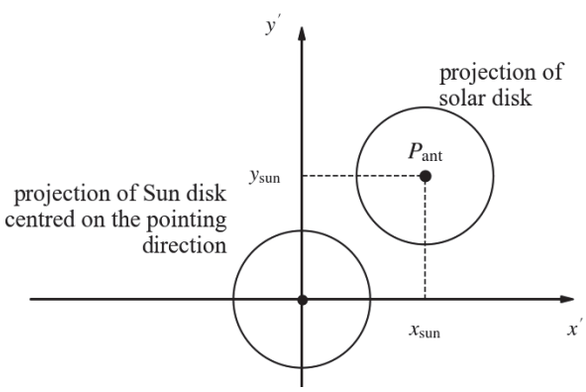


Fig. 5. Projection of the Sun on the $x' - y'$ plane.

$$\Delta T_A = \frac{\iint_{S_{disk}} T_{sun} \cdot G(\Omega)\, d\Omega}{\iint_{4\pi} G(\Omega)\, d\Omega} \tag{4}$$

The denominator represents the pattern solid angle, denoted as Ω_{ant}. To ensure accuracy, this value was calculated via rigorous spherical integration over the full spatial domain ($0 \leq \theta \leq \pi$). Considering the rotational symmetry of the antenna pattern, Ω_{ant} is expressed as:

$$\Omega_{ant} = \int_0^{\pi} G(\theta) \cdot 2\pi \sin\theta \, d\theta \tag{5}$$

For the numerator, the planar projection model derived in the previous section was applied for computational efficiency. Let (x_{sun}, y_{sun}) be the projected coordinates of the solar center. The numerical integration is performed on the 2D plane using local coordinates (u, v) relative to the solar center:

$$\Delta T_A \approx \frac{1}{\Omega_{ant}} \int_{-\theta_s}^{\theta_s} \int_{-\sqrt{\theta_s^2 - u^2}}^{\sqrt{\theta_s^2 - u^2}} T_{sun} \cdot G(\theta_{off})\, dv\, du \tag{6}$$

where θ_s is the solar angular radius defined in Section II-B. The variable θ_{off} represents the effective off-axis angle for each differential area (u, v) within the solar disk. Based on the small angle approximation, it is defined by the Euclidean distance on the projected plane:

$$\theta_{off} = \sqrt{(x_{sun} + u)^2 + (y_{sun} + v)^2} \tag{7}$$

Finally, the magnitude of Carrier-to-Noise ratio (C/N) degradation is derived by incorporating the increased antenna temperature into the baseline system noise temperature (T_{sys}):

$$\Delta(C/N) = 10 \log_{10} \left(\frac{T_{sys} + \Delta T_A}{T_{sys}} \right) \quad [\text{dB}] \tag{8}$$

III. Simulation Environment

A. Network Topology

The simulation network is configured as a hypothetical terrestrial microwave link located near the Jang Bogo Station in Antarctica. The node locations were selected based on the polar control point information specified in [1]. To verify Line-of-Sight (LoS) availability between control points for long-distance transmission, the terrain altitude was analyzed using Google Earth [11] as shown in Fig. 7. Based on this analysis, relay nodes were heuristically placed to ensure LoS.

In the configured network, the maximum elevation difference between nodes reaches 2,100 m, and the maximum link distance is approximately 26 km. Accordingly, it was assumed that each node is equipped with an antenna system capable of supporting a transmission range of up to 50 km.

For geometric analysis, the Campbell Glacier node was set as the observation point within the horizon coordinate system described in Section II-A. The positions of surrounding nodes relative to this observation point, including their azimuth and elevation, are listed in Table I.

TABLE I
Node Locations and Horizon Coordinates relative to Campbell Glacier

Node Name	Lat. (DMS)	Lon. (DMS)	Alt. (m)	Az. (°)	El. (°)
Melbourne	74°20′59″ S	164°41′27″ E	2624.8	57.33	10.24
Temp Node 1	74°36′26″ S	164°05′22″ E	624.4	315.01	3.64
Temp Node 2	74°30′49″ S	164°46′24″ E	511.8	135.52	-0.23
Temp Node 3	74°17′45″ S	163°58′16″ E	1673.4	200.39	0.05
Temp Node 4	74°33′32″ S	164°30′14″ E	276.0	167.22	-1.00
Temp Node 5	74°23′41″ S	163°57′40″ E	1953.2	275.10	6.35

B. System Parameters

The simulation frequency was selected as 11 GHz, a band commonly used for long-distance terrestrial microwave communications. This frequency serves as a reference value to analyze geometric interference patterns in terrestrial links, independent of specific propagation suitability for the Antarctic environment.

Antenna specifications were based on the reference radiation pattern specified in ITU Radio Regulations Appendix 30 Annex 5, Figure 8 [10]. To investigate the impact of beamwidth

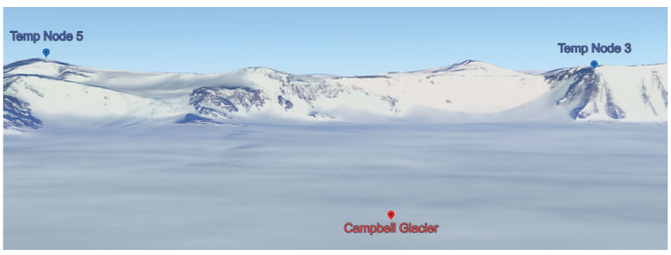


Fig. 7. 3D terrain visualization for LoS verification between Campbell Glacier and relay nodes using Google Earth.

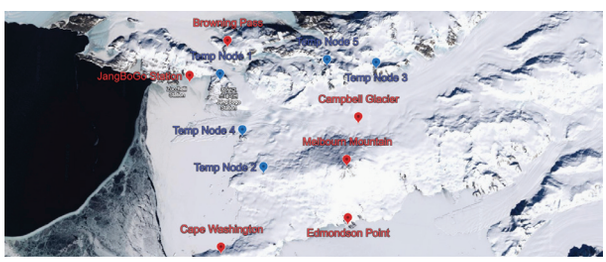


Fig. 8. Simulation network topology. Red markers indicate control points, and blue markers indicate hypothetical relay nodes (labeled as 'Temp Node' in the figure).

on the intensity and duration of interference, two scenarios were defined:

- **Scenario 1 (Small Aperture):** Antenna diameter $D = 0.6$ m, corresponding to a half-power beamwidth $\phi_{3dB} \approx 3.180°$.
- **Scenario 2 (Large Aperture):** Antenna diameter $D = 2.4$ m, corresponding to a half-power beamwidth $\phi_{3dB} \approx 0.795°$.

The baseline system noise temperature (T_{sys}) was assumed to be **155 K** under clear sky conditions, as referenced in [9].

C. Analysis Period

In high-latitude terrestrial communications, the timing of interference is determined by the relative geometric arrangement (azimuth and elevation) of the nodes.

Monthly solar trajectories projected onto the horizon coordinate system centered at the receiver (Campbell Glacier) were analyzed. The results indicated a potential for collinear alignment (Receiver-Transmitter-Sun) during the periods of February–April and August–November.

as shown in Fig. 3, frequent geometric alignments were specifically identified in the link between Campbell Glacier and Melbourne Mountain during March. Consequently, this study selected March 2025 as the analysis period to perform a detailed investigation of Sun Outage on this link.

IV. Simulation Results and Discussion

In the simulation analysis presented in this section, unless specified otherwise, the term off-axis angle (θ) refers to the angular separation between the center of the solar disk (Solar Center) and the antenna boresight.

A. Simulation Results

1) Scenario 1: Small Aperture (0.6 m): Calculated based on the simulation parameters, the half-power beamwidth (ϕ_{3dB}) of the 0.6 m antenna is approximately $3.180°$, indicating relatively broad directivity. For analysis, θ_{min} is defined as the minimum angular separation between the solar center and the antenna boresight during the daily solar transit. The maximum C/N degradation at this moment is adopted as the performance metric.

Table II summarizes the Sun Outage events observed in March 2025 for Scenario 1, and Fig. 6 presents the daily trends.

The simulation results indicate that Sun Outage occurred for a total of **19 days** in March. Due to the relatively wide beamwidth ($3.180°$), the solar transit time through the antenna pattern was prolonged. Consequently, on the day with the smallest θ_{min} (March 17), the interference duration reached a maximum of **29.7 minutes**, exhibiting a long-duration characteristic.

As shown in the table, there is a distinct trend where a smaller θ_{min} (closer alignment to the antenna center) corresponds to an increase in the influence of antenna gain, thereby

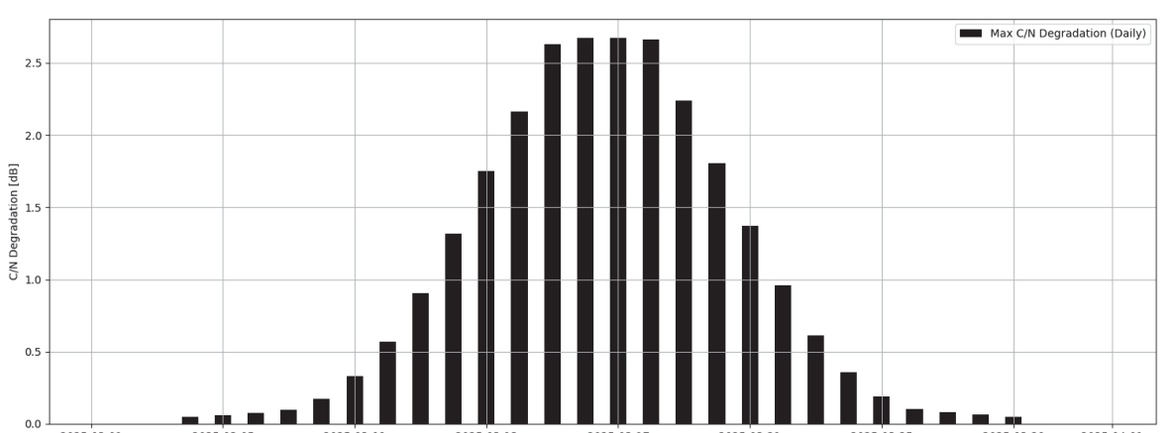


Fig. 6. Daily maximum C/N degradation for Scenario 1 (0.6 m antenna).

TABLE II
SUN OUTAGE EVENTS FOR SCENARIO 1 (0.6 M ANTENNA)

Date	Start (Local)	Duration (min)	θ_{min} (°)	Max Deg. (dB)
2025-03-08	10:23	12.8	3.350	0.17
2025-03-09	10:20	18.0	2.961	0.33
2025-03-10	10:18	21.5	2.571	0.57
2025-03-11	10:16	24.0	2.181	0.91
2025-03-12	10:15	26.0	1.790	1.32
2025-03-13	10:14	27.5	1.399	1.75
2025-03-14	10:13	28.7	1.007	2.16
2025-03-15	10:13	29.2	0.615	2.63
2025-03-16	10:12	29.5	0.223	2.67
2025-03-17	**10:12**	**29.7**	**0.170**	**2.67**
2025-03-18	10:12	29.3	0.562	2.66
2025-03-19	10:12	28.7	0.954	2.24
2025-03-20	10:12	27.5	1.346	1.81
2025-03-21	10:12	26.2	1.737	1.37
2025-03-22	10:13	24.3	2.128	0.96
2025-03-23	10:14	21.7	2.518	0.61
2025-03-24	10:15	18.3	2.908	0.36
2025-03-25	10:17	13.7	3.296	0.19
2025-03-26	10:22	4.2	3.684	0.10

increasing C/N degradation and extending the interference duration.

2) Scenario 2: Large Aperture (2.4 m): Based on the simulation parameters, the half-power beamwidth (ϕ_{3dB}) of the 2.4 m antenna is calculated to be approximately $0.795°$, which is narrower than that of Scenario 1.

Table III summarizes the Sun Outage events observed in March 2025 for Scenario 2, and Fig. 9 illustrates the daily maximum C/N degradation.

Simulation results for March indicate that Sun Outage events were observed for a total of **8 days**. Due to the narrower beamwidth, the solar transit time through the antenna pattern was reduced. Even on the day of most precise alignment (minimum θ_{min}), the interference duration was limited to **11.6 minutes**, which is shorter than the approximately 30 minutes observed in Scenario 1.

However, as shown in Table III and Fig. 9, the high-gain characteristic associated with the narrower beamwidth resulted in a sharp increase in C/N degradation as θ_{min} decreased.

TABLE III
SUN OUTAGE EVENTS FOR SCENARIO 2 (2.4 M ANTENNA)

Date	Start (Local)	Duration (min)	θ_{min} (°)	Max Deg. (dB)
2025-03-13	10:26	3.3	1.399	0.19
2025-03-14	10:23	8.4	1.007	0.48
2025-03-15	10:22	10.6	0.615	2.84
2025-03-16	10:21	11.5	0.223	6.31
2025-03-17	**10:21**	**11.6**	**0.170**	**6.59**
2025-03-18	10:21	10.8	0.562	3.37
2025-03-19	10:21	8.8	0.954	0.61
2025-03-20	10:23	4.5	1.346	0.21

Specifically, when the solar center approached the boresight ($\theta_{min} \approx 0.170°$), **6.59 dB** was recorded, representing a greater impact compared to Scenario 1.

Consequently, these results confirm that while increasing the antenna aperture reduces the frequency and duration of Sun Outage events, the instantaneous impact on the communication link becomes more severe during occurrence.

B. Discussion and Implications

The maximum C/N degradation of 6.59 dB derived from this simulation can affect link availability. Depending on the link fading margin, this may result in Link Outage, a downgrade in Adaptive Coding and Modulation (ACM) levels, or an increased Packet Loss Rate (PLR).

Unlike stochastic weather-related attenuation such as heavy snow or storms, Sun Outage is a deterministic phenomenon based on celestial mechanics. Since the timing, duration, and intensity of interference can be precisely predicted, the results enable proactive network management strategies.

To maintain Quality of Service (QoS), proactive measures can be implemented based on these predictions. Such measures include rescheduling the transmission of critical data during predicted interference intervals or pre-activating detouring paths.

Finally, this study confirms that geometric collinear alignment in high-gain antenna systems is the primary cause of

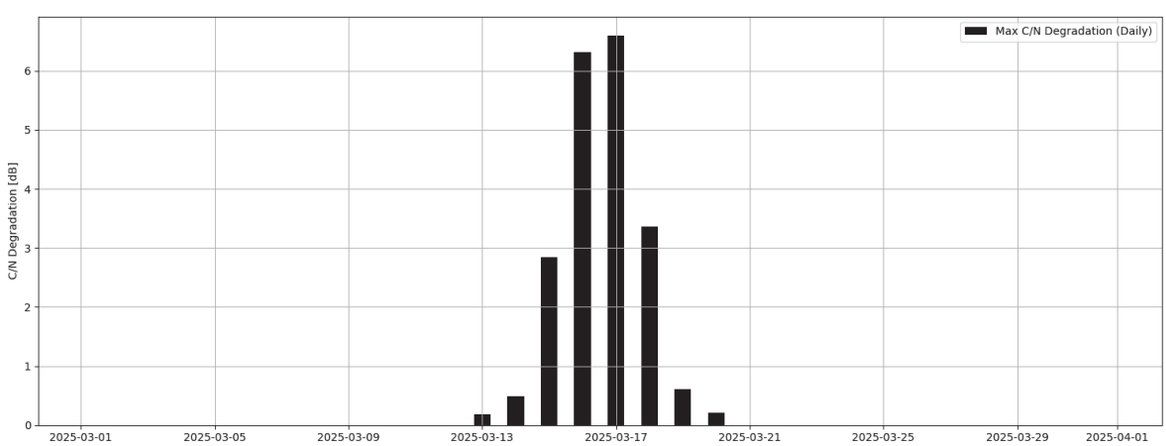


Fig. 9. Daily maximum C/N degradation for Scenario 2 (2.4 m antenna).

Sun Outage in high-latitude terrestrial networks. The proposed analysis framework is applicable to other communication systems where similar geometric conditions are met, such as Non-Geostationary Orbit (NGSO) satellite networks.

V. Conclusion

This paper demonstrates that the Sun Outage phenomenon, which has primarily been discussed in the context of satellite communications, can also serve as a significant interference factor in high-latitude terrestrial microwave networks. The unique geometric interference conditions resulting from the combination of low solar trajectories characteristic of high latitudes and the horizontal directivity of terrestrial antennas were analyzed. Furthermore, a quantitative analysis framework based on coordinate transformation and numerical integration was proposed to evaluate this impact. Simulation results confirmed a maximum C/N degradation of 6.59 dB, indicating a potential threat to link availability. In addition, the technical trade-off relationship, in which increasing the antenna aperture leads to higher interference intensity but reduces its duration, was quantitatively verified.

In polar environments where power resources are extremely limited, overcoming interference by simply increasing transmission power is practically unfeasible. However, as Sun Outage is a deterministic phenomenon governed by celestial mechanics, the timing and intensity of the interference can be precisely predicted. Therefore, this study suggests that proactive avoidance strategies and prediction-based network management are the most effective measures for ensuring the reliability of polar communication networks.

The analytical model proposed in this study is applicable to various systems sharing similar geometric constraints, such as Non-Geostationary Orbit (NGSO) satellite networks and other high-gain directional communication infrastructure. Based on the established prediction model, future research will focus on developing a power-efficient routing mechanism that maximizes energy efficiency by rerouting communication paths during predicted interference intervals.

Acknowledgment

This research was supported by the Korea Institute of Marine Science & Technology Promotion (KIMST) grant funded by the Ministry of Oceans and Fisheries (RS-2021-KS211509 and RS-2022-KS221571).

References

[1] Telecommunications Technology Association (TTA), "Extreme cold region data acquisition system – part 1: Overview and functional requirements," Telecommunications Technology Association (TTA), Standard TTAK.KO-06.0609-Part1, Dec. 2024.

[2] M. de Witt, C. Chung, and J. Lee, "Mapping renewable energy among antarctic research stations," *Sustainability*, vol. 16, no. 1, p. 426, 2024.

[3] G. Muratore, T. Giannini, and D. Micheli, "Solar radio emission as a disturbance of radiomobile networks," *Scientific Reports*, vol. 12, p. 9324, 2022.

[4] M. Ishii, J. Berdermann, B. Forte *et al.*, "Space weather impact on radio communication and navigation," *Advances in Space Research*, 2024, in press.

[5] V. K. Srivastava and P. Mishra, "Sun outage prediction modeling for earth orbiting satellites," *Applied Sciences (Switzerland)*, vol. 5, pp. 545–552, 2022.

[6] J. Jin, L. Shang, Z. Yang, H. Wang, and G. Li, "A local pre-rerouting algorithm to combat sun outage for inter-satellite links in low earth orbit satellite networks," *Applied Sciences*, vol. 14, no. 4, p. 1625, 2024.

[7] International Telecommunication Union (ITU-R), "Impact of interference from the sun into a geostationary-satellite orbit fixed-satellite service link," International Telecommunication Union (ITU), Tech. Rep. Recommendation ITU-R S.1525, 2001.

[8] K. S. Anderson, C. W. Hansen, W. F. Holmgren, A. R. Jensen, M. A. Mikofski, and A. Driesse, "pvlib python: 2023 project update," *Journal of Open Source Software*, vol. 8, no. 92, p. 5994, 2023. [Online]. Available: https://doi.org/10.21105/joss.05994

[9] International Telecommunication Union (ITU-R), "A methodology to evaluate the impact of solar interference on geostationary (gso) broadcasting-satellite service (bss) link performance," International Telecommunication Union (ITU), Recommendation ITU-R BO.1506, 2000.

[10] International Telecommunication Union (ITU), "Appendix 30 – provisions for all services and associated plans and list for the broadcasting-satellite service," in *Radio Regulations*. International Telecommunication Union (ITU), 2020, rev. WRC-19.

[11] Google LLC. (2024) Google earth pro (version 7.3.6). Accessed: Mar. 2025. [Online]. Available: https://www.google.com/earth/versions/

Deep Learning-Based Clustering Protocol for Digital Twin–Enabled IoT-MANETs

Junaedi Adi Prasetyo*, Ridho Hendra Yoga Perdana †, Yushintia Pramitarini ‡, and Axel Sikora§
* Department of Business and Informatics, Politeknik Negeri Banyuwangi, Banyuwangi 68461, Indonesia
† School of Information and Communication Engineering, Chungbuk National University, Cheongju 28644, Republic of Korea
‡ Institute of Industrial Technology, Korea University, Sejong 30019, Republic of Korea
§ Institute of Reliable Embedded Systems and Communication Electronics, Hochschule Offenburg, Germany
Emails: *junaedi.prasetyo@poliwangi.ac.id, †perdanarhy@ieee.org, ‡shintia88@korea.ac.kr, §axel.sikora@hs-offenburg.de

***Abstract*—This paper proposes a deep learning–based clustering (DLC) protocol for digital twin–enabled IoT-MANETs to ensure high quality of service (QoS), addressing the challenges posed by highly dynamic topologies, limited battery capacity, and fluctuating link conditions that often lead to unstable and unreliable network performance. To address these challenges, the proposed framework integrates deep learning (DL) with a digital twin, enabling real-time prediction of clustering decisions using both physical and network layer information from real and simulation measurements. The formulates a weight sum cross-layer function maximization problem that jointly considers physical-layer features such as node mobility, remaining battery energy, received signal strength indicator (RSSI) from real measurements, and channel quality together with network-layer metrics such as connectivity. Cluster-head (CH) selection is performed by the DL model using both residual energy and mobility constraints, ensuring stable and long-lasting cluster formation. In parallel, the digital twin maintains a continuously synchronized virtual representation of the IoT-MANET, enabling the learning model to update its predictions using real-time cross-layer data such as mobility, energy levels, channel quality, and connectivity. This integration enhances decision accuracy and improves overall network stability. Simulation results demonstrate that the proposed DLC protocol achieves low control overhead, high cluster stability, and strong scalability in diverse IoT-MANET scenarios, outperforming LEACH and Lowest-ID clustering schemes.**

***Index Terms*—Clustering, Internet of Things,**

I. Introduction

Mobile ad hoc networks (MANETs) represent a decentralized class of wireless networks composed of mobile nodes that communicate directly without fixed infrastructure. Each node is equipped with sensing, processing, and wireless communication modules, enabling seamless device-to-device coordination over highly dynamic links [1]. In many modern deployments, MANET nodes operate as part of the Internet of Things (IoT), where heterogeneous devices such as sensors, wearable systems, and smart machines collaborate autonomously to collect, process, and exchange data; this integration enhances flexibility and situational awareness but simultaneously intensifies challenges related to mobility, energy constraints, and scalable topology management. Recently, digital twin (DT) technology has emerged as a promising tool to mirror real-time network states in a virtual environment, offering new opportunities for predictive, data-driven clustering and improved network management in dynamic IoT-MANETs. MANETs are widely applied in scenarios such as environmental monitoring, public safety missions, battlefield operations, and disaster response, where rapid deployment, flexibility, and autonomous operation are essential [2]. However, the inherent mobility of nodes, energy constraints, and environmental variability often lead to unstable routing paths and frequent link disruptions. To address these issues, clustering has emerged as an effective mechanism to enhance connectivity, reduce routing overhead, and maintain network robustness under dynamic and energy-limited conditions [3], [4].

A. State of the art

Enhancing clustering stability and efficiency in MANETs remains a central challenge due to node mobility, fluctuating channel conditions, and limited battery capacity. In [5], a clustering strategy based on the K-means algorithm was proposed to mitigate mobility and energy issues; however, its performance depends heavily on the predefined number of clusters, making it unsuitable for highly dynamic network topologies. A 3D clustering approach presented in [6] further demonstrated that the spatial distribution and density of nodes significantly influence cluster formation quality, stability, and energy usage. Another method in [1] employed non-uniform clustering, where cluster-head (CH) election adapts to node-level features such as residual energy. Although this enhances communication efficiency, nodes positioned near the base station (BS) may suffer accelerated energy depletion due to uneven traffic loads. Classic clustering schemes such as LEACH [7] introduced probabilistic CH rotation to balance energy consumption, but its periodic re-election process makes it highly unstable in mobile scenarios. Likewise, Lowest-ID–based clustering [8] relies solely on node identifiers for CH selection, resulting in biased and energy-imbalanced cluster structures that degrade quickly under mobility. While traditional TCP/IP layering simplifies system design, the independence of network layers restricts adaptability in fast-changing MANET environments. As a result, several studies

have investigated cross-layer clustering methods, leveraging physical-layer indicators—such as signal strength, channel conditions, and mobility patterns to enhance CH selection [9]. Although cross-layer integration improves clustering decisions, it also introduces additional computational burden and requires continuous real-time optimization, which is difficult to sustain in resource-constrained MANETs.

Deep learning (DL) has emerged as an alternative for intelligent clustering, offering the ability to learn complex relationships among mobility, energy, and channel parameters [**?**]. However, existing DL-based clustering solutions have several disadvantages. They typically require large labeled datasets for training, consume substantial computational resources for frequent model updates, and rely on offline training processes that may fail to generalize when network conditions shift rapidly. Furthermore, decentralized MANET nodes often lack the processing power needed to execute or retrain DL models in real time, limiting practical deployment.

B. Motivation and Main Contribution

Most existing studies on MANET clustering rely on traditional methods such as K-means, Lowest-ID, and LEACH, or use cross-layer optimization and deep learning independently. However, these approaches struggle with mobility, energy constraints, and real-time adaptability. Emerging digital twin (DT) technology offers new opportunities for dynamic network representation, yet its integration with deep learning–based clustering remains largely unexplored. Next-generation IoT-MANETs therefore require a unified framework that combines cross-layer information, DL prediction, and DT synchronization. This raises several key research questions: 1) *How can mobility, energy, RSSI, and channel quality be jointly learned by a deep model for accurate CH prediction?* 2) *How can a digital twin efficiently mirror real-time node states to support low-overhead, adaptive clustering?* 3) *How can clustering remain stable under high mobility and realistic mobility models?* 4) *How can control overhead be minimized while outperforming traditional schemes such as LEACH and Lowest-ID?*

Motivated by these gaps, this work develops a digital twin–enabled deep learning clustering framework for IoT-MANETs, aiming to achieve stable cluster formation, reduced overhead, and improved adaptability to dynamic network conditions. The main contributions of the paper are as follows:

- We propose a novel deep learning–based clustering (DLC) protocol for digital twin–enabled IoT-MANETs, where the DT continuously mirrors real-time network states and assists the learning model in predicting optimal clustering decisions. The proposed framework enhances network stability and significantly reduces control overhead during cluster formation.
- We formulate a weighted cross-layer cost maximization problem that jointly considers physical-layer features such as node mobility (position and speed), remaining battery energy, received signal strength indicator (RSSI) from real measurements, and channel quality together with network layer metrics such as connectivity.
- To address the formulated problem, the proposed DLC protocol employs a deep learning model to accurately predict the optimal CH. The CH is selected using both residual-energy criteria and mobility constraints (velocity threshold), resulting in improved connectivity, longer link lifetime, and more stable cluster formation.
- Extensive simulations demonstrate that the proposed DLC protocol supported by a continuously updated digital twin, forms highly stable clusters under diverse node densities and mobility patterns. The results show significant improvements in cluster stability, link lifetime, and control overhead compared with LEACH and Lowest-ID clustering schemes.

II. System Model

In this paper, we introduce a deep learning–based clustering protocol designed for digital twin–enabled IoT-MANETs as illustrated in Fig. 1 to enhance network stability and connectivity. Unlike traditional clustering schemes, the proposed

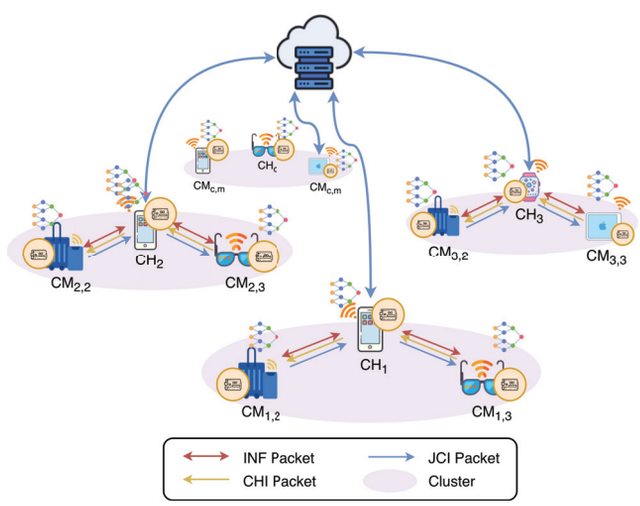


Fig. 1. The Concept of the Proposed DLC protocol

framework first selects the optimal cluster-head (CH) using a deep learning model trained inside the digital twin, and then performs cluster formation. This top–down approach significantly reduces control overhead because the network does not need to identify cluster members before choosing the CH. The proposed protocol formulates a weighted cross-layer cost function that integrates physical-layer information such as node mobility (position and speed), remaining battery energy, received signal strength (RSSI), and channel quality with network-layer information such as connectivity. The mobility information is derived from hardware modules, where each node obtains its real-time position and velocity through an onboard GPS module, while RSSI and energy values are directly measured from the device's communication and sensing interfaces. CH candidates are selected using the deep learning model by jointly evaluating residual energy,

mobility constraints (e.g., velocity threshold), and connectivity, ensuring long-term stability and prolonged link lifetime. In contrast to conventional clustering protocols that form clusters first and select a CH later—which often results in considerable control overhead due to repeated member discovery and dissemination—the proposed approach is more efficient. By performing CH selection first and allowing nodes to join the cluster afterward, the protocol requires only a small number of control packets, thereby reducing signaling overhead and improving scalability.

As shown in Fig. 1, we consider a dynamic IoT-MANET environment comprising L mobile nodes M forming a set of clusters $\mathcal{K} = \{\mathsf{C}_k\}_{k=1}^{K}$, where node movement follows the RWP mobility model. Each node is characterized by its position $\mathsf{P} = (x_l, y_l, z_l)$, speed S, movement direction D, and remaining battery energy B, all of which are continuously mirrored by the digital twin to support real-time and predictive clustering decisions. Furthermore, all parameters are measurement from real environment from the designed hardware which can be illustrated as Fig. 2.

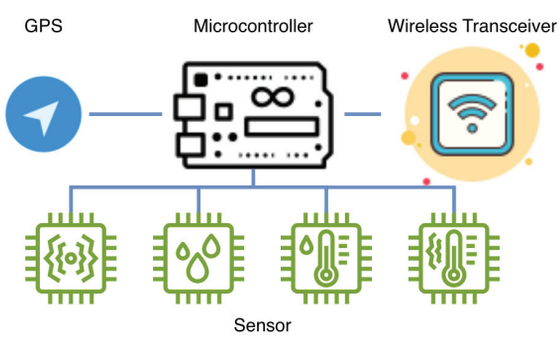


Fig. 2. The Architecture of the Proposed Hardware Nodes

The Fig. 2 illustrates the proposed hardware architecture used to measure real-world physical parameters in an IoT-MANET environment. The architecture consists of two main components: a navigation/positioning module on the left and a wireless communication module on the right. These two modules interface with a set of distributed sensor nodes, each equipped with dedicated sensing hardware. On the sensing side (bottom), multiple sensor chips are deployed to capture diverse environmental parameters such as vibration, humidity, temperature, and thermal variations. Each sensor module represents a physical-layer parameter that contributes directly to the digital twin and deep learning–based clustering process. The sensor readings including GPS position, mobility indicators, energy, signal measurements, and environmental conditions—are transmitted to the wireless communication module. This module ensures reliable data transfer to the network layer and synchronizes these measurements with the digital twin for real-time analysis. Overall, the architecture demonstrates how hardware-based sensing (via GPS and environmental sensors) and wireless connectivity work together to provide accurate, real-time network and environmental information that supports clustering decisions in the proposed deep learning–driven protocol.

A. The Proposed DLC Protocol Algorithm

In this subsection, we describe in detail the proposed DLC protocol. The protocol leverages a cross-layer design where physical-layer measurements—such as GPS-based mobility information, remaining battery energy, RSSI, and channel quality and network-layer metrics such as connectivity are jointly processed through a weighted cost model within the digital twin. Cluster-head selection is then performed by the deep learning model, which evaluates residual energy and mobility constraints (e.g., velocity threshold) to identify the most stable and sustainable CH candidate. Based on these combined features, the clustering and CH election procedures proceed through the following steps:

- **Step 0: Initialization**
  The real measurement and simulation begins with each node turning on and operating independently.
- **Step 1: Dissemination of Node Information**
  Each node M_l periodically measures its information, such as its position and remaining battery energy. Additionally, the velocity (speed and direction) is generated by following the mobility model. Then, M_l generates and broadcasts information (INF) packets to its neighbor node. The `INF` packet contains the following fields:
  $$[\mathsf{Tpe}, \mathsf{S_{ID}}, \mathsf{D_{ID}}, \mathsf{Pos}, \mathsf{E}, \mathsf{V}]$$
  where Tpe represents packet type, $\mathsf{S_{ID}}$ represents sender node ID, $\mathsf{D_{ID}}$ represents destination node ID, Pos, $\mathtt{E}$, and V represents GPS position, remaining battery energy, and velocity of l-th node, respectively.
  - if M_l receives `INF` packet, then goes to **Step 2**.
  - Otherwise, M_l waits receives `INF` packet.
- **Step 2: CH Election**
  To improve the stability of each cluster, the CH is selected from the candidate nodes based on the highest remaining energy, while also satisfying predefined mobility constraints. In particular, the node's velocity must lie within specified upper and lower thresholds. This constraint can be mathematically represented as:
  $$l^* = \arg \max_{l \in \mathsf{NB}_l^* \cup \{i\}} \{\mathsf{E}_l\} \tag{1a}$$
  $$\text{s.t} \quad \mathsf{V}^{\mathsf{L}} \leq \mathtt{V}_l \leq \mathsf{V}^{\mathsf{U}}, \tag{1b}$$
  where V^{U} and V^{U} represent the lower and upper bounds of velocity, with a predefined value, representing the allowable range of velocities for candidate nodes.
  - if $l = l^*$, M_l is elected as CH, then go to **Step 3**.
  - Otherwise, M_l waits receives `CHI` packet and go to **Step 4**.
- **Step 3: Dissemination of CH Information**
  When M_l becomes the CH, the M_l, or we can call CH_c, generates and broadcasts the `CHI` packet to be announced to its neighbor nodes. The `CHI` packet contains:
  $$[\mathsf{Tpe}, \mathsf{S_{ID}}, \mathsf{D_{ID}}, \mathsf{Pos}, \mathsf{V}, \mathsf{E}, \mathsf{Stat}]$$

where stat denotes the status CH of the node. Then, go to **Step 7**.

- **Step 4: Decision of Gateway**
  When node M_l receives the CHI packet, go to **Step 4.1**. Otherwise, M_l waits until receives CHI packet.

  - **Step 4.1:** If M_l receives more than one CHI packet, M_l becomes a gateway node, and then go to **Step 5**; Otherwise, M_l becomes CM then go to **Step 6**.

- **Step 5: Cluster Forming**
  When node M_l receives CHI packet from the candidate CH nodes, the node predicts the optimal values of weight (w_1, w_2, w_3) and cost Ct values. The objective of this paper is to form a cluster by selecting the best CH by maximizing cost value to improve connectivity by optimizing weight value subject to the remaining energy E, channel capacity SE, and cosine similarity CSS, which can be formulated as

$$\max_{w_1,w_2,w_3} \ \mathsf{Ct}_l = w_1\mathsf{E}_l + w_2\mathsf{SE}_l + w_3 k, \tag{2a}$$

$$\text{s.t } w_1 + w_2 + w_3 = 1, \tag{2b}$$

$$w_1 > 0, w_2 > 0, w_3 > 0, \tag{2c}$$

$$\mathsf{E}_l \geq \mathsf{E}^{\mathsf{th}}, \tag{2d}$$

$$\mathsf{SE}_l \geq \mathsf{SE}^{\mathsf{th}}, \tag{2e}$$

$$\mathsf{CSS}_l \leq \mathsf{CSS}^{\mathsf{th}}, \tag{2f}$$

where constraints (2b) and (2c) represent weight values that must be equal to one and have a positive value, while constraints (2d), (2e), and (2f) represent the remaining battery, spectral efficiency SE and cosine similarity must be greater than predefined threshold $(\mathsf{E}^{\mathsf{th}}, \mathsf{SE}^{\mathsf{th}}, \mathsf{CSS}^{\mathsf{th}})$. From problem (2), it is clear that the problem (2) is non-convex since both the objective function and constraint are nonlinear, requiring multiple iterations to solve it by using alternative optimization (AO), which can be studied in []. In this paper, a DL-based method is proposed to solve the problem (2) without requiring iterations, as explained in Section III. From problem (2), the SE_l can be obtained as

$$\mathsf{SE}_l = \log_2(1 + \mathrm{RSSI}_l), \quad [\mathrm{bps/Hz}]. \tag{3}$$

While the cosim similarity between node l and its neighbor nodes l', which can be expressed as

$$\mathsf{CoSim}(l, l') = \frac{\sum_{l'}^{L^{\mathsf{NB}}} \sum_i^I \overrightarrow{V}_{l,i} \overrightarrow{V}_{l',i}}{\sqrt{\sum_i^I \overrightarrow{V}_{l,i}^2} \sqrt{\sum_i^I \overrightarrow{V}_{l',i}^2}}, \tag{4}$$

where $\overrightarrow{V}_{l,i}$ is the i-th vector information of node l. Each node $\overrightarrow{V}_{l,i}$ is related to a mobility vector information metric value, such as velocity and position. Furthermore, the distance between node l and l' can be calculated by using

$$\mathsf{d}(l, l') = \sqrt{(x_l - x_{l'})^2 + (y_l - y_{l'})^2 + (z_l - z_{l'})^2}, \tag{5}$$

Thus, each node selects the following CH candidate by the criteria, which can be formulated as

$$l^* = \arg\max_l \ \{\mathsf{Ct}_{l'}\}, \tag{6a}$$

$$\text{s.t} \quad d(l, l') \leq \bar{d}, \tag{6b}$$

where constrain (6b) represent the distance between node should less than maximum distance $\bar{d}$. Then, go to step **6**.

- **Step 6: CM's Table Updates**
  After M_l select the optimal CH, then M_l updates the CM table as shown in Table I where $\mathsf{CM_{ID}}$, $\mathsf{CH_{ID}}$, and St denotes the CM ID and CH ID, status of the its node (CM or GW) and send a JCI packet which contains the following fields:

$$[\mathsf{Tpe}, \mathsf{S_{ID}}, \mathsf{D_{ID}}, \mathsf{Stat}]$$

  Then, go to step **7**.

TABLE I
CM TABLE OF THE DLC PROTOCOL.

$\mathsf{CH_{ID}}$	$\mathsf{CM_{ID}}$	St

- **Step 7: CHs Table Updates**
  When M_l who is the CH candidate receives the JCI packet from CM, then M_l updates the CH table as shown in Table II.

TABLE II
CH TABLE OF THE DLC PROTOCOL

$\mathsf{CH_{ID}}$	$\mathsf{CM_{ID}}$	Stat

III. DEEP LEARNING DESIGN

In this section, a deep neural network (DNN) model is designed to predict the optimal candidate CH and the corresponding cost value Ct_l used in the proposed DLC protocol. The DNN architecture consists of an input layer, multiple hidden layers, and an output layer, as illustrated in Fig. 3. The input layer receives GPS-based position information,

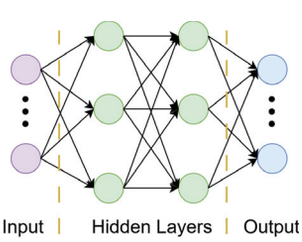


Fig. 3. The architecture of the proposed DNN model

remaining battery energy, RSSI, channel quality, and node speed, arranged in an input dimension of $\mathsf{NB} \times 7$. The network includes three hidden layers with 64, 128, and 64 neurons, respectively, while the output layer has a dimension of 2×1, representing the predicted CH score and Cs value. We employ an offline training approach with 10,000 datasets, where network weights and biases are optimized using the

backpropagation algorithm [10]. Once offline training is completed, the computational burden shifts to the online inference stage, enabling lightweight and efficient real-time execution of the proposed DLC protocol.

IV. Performance Evaluation

In this section, we evaluate the performance of the proposed DLC protocol under various mobility scenarios with different maximum node speeds using the NS-3 simulator. All IoT-MANET nodes were randomly deployed in a $1,000\text{m} \times 1,000\text{m}$ area, and their movement followed a random way point (RWP) and reference point group mobility (RPGM) models. Realistic physical-layer parameters, such as position, speed, and RSSI, were obtained through the hardware module (e.g., GPS and RF measurements) and synchronized with the digital twin during simulation. This setup enabled a realistic evaluation of how mobility and cross-layer dynamics influenced clustering stability, CH selection, and overall network performance. The simulation environment and parameters of this paper can be summarized in Table III. Besides, LEACH

TABLE III
Simulation Environment and Parameters

Parameters	Value
Microcontroller	Arduino Uno
Mobility model	RWP and RPGM [11]
Number of nodes	[30,50]
Range speed of node	[10:10:40] (km/h)
RSSI threshold	-80 dBm
Session length	5 s
Simulation area (3D)	1000 m × 1000 m × 0 m
Simulation MAC protocol	802.11a
Simulation time	100 s
Simulator	NS-3 Simulator
Tranceiver	nRF24L01

in [7] and Lowest-ID in [8] are also provides as benchmark schemes to compare with the proposed clustering protocol. Moreover, the LEACH and Lowest-ID protocols can be simulated with same simulation environment and parameters.

A. Performance Metric

To show the effectiveness of the proposed DLC protocol in handling the dynamic behavior of nodes in digital twin-enabled IoT-MANET environments, we evaluate its performance using three key stability metrics.

1) Number of cluster heads: which indicates the average number of CHs formed in each session and reflects how well the clustering structure is maintained under mobility.
2) Average number of cluster head changes: which measures how frequently CHs are replaced within a cluster, providing insight into the protocol's resilience to mobility-induced topology variations.
3) Control overhead: defined as the average number of control packets exchanged during the clustering process per session and per node, capturing the communication efficiency of the protocol in highly dynamic scenarios.

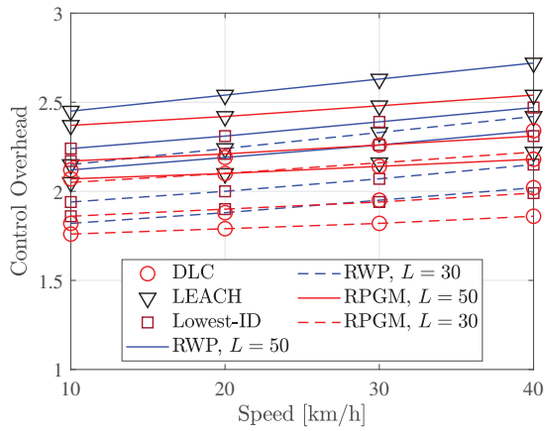


Fig. 4. Impact of variation of speed on the control overhead

B. Simulation Results

Fig. 4 presents the control overhead as a function of node speed. As can be observed, the control overhead remained relatively constant across different speeds for both $L = 30$ and $L = 50$. This behavior indicates that DLC did not require additional control packets to maintain cluster structure, confirming that node mobility had minimal influence on cluster formation. When compared with benchmark schemes, DLC consistently achieved lower control overhead than both Lowest-ID and LEACH. Lowest-ID required more signaling because CH stability depended entirely on ID ordering rather than mobility or energy conditions, causing periodic updates when nodes moved. LEACH showed the highest overhead due to its randomized periodic CH re-selection, which triggers frequent control message broadcasts regardless of mobility. Furthermore, the RPGM mobility model outperformed RWP since its group-based movement kept nodes closer together, reducing unnecessary re-clustering events. This stability further strengthened DLC's ability to maintain low overhead, while LEACH and Lowest-ID remained more sensitive to mobility variations.

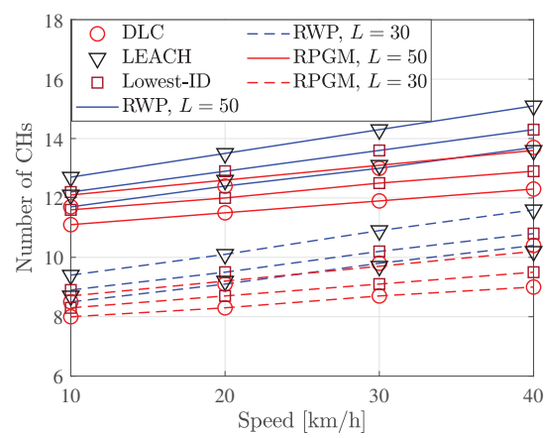


Fig. 5. Number of CHs versus variation of speed

Fig. 5 shows the number of CHs versus variation of node speed with $L = 30$ and $L = 50$. Overall, the DLC protocol maintained a stable CH count across all speeds, with only a slight increase as mobility increased. This minor variation suggests that DLC can preserve cluster organization even at higher speeds due to its deep learning–based CH prediction and cross-layer decision-making. In comparison, Lowest-ID produced a slightly higher number of CHs, as its identifier-based selection does not adapt well to mobility or energy variations. Nodes with lower IDs repeatedly became CHs even when not ideal, leading to additional clusters forming when mobility increased. LEACH produced the highest number of CHs, as its probabilistic CH rotation frequently elects new CHs regardless of network conditions, causing more clusters to emerge at higher speeds. Once again, RPGM outperformed RWP by maintaining stable group movement, reducing fragmentation, and requiring fewer CH adjustments. This advantage aligned well with DLC's learning-based CH decision, resulting in the most stable cluster structure among all tested protocols.

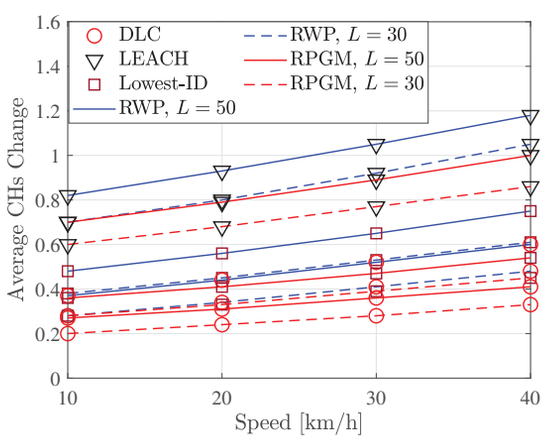


Fig. 6. Average CH changing versus variation of speed

Figure 6 illustrates the impact of the variation of speed on the average number of CH changes. For the proposed DLC protocol, CH changes remained consistently close to zero across all speeds. This indicates that once a CH was elected, its role remained stable even under higher mobility, thanks to the deep learning model's use of mobility constraints, residual energy, and RSSI measurements. In contrast, Lowest-ID experienced moderate CH changes, as mobility caused nodes with the lowest IDs to periodically move out of range, forcing new CH selections. LEACH showed the highest number of CH changes, due to its built-in periodic CH rotation mechanism that triggers re-election even without mobility. The RPGM model again offered higher stability than RWP, as nodes moved cohesively in groups, reducing the likelihood of CH displacement. This stability particularly benefited the DLC protocol, allowing it to maintain near-zero CH changes, whereas Lowest-ID and LEACH were much more affected by the randomness of RWP mobility.

V. Conclusions

This paper proposed a DLC protocol for digital twin–enabled IoT-MANETs to address the challenges of highly dynamic topologies, energy constraints, and fluctuating wireless conditions. The framework integrated deep learning with a digital twin to predict CH candidates using cross-layer information, including mobility, residual energy, RSSI, channel quality, and connectivity. A weighted cross-layer cost function was formulated to guide CH selection, while the digital twin provided a synchronized virtual environment that continuously updated the learning model with real-time measurements. This integration ensured stable cluster formation, low control overhead, and robust adaptation to network dynamics. Simulation results showed that the proposed DLC protocol achieved high cluster stability, reduced overhead, and strong scalability, outperforming classical clustering schemes such as LEACH and Lowest-ID.

References

[1] A. Amalia, Y. Pramitarini, R. H. Y. Perdana, K. Shim, and B. An, "Energy Efficient Multicast Routing Protocol Using FL-Based Optimal Route Selection in IoT-Enabled MANETs With RIS and CF-mMIMO," *IEEE Internet Things J.*, vol. 12, no. 22, pp. 47 588–47 606, 2025.

[2] E. Yaacoub and M. S. Alouini, "A Key 6G Challenge and Opportunity-Connecting the Base of the Pyramid: A Survey on Rural Connectivity," *Proc. IEEE*, vol. 108, no. 4, pp. 533–582, 2020.

[3] Y. Pramitarini, R. H. Y. Perdana, T.-n. Tran, K. Shim, and B. An, "A Hybrid Price Auction-Based Secure Routing Protocol Using Advanced Speed and Cosine Similarity-Based Clustering against Sinkhole Attack in VANETs," *Sensors*, vol. 22, no. 15, pp. 1–22, 2022.

[4] A. Ahmad, M. A. Pasha, S. Masud, M. Schappacher, A. Belhouichet, and A. Sikora, "Multi-hop routing implementation in spatially distributed wireless sensor networks for large-scale environmental monitoring applications," in *Proc. 2019 10th IEEE Int. Conf. Intell. Data Acquis. Adv. Comput. Syst. Technol. Appl. IDAACS 2019*, vol. 1. IEEE, 2019, pp. 111–116.

[5] A. Raza, M. F. Khan, M. Maqsood, B. Haider, and F. Aadil, "Adaptive k-means clustering for Flying Ad-hoc Networks," *KSII Trans. Internet Inf. Syst.*, vol. 14, no. 6, pp. 2670–2685, 2020.

[6] Y. Pramitarini, R. H. Y. Perdana, K. Shim, and B. An, "Federated Blockchain-Based Clustering Protocol for Enhanced Security and Connectivity in FANETs With CF-mMIMO," *IEEE Internet Things J.*, vol. 12, no. 10, pp. 14 016–14 035, 2025.

[7] A. Jayapalan and S. Anandan, "LEACH-ERA: An Energy-Efficient and Reliable Cluster Head Selection Algorithm for Prolonged Wireless Sensor Network Lifetime," in *2024 Int. Conf. Smart Electron. Commun. Syst. ISENSE 2024*. Kottayam, India: IEEE, 2024, pp. 1–6.

[8] D. Gavalas, G. Pantziou, C. Konstantopoulos, and B. Mamalis, "Lowest-ID with adaptive ID reassignment: A novel mobile ad-hoc networks clustering algorithm," in *2006 1st Int. Symp. Wirel. Pervasive Comput.*, vol. 2006, Phuket, Thailand, 2006, pp. 1–5.

[9] A. Amalia, Y. Pramitarini, R. H. Y. Perdana, K. Shim, and B. An, "A Cross-Layer FL-Based Clustering Protocol to Support Multicast Routing in IoT-Enabled MANETs With CF-mMIMO," *IEEE Access*, vol. 13, no. January, pp. 3881–3899, 2025.

[10] R. H. Y. Perdana, T.-V. Nguyen, and B. An, "Adaptive User Pairing in Multi-IRS-aided Massive MIMO-NOMA Networks: Spectral Efficiency Maximization and Deep Learning Design," *IEEE Trans. Commun.*, vol. 71, no. 7, pp. 4377–4390, 2023.

[11] Y. Pramitarini, R. H. Y. Perdana, K. Shim, and B. An, "DLSMR: Deep Learning-Based Secure Multicast Routing Protocol against Wormhole Attack in Flying Ad Hoc Networks with Cell-Free Massive Multiple-Input Multiple-Output," *Sensors*, vol. 23, no. 18, pp. 1–28, 2023.

A Study on a Privacy-Preserving Platform for Protecting and Using Medical Data

Muneeb MUHAMMAD
Dept. of Computer Engineering
Sang-Ji University
Wonju, Republic of Korea
2019015001@sj.sangji.ac.kr

Oudone PHENGKHAMLAR
Dept. of Computer Engineering
Sang-Ji University
Wonju, Republic of Korea
2025015105@sangji.ac.kr

Kwang-Man Ko
Dept. of Computer Engineering
Sang-Ji University
Wonju, Republic of Korea
kkman@sangji.ac.kr

***Abstract*—Medical data (EHR, images, and vital signs) is essential for healthcare AI, but it is hard to share because of privacy laws and strict hospital policies. This creates a trade-off: strong models need diverse data, while sharing sensitive records increases privacy and compliance risks. We propose an end-to-end platform that connects (1) differentially private (DP) synthetic medical data generation using diffusion models with DP-SGD, (2) multi-hospital federated learning (FL) for Non-IID settings so data stays local during training and inference, and (3) permissioned blockchain governance for consent, access control, and auditable life-cycle provenance. We present a prototype evaluation on public datasets and report key measurements for utility (AUROC/F1), privacy risk (membership inference), FL fairness/stability, and audit latency. Results suggest the integrated design can support practical medical AI development with improved privacy protection and accountability.**

***Index Terms*—Synthetic data, Differential privacy, Federated learning, Non-IID, Blockchain, Healthcare AI**

I. Introduction

Medical data is a valuable resource for precision medicine and medical AI, including EHR, lab tests, vital signs, and medical images. With recent progress in deep learning and LLMs, AI performance in diagnosis support, prediction, and clinical workflow optimization has improved. However, in real hospitals, data sharing and reuse are still heavily constrained by strict regulations (e.g., PIPA/GDPR/HIPAA) and internal security policies.

Medical data also contains highly sensitive information (health status, disease history, genetics). This makes cross-institution collaboration difficult, especially for rare diseases, minority groups, or long-term follow-up studies. Practical barriers such as Non-IID data across hospitals, different formats, interoperability issues, and legal/ethical accountability further limit "sufficient, diverse, and safe" medical data utilization.

To address these challenges, three technical directions are often discussed: (i) *synthetic data* to reduce direct exposure of raw data, (ii) *federated learning (FL)* to train models without exporting patient data, and (iii) *blockchain governance* to provide immutable logs and policy-based control. However, using these technologies separately is often not enough. Synthetic data may still leak privacy; FL can be unstable under Non-IID settings and faces security risks; and blockchain solutions may only log access events without covering the full end-to-end life cycle.

Therefore, this paper proposes an end-to-end platform that connects: (1) diffusion-based DP synthetic medical data generation, (2) Non-IID-aware federated learning (e.g., clustering + FedProx + robust aggregation), and (3) permissioned blockchain governance (consent, access rules, and provenance logging) to support privacy-preserving medical AI across the full pipeline.

Goals. We aim to (a) enable safe data use when real data access is limited, (b) train strong models without sending raw data outside hospitals, and (c) keep auditable records for compliance (consent, access, model versions, and usage history).

Main contributions. Our contributions include: (i) DP diffusion-based synthetic data with an explicit privacy–utility trade-off, (ii) a robust FL design for Non-IID medical data, (iii) a permissioned blockchain layer for consent-aware governance and life-cycle provenance, and (iv) an illustrative prototype evaluation on public datasets.

Paper organization. Section II reviews related work; Section III describes the platform; Sections IV–VII present the system/threat model, prototype, and evaluation; and Sections VIII–X discuss limitations and future work.

II. Background and Related Work

A. Synthetic Medical Data and Privacy

Synthetic medical data is widely studied because it can support model development when direct access to real patient data is restricted by regulations and hospital policies. Early work mainly used GANs or VAEs to generate EHR-like tables or medical images, aiming to match the statistics and downstream performance of real data [1]. More recently, diffusion models have become popular because they often provide stable training and high-quality samples [2].

However, synthetic data is not automatically privacy-safe. Prior studies report that attackers may still infer whether a specific record was used for training (membership inference) [3]. Therefore, privacy-preserving training is often required. A common approach is Differential Privacy (DP), where DP-SGD clips per-example gradients and adds noise so that each training record has limited influence [4]. In this work, we adopt diffusion-based generation with DP-SGD and treat

privacy–utility as an explicit trade-off rather than assuming that "realistic" samples are safe.

B. Federated Learning in Healthcare and Security Issues

Federated Learning (FL) enables multiple hospitals to train a shared model while keeping raw data local [5]–[7]. Federated learning has also been studied for practical digital health settings [8]. This fits healthcare collaboration, but real deployments face two major difficulties. First, hospitals are typically Non-IID: patient populations, devices, and clinical protocols differ across sites, so simple FedAvg may be unstable and can harm some hospitals. Second, privacy and security threats still exist even when only gradients/weights are shared [6]. In addition, insider threats such as poisoning or label-flipping may degrade model quality or introduce backdoors.

To improve practical robustness, Non-IID-oriented optimizers such as FedProx can stabilize convergence [9]. Robust aggregation methods can reduce the impact of outlier or malicious updates, and DP-style protection can also be applied during local training. Our platform combines these ideas in a multi-hospital setting and connects training with governance and auditing.

C. Blockchain-based Governance and BCFL

Even if raw data stays inside hospitals, governance is still required: who is allowed to use which data or model, for what purpose, and how to prove compliance later. Blockchain has been used for consent management, access logging, and integrity protection in healthcare [10]. A remaining gap is full life-cycle governance: many systems do not manage the entire pipeline from synthetic data generation and FL training to model deployment and inference under policy constraints. To address this, our platform uses a permissioned blockchain to record key events end-to-end and introduces a lightweight trust score for each model version, combining performance, privacy-test results, and compliance checks.

III. Proposed Platform Architecture

The proposed platform connects three layers so that privacy protection and data use work together from start to finish:

- **Layer 1: Synthetic Data Layer** — create privacy-preserving synthetic medical data (diffusion + DP).
- **Layer 2: Federated Learning Layer** — train and use models across hospitals without sharing raw data.
- **Layer 3: Blockchain Governance Layer** — manage consent, access rules, and full life-cycle records (provenance and audit).

Workflow (overview). Hospitals keep raw data locally. They may generate DP synthetic data for safer testing or limited sharing. For real training, hospitals join FL rounds to build a global model. The blockchain records key events: consent, policies, training rounds, model versions, and inference requests.

A. Layer 1: Synthetic Data with Diffusion + Differential Privacy

Goal: generate synthetic data that is (a) useful for model development and evaluation and (b) safer for privacy than real data.

Why diffusion? A diffusion model learns data by adding noise and then learning to remove it step-by-step, often producing high-quality samples and stable training [2].

Why DP-SGD? DP-SGD reduces privacy leakage by limiting the influence of each training example (gradient clipping) and adding controlled noise before updating model parameters [4]. The amount of privacy protection can be tracked by a privacy budget (e.g., ϵ).

B. Layer 2: Federated Learning for Multi-Hospital Training (Non-IID)

Goal: train a strong medical AI model across multiple hospitals without sending raw data outside each hospital.

Core ideas used in this platform:

- Clustering by disease/cohort to reduce distribution gaps.
- FedProx to improve stability under Non-IID [9].
- DP-style protection during local training (optional, depending on policy).
- Robust aggregation to reduce the impact of outlier updates.

C. Layer 3: Blockchain Governance (Consent, Access, Provenance)

Goal: make data/model use transparent and auditable. The platform uses a permissioned blockchain (Hyperledger Fabric) to record key life-cycle events and enforce rules [10].

Purpose-based consent. Consent is stored as a structured template: what data, for what purpose, for how long, and under which limits. When a user requests access or model inference, the system checks whether the request matches valid consent.

Hybrid access control (RBAC + ABAC). RBAC gives base permissions by role (e.g., hospital researcher, regulator). ABAC adds extra conditions (e.g., data sensitivity, purpose, time). Every important action is logged with a digital signature.

D. Training Objective: Utility–Privacy Trade-off

A key challenge is balancing usefulness and privacy in synthetic data generation. We treat this as a multi-objective optimization problem that considers (i) diffusion reconstruction quality, (ii) downstream predictive utility, and (iii) privacy risk.

We define the overall objective as:

$$L_{\text{total}}(\psi, \theta) = \lambda_1 L_{\text{DDPM}}(\psi) + \lambda_2 L_{\text{task}}(\theta; \psi) + \lambda_3 L_{\text{privacy}}(\psi). \tag{1}$$

We denote the diffusion generator parameters by ψ and the downstream task model parameters by θ. The weights $(\lambda_1, \lambda_2, \lambda_3)$ control the utility–privacy trade-off.

IV. System and Threat Model

A. System model

We consider a cross-silo setting where multiple hospitals collaboratively build a global model. Each hospital keeps raw data locally and may (i) generate DP synthetic data for safer sharing or testing, and (ii) participate in FL rounds by sending model updates instead of raw data. A permissioned blockchain maintains policy and audit records shared among stakeholders (hospitals, data controllers, and auditors).

B. Threat model (prototype scope)

We focus on common privacy and integrity risks that motivate our design. For synthetic data, we consider membership inference attacks (MIA) that attempt to infer whether a patient's record was used for training [3]. For FL, we consider inference risks from shared updates and robustness issues such as abnormal or malicious client updates. The blockchain network is permissioned and authenticated; our goal is tamper-resistant audit logs and policy verifiability (not anonymity). Attacks outside the scope include physical compromise of hospital servers and fully malicious blockchain administrators.

V. Implementation Overview (Prototype)

A. Platform components

The prototype platform consists of three components: (i) a DP diffusion generator for synthetic EHR/image/signal data, (ii) an FL coordinator and hospital clients for cross-silo training, and (iii) a blockchain service (Hyperledger Fabric) for consent, policy checks, and event logging.

B. Logged events and provenance

We log the minimum set of events required for compliance and traceability: consent updates, access requests, synthetic data generation jobs, FL round metadata, model version registration, and inference requests. Provenance links datasets (real or synthetic), training rounds, and model versions to support later audits and reproducibility.

C. End-to-End Prototype Workflow and Audit Queries

To clarify how the platform works in practice, we describe a simple end-to-end prototype workflow. The goal is to show (i) how data and model actions are constrained by policy, and (ii) how an auditor can later verify compliance without accessing raw patient data.

Step 1: Consent and policy registration. A data controller registers a purpose-based consent template and organizational policies. Each consent specifies allowed purpose (e.g., research vs. clinical decision), validity period, and constraints such as model type and privacy requirements (e.g., DP budget range).

Step 2: Synthetic data generation (optional). If real data access is limited, a hospital can generate DP synthetic datasets for prototyping or limited sharing. The system records job metadata (dataset type, generator version, and privacy settings) so results can be traced to the exact generation conditions.

Step 3: Federated training across hospitals. Hospitals participate in FL rounds by training locally and sending model updates. The coordinator aggregates updates (e.g., FedProx + robust aggregation). For each round, the platform records the participant list, round number, and resulting model hash/version.

Step 4: Model release and controlled inference. A released model is registered with a version ID, validation metrics, and privacy-test summaries. For inference, a user submits a request with declared purpose and role attributes; the policy engine checks consent and access rules before allowing inference. Inference requests are logged as auditable events.

Audit queries. An auditor can later verify which consent/policies were active at the time of training or inference, which hospitals contributed to a model version, whether real vs. synthetic data was involved, and whether any required events are missing in the ledger.

VI. Prototype Configuration and Policy Enforcement

This section provides concrete configuration choices to make the prototype more reproducible and to clarify what is enforced by policy. While the exact values may vary by institution, reporting typical settings improves transparency and helps readers understand the end-to-end pipeline.

A. DP Synthetic Data Configuration (Illustrative)

We use DP-SGD for training the diffusion generator. Each training step applies per-example gradient clipping with norm bound C and adds Gaussian noise with multiplier σ. The privacy budget ϵ is tracked using a DP accountant [4]. In a practical deployment, the data controller can set an acceptable ϵ range based on risk tolerance. In our prototype, we treat DP strength as a configurable knob and report both utility and privacy risk (MIA success rate) to show the expected trade-off.

B. Federated Learning Protocol (Illustrative)

We simulate a cross-silo setting with multiple hospitals as clients. Non-IID is constructed by splitting patients so that each site has different label proportions and dataset sizes. Each FL round samples a subset of hospitals, performs local training for a small number of epochs, and returns model updates for aggregation. To stabilize training under heterogeneity, we include FedProx with proximal coefficient μ [9]. Robust aggregation is applied as a safety layer to reduce the impact of abnormal updates. In addition to global AUROC/F1, we report inter-site standard deviation as a fairness indicator and round-to-round variation as a stability indicator.

C. Policy Engine and On-chain Event Schema

The platform enforces purpose-based consent and hybrid access control (RBAC + ABAC). Each access or inference request includes: *(i)* requester role, *(ii)* declared purpose, *(iii)* resource type (dataset/model), and *(iv)* time window. A request is approved only if it matches active consent and organizational policies. To support audits, the blockchain records a concise event schema with: `eventType`, `actorID`, `resourceID`, `purpose`, `timestamp`, and `hashPointer` (off-chain metadata hash).

D. Trust Score (Lightweight Prototype)

To help stakeholders select a safe model version, we define a simple trust score that combines utility, privacy risk, and regulatory/policy compliance:

$$T(\mathbf{M}_v) = \alpha_1 \cdot \mathrm{AUROC}(\mathbf{M}_v) + \alpha_2 \cdot \big(1 - \mathrm{MIA}_{rate}(\mathbf{M}_v)\big) - \alpha_3 \cdot \mathrm{RegViolation}(\mathbf{M}_v). \tag{2}$$

Here, $\mathrm{AUROC}(\mathbf{M}_v)$ measures predictive utility, $\mathrm{MIA}_{rate}(\mathbf{M}_v)$ is the membership inference attack success rate (lower is better), and $\mathrm{RegViolation}(\mathbf{M}_v)$ is a normalized violation score (0 means fully compliant; higher means more violations). The weights $\alpha_1, \alpha_2, \alpha_3 \geq 0$ can be set by a governance committee. In our prototype, the trust score is stored together with the model version hash for auditability.

E. Why this section matters

Adding explicit configuration and policy details reduces ambiguity. Readers can see what is measured, what is enforced, and what is logged. This also clarifies that the evaluation is *prototype/illustrative*, while still following a realistic compliance-oriented workflow.

VII. Experiments and Evaluation

We present an illustrative prototype evaluation using public medical datasets to demonstrate three parts: (1) synthetic data usefulness and privacy, (2) FL performance under Non-IID, and (3) blockchain audit and latency.

A. Experimental environment and evaluation protocol

We use public datasets that cover different modalities: EHR/time-series (MIMIC-IV [11]), chest X-ray images (MIMIC-CXR [12]), and biosignals (PhysioNet [13]). We use representative tasks such as ICU mortality prediction, readmission prediction, and binary classification for imaging. All models are implemented in Python/PyTorch.

For cross-silo FL, we simulate multiple hospitals as clients and construct Non-IID settings by splitting data so that sites have different label distributions and different sample sizes. We report utility using AUROC and F1-score on a held-out real test set. We quantify privacy risk using membership inference attack (MIA) success rate (lower is better) [3]. For FL, we measure fairness and stability using inter-site performance deviation (standard deviation across hospitals) and round-to-round variation. For governance, we measure audit completeness and transaction latency. The Table I shows the experimental setup summary as a prototype for readrers understanding.

Table I: Experimental setup summary (prototype)

Category	**Description**
Datasets	MIMIC-IV (EHR/time-series), MIMIC-CXR (chest X-ray), PhysioNet (biosignals)
Prediction Tasks	ICU mortality prediction, readmission prediction, chest X-ray binary classification
Dev Language / Framework	Python, PyTorch
FL Simulation	PySyft-based cross-silo FL simulation environment
Synthetic Data Model	Diffusion-based generator + DP-SGD
Blockchain Environment	Hyperledger Fabric, Multi-organization setup with ordering and peer nodes
Evaluation Metrics	AUROC, F1-score, MIA success rate, audit completeness, transaction latency

B. Illustrative Results

Note: The following numbers are illustrative outputs from a prototype evaluation to demonstrate the measurement pipeline and expected trends. They should not be interpreted as final clinical evidence.

1) Real vs. DP synthetic data (illustrative): This Table II shows the comparison of real and sythetic data for performance and privacy metrics.

Table II: Performance and Privacy Metrics: Real vs. DP Synthetic Data (Illustrative)

Model Type	**Training Data**	**AUROC**	**F1**	**MIA (↓)**
Baseline	Real (MIMIC-IV)	0.98	0.91	0.72
Proposed	Synthetic (DP)	0.97	0.90	0.36

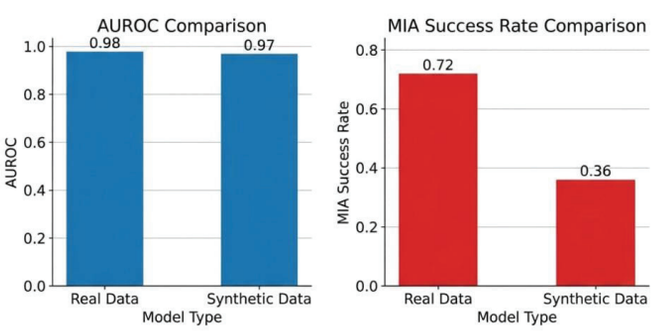


Figure 1: Performance and Privacy Comparison of Real vs. Synthetic Data Models (Illustrative)

2) FedAvg vs. FedProx under Non-IID (illustrative): To illustrate the comparison of FedAvg and FedProx under Non-IID scenario, we have presented this in Table III.

Table III: FedAvg vs. FedProx in a Non-IID Setting (Illustrative)

Alg.	**AUROC**	**F1**	**Inter-site Std. Dev.**	**Round Var. (Max–Min)**
FedAvg	0.93	0.86	0.081	0.12
FedProx	0.95	0.88	0.028	0.04

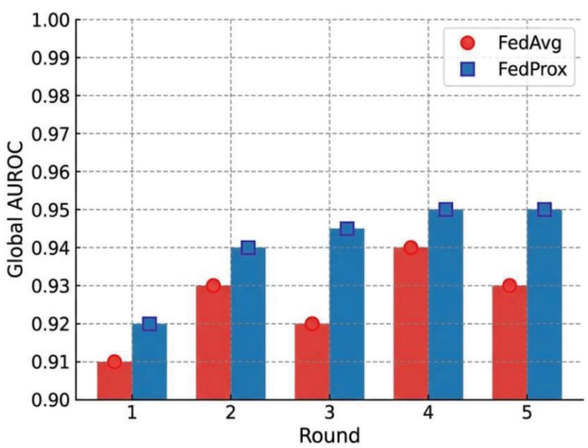


Figure 2: Round-wise Global AUROC Convergence under a Non-IID Setting (Illustrative)

3) Blockchain provenance management (illustrative):

Table IV: Blockchain-based Audit/Traceability Metrics (Illustrative)

Metric	**Value**	**Description**
Audit Completeness	100%	Missing rate of original ledger records among configured key events: 0%
Average Tx Latency	250 ms	Average time from event record request to commit completion
95th Percentile Latency	420 ms	Latency for the slowest top 5% of transactions
Logged FL Rounds	100 rounds	Metadata logged for all rounds
Logged Model Versions	20 versions	Version metadata stored at checkpoints

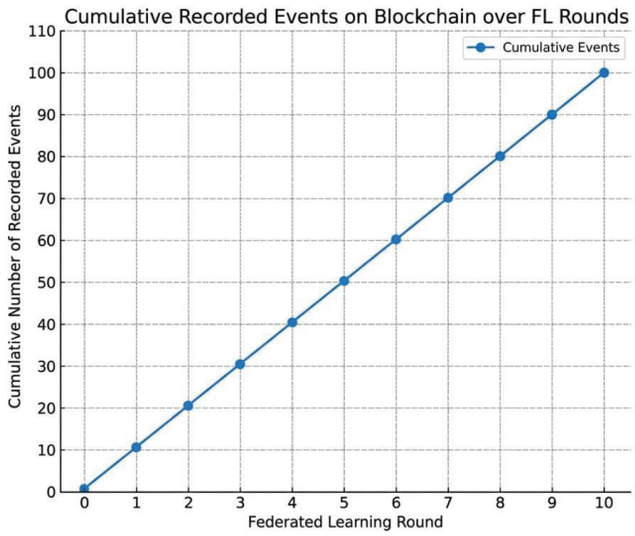


Figure 3: Cumulative Recorded Events on Blockchain over FL Rounds (Illustrative)

VIII. Discussion

The prototype results highlight practical advantages of integrating the three layers:

- DP synthetic data supports safer experimentation and limited sharing, while maintaining model quality close to real-data training.
- Federated learning enables multi-hospital training without moving raw data; FedProx and robust aggregation help with Non-IID stability and fairness.
- Permissioned blockchain governance provides consent-aware access control and tamper-resistant audit trails, improving accountability across institutions.

IX. Limitations

- DP strength is a trade-off: stronger privacy can reduce synthetic data/model utility.
- Operating a permissioned blockchain and policy system has real deployment costs (nodes, keys, governance).
- A trust score must be carefully defined and validated by stakeholders.

Prototype limitation. This work is a prototype/illustrative study. The evaluation demonstrates the integrated workflow and measurement design, but it does not claim clinical-level validation. Future work should include real multi-institution pilots, clearly reported DP budgets (e.g., ϵ), and stronger adversarial evaluations.

X. Conclusion and Future Work

This paper proposed an end-to-end platform that connects DP synthetic data, federated learning, and permissioned blockchain governance for privacy-preserving medical AI. The prototype results suggest that (i) DP synthetic data can keep model utility close to real-data training while reducing privacy leakage risk, (ii) a FedProx-based FL design can improve Non-IID stability and inter-site fairness without moving raw

patient data, and (iii) blockchain provenance logging can provide tamper-resistant audit trails across the full lifecycle (data generation–training–deployment). Overall, the platform offers a practical way to balance medical data utility, privacy protection, and cross-institution accountability in a single workflow.

Future work will expand and validate the platform in more realistic settings. First, we will conduct pilot studies with real hospitals by integrating with hospital information systems (e.g., EHR/PACS) and evaluating the platform under real operational constraints. Second, we will develop *regulation-aware* policies (e.g., GDPR/HIPAA-aligned) that can automatically adjust the utility–privacy profile by country, institution, and use case. Third, we will extend the platform to multimodal medical data (images, structured EHR, free-text notes, and genomics) to support broader clinical and research scenarios. Finally, we will enhance the blockchain-based trust framework into an end-to-end trust mechanism that covers model verification, deployment, version certification, and stronger threat models for both synthetic data and FL updates.

ACKNOWLEDGMENT

This research was supported by the Regional Innovation System & Education(RISE) program through the Gangwon RISE Center, funded by the Ministry of Education(MOE) and the Gangwon State(G.S.), Republic of Korea.(2025-RISE-10-005)

REFERENCES

[1] J. Yoon *et al.*, "PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees," in *Proc. ICLR*, 2019.
[2] J. Ho *et al.*, "Denoising Diffusion Probabilistic Models," in *Proc. NeurIPS*, 2020.
[3] R. Shokri *et al.*, "Membership Inference Attacks Against Machine Learning Models," in *Proc. IEEE S&P*, pp. 3–18, 2017.
[4] C. Xu *et al.*, "Differentially Private Diffusion Models for Tabular Data Synthesis," in *Proc. ICML*, 2023.
[5] B. McMahan *et al.*, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proc. AISTATS*, 2017.
[6] P. Kairouz *et al.*, "Advances and Open Problems in Federated Learning," *Foundations and Trends in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.
[7] K. Bonawitz *et al.*, "Towards Federated Learning at Scale: System Design," in *Proc. MLSys*, 2019.
[8] N. Rieke *et al.*, "The Future of Digital Health with Federated Learning," *npj Digital Medicine*, vol. 3, no. 119, 2020.
[9] T. Li *et al.*, "Federated Optimization in Heterogeneous Networks," in *Proc. MLSys*, 2020.
[10] T.-T. Kuo and L. Ohno-Machado, "Blockchain and Healthcare," *JAMIA*, vol. 24, no. 6, pp. 1211–1220, 2017.
[11] A. E. W. Johnson *et al.*, "MIMIC-IV, a Freely Accessible Electronic Health Record Dataset," *Scientific Data*, vol., 8, 2021.
[12] A. E. W. Johnson *et al.*, "MIMIC-CXR, a De-identified Publicly Available Database of Chest Radiographs with Free-text Reports," *Scientific Data*, 2019.
[13] A. L. Goldberger *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

Design of a Security Event-Driven Monitoring System Architecture for Software-Defined Vehicles

1st Tai-Gil Kwon *, 2nd Yongseong Kim †
* *Smart Network Research Center, Korea Electronics Technology Institute, Seongnam, Korea*
† *Smart Network Research Center, Korea Electronics Technology Institute, Seongnam, Korea*
Emails: *tgkwon@keti.re.kr, yskim@keti.re.kr*

***Abstract*—The proliferation of Software-Defined Vehicles (SDVs) has increased the complexity of in-vehicle software architectures and expanded the scope of security threats due to Over-the-Air (OTA) updates and Vehicle-to-Everything (V2X) communication. Traditional ECU-centric architectures are insufficient for real-time monitoring and response to security events arising in such dynamic environments. This study proposes a security event-based monitoring architecture built on the Adaptive AUTOSAR platform, which collects and normalizes security events and prioritizes them for real-time analysis at a central monitoring system. The proposed architecture provides design guidelines for integrated security monitoring and real-time threat response in SDV environments.**

Keywords—Security Event Monitoring, Adaptive AUTOSAR, Software-Defined Vehicle

I. Introduction

With the accelerated adoption of Software-Defined Vehicles (SDVs) across the automotive industry, the internal software architecture of vehicles has become significantly more complex[1]. In traditional ECU-centric architectures, functionalities are fixed in hardware, allowing for relatively stable security boundaries[2]. However, in SDV environments, software-centric operations such as Over-the-Air (OTA) updates, Vehicle-to-Everything (V2X) communication, and dynamic feature deployment based on Service-Oriented Architecture (SOA) are increasingly prevalent. These changes imply that new security vulnerabilities may continuously emerge across various points of interaction within and outside the vehicle, resulting in a wider range of potential attack vectors[3][4]. In such an environment, where software-based functionalities are rapidly increasing, managing security events in a distributed manner across the vehicle becomes challenging.

Therefore, an integrated security monitoring framework capable of real-time threat detection and response is essential. In particular, a structured approach is required to collect and normalize security-related data generated by various sensors, ECUs, and communication modules, and to prioritize events based on their severity or criticality[5].

To address these challenges, this study proposes a security event-based monitoring architecture centered on the collection, transmission, processing, and visualization of security events using the Adaptive AUTOSAR platform. The proposed architecture clearly defines the data flows and operational mechanisms among the key components required in SDV environments, providing a systematic foundation for real-time security monitoring and automated threat response across vehicle fleets.

II. Architecture and Functional Overview

A. Overview of the Adaptive AUTOSAR Architecture

Adaptive AUTOSAR differs from the static, ECU-centric architecture of Classic AUTOSAR by providing high-performance processors and a dynamic application execution environment. It applies a Service-Oriented Architecture (SOA) to flexibly support OTA updates, V2X communication, and cloud-integrated services[6].

Application Layer: Implements vehicle functionalities, security support modules (IDSR), and communication services, enabling service invocation and event-driven interactions.

Runtime Environment (ARA): Connects applications to platform services, manages application lifecycles, and provides messaging and security services.

Function Cluster Layer: Provides common functionalities including intrusion detection system management (IDSM), SOA-based communication management, cryptography and security services, and persistency.

Adaptive AUTOSAR supports the integration of IDSM and IDSR, and through service-based communication and dynamic application execution, it provides a foundation for real-time processing of large-scale security events and integrated monitoring in SDV environments[7].

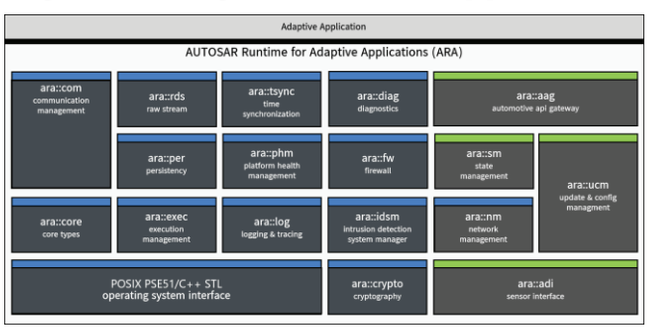


Fig. 1. Adaptive AUTOSAR Architecture

B. Proposed Security Monitoring System Architecture

This section introduces the proposed monitoring system architecture designed to collect, normalize, and analyze various security events generated within SDV (Software-Defined Vehicle) environments in real time. The proposed architecture consists of three key components: (1) IDSM, (2) Security Reporting Software, and (3) Security Monitoring Software. Each component is hierarchically integrated to enable real-time monitoring and management of the vehicle's internal security state. The overall architecture is structured around the workflow of security event collection, transmission,

processing, and visualization, with events being centrally managed through an Adaptive AUTOSAR-based service environment. This design provides a systematic monitoring framework that enables real-time security oversight and efficient threat response in SDV environments.

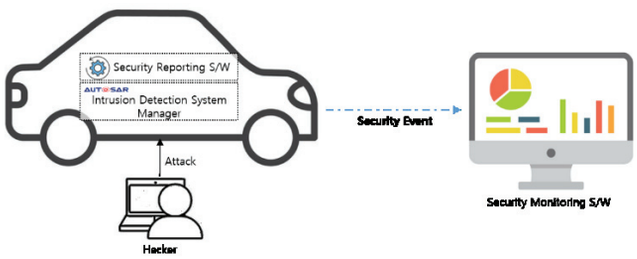


Fig. 2. Security Monitoring System Architecture

C. Detailed Architecture of Key Components

IDSM serves as a core Functional Cluster (FC) of the intrusion detection system within the Adaptive AUTOSAR platform. Positioned between security sensors and the Security Reporting Software, it functions as a central data hub by collecting, processing, filtering, and propagating security events generated from various security sensors[8].

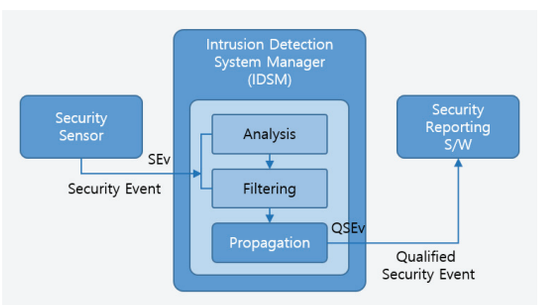


Fig. 3. IDSM Architecture

The Security Reporting Software serves as a key component responsible for transmitting security events detected within the vehicle to higher-level domains, specifically the SoC-based Security Monitoring Software. It receives normalized and filtered eligible security events from the IDSM, enriches them with metadata through security event analysis, reconstructs them into a transmittable extended format, and ensures reliable delivery. The architecture of the reporting software is organized into three main functional blocks: (1) blocks for integration with the AUTOSAR platform, (2) blocks for additional data processing of received security events, and (3) blocks for transmitting fully processed security events to the upper-level domain. This design enables structured, reliable, and extendable reporting of vehicle security events.

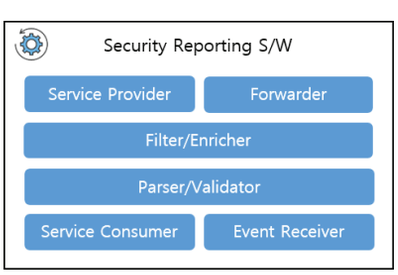


Fig. 4. Security Reporting Software Architecture

Before The Security Monitoring Software is designed to receive security events remotely transmitted from the in-vehicle Security Reporting Software, perform additional data processing, store backup data required for future security analysis, and enable real-time monitoring of the vehicle's security state. The core backend of the software is organized into three main functional blocks: (1) blocks for interfacing with the Security Reporting Software, (2) blocks for additional processing of received security events, and (3) blocks for managing fully processed security events in a database or transmitting them to the frontend for visualization. This architecture ensures comprehensive data management and real-time monitoring of security events within the vehicle.

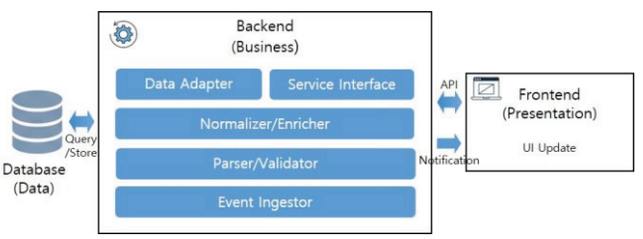


Fig. 5. Security Monitoring Software Architecture

III. Conclusion

This study proposed an Adaptive AUTOSAR-based monitoring architecture to address the increasing security requirements in SDV environments, enabling the collection, normalization, and analysis of in-vehicle security events. The proposed architecture clearly defines the functional roles and data flows among its key components, demonstrating the feasibility of implementing an integrated security monitoring system. Future work will focus on developing a prototype based on the designed monitoring architecture using the Adaptive AUTOSAR platform and conducting further research to validate its operational performance.

Acknowledgment

This work was supported by the Technology Innovation Program(RS-2025-02633228, Development of Access Control Technology for Future Mobility Cybersecurity) funded by the Ministry of Trade Industry & Energy(MOTIE, Korea).

References

[1] M. Kramer, M. Armbruster, and R. Johansson, "Software-Defined Vehicles: Challenges and Opportunities," *IEEE Vehicular Technology Magazine*, vol. 17, no. 3, pp. 45–55, 2022.

[2] AUTOSAR Consortium, *AUTOSAR Foundation—Overview and Concepts*, Version 1.9.0, 2021.

[3] P. García, A. Laya, and J. Alonso-Zarate, "V2X Communications in Next-Generation Vehicles: Architecture and Security Considerations," *IEEE Communications Standards Magazine*, vol. 6, no. 2, pp. 12–20, 2022.

[4] K. Koscher et al., "Experimental Security Analysis of a Modern Automobile," *IEEE Symposium on Security and Privacy*, pp. 447–462, 2010.

[5] S. Chattopadhyay and P. Mishra, "Automotive Security: A Survey of Vulnerabilities and Countermeasures," *ACM Computing Surveys*, vol. 54, no. 5, pp. 1–37, 2021.

[6] AUTOSAR Consortium, *AUTOSAR Adaptive Platform Overview*, AUTOSAR Release 23-11, 2023.

[7] J. Müller, "Security Event Processing in Adaptive AUTOSAR-Based Vehicle Architectures," *IEEE Vehicular Technology Magazine*, 2023.

[8] AUTOSAR Consortium, *Adaptive Platform – Intrusion Detection System Management (IDSM) Specification*, 2023.

Evaluation of Acceptance and Fidelity Scaling for Topology-Aware Zero-Level Magic-State Distillation in Lattice and Heavy-Hex Architecture

Leanghok Hour, Dongmin Kim, Sanghyeon Lee, Youngsun Han*
AI Convergence, Pukyong National University, Busan, South Korea
*Corresponding Author: *youngsun@pknu.ac.kr*

***Abstract*— We evaluate zero-level magic-state distillation on lattice and heavy-hex qubit topologies using Steane-encoded non-fault-tolerant circuits with verification rounds $r = 3$ and $r = 5$. Physical error rates range from 10^{-4} to 10^{-2}, and we track logical fidelity, acceptance probability, and routing overhead under realistic depolarizing noise. Across all configurations, logical fidelities remain extremely high, above 0.996 even at $p = 10^{-2}$, reflecting the shallow depth of zero-level distillation. Differences between layouts arise primarily in acceptance behavior and routing cost. Lattice layouts consistently sustain higher acceptance across the entire noise range, with acceptance at $p = 10^{-3}$ remaining around 0.72 for $r = 3$ and 0.49 for $r = 5$, Heavy-hex circuits show substantially lower acceptance (approximately 0.51 for $r = 3$ and 0.24 for $r = 5$ at the same error point), a direct consequence of their deeper circuits and larger CNOT counts. Routing analysis confirms this: heavy-hex requires 10–20% higher circuit depth and 30–40% more CNOTs than lattice layouts (e.g., 243 vs. 130 CNOTs for $r = 3$). This additional circuit volume does not significantly degrade logical fidelity but sharply increases the likelihood of failing verification rounds, causing acceptance to collapse more rapidly on heavy-hex than on lattice. Increasing verification rounds from $r = 3$ to $r = 5$ offers only marginal fidelity improvements but drastically reduces acceptance due to the cumulative rejection probability of repeated stabilizer checks. Overall, the results indicate that at zero-level distillation the limiting factor is acceptance rather than logical fidelity, and that hardware-dependent routing overhead plays a decisive role in determining the achievable throughput of near-term distillation pipelines.**

Keywords—quantum computing, quantum error correction, magic state distillation, fault-tolerant quantum computer.

I. Introduction

Magic-state distillation (MSD) [1] is the standard mechanism for supplying high-fidelity non-Clifford resources required for universal fault-tolerant quantum computation. At the bottom of the MSD hierarchy, zero-level distillation (ZLD) [2] is performed without concatenation or higher-order encoding, making it a clean probe of how hardware topology, gate noise, and routing overhead shape the very first stage of resource preparation. Because this stage operates directly on physical qubits, its effective throughput is controlled by concrete architectural factors such as two-qubit connectivity, path length, and the cost of repeated verification.

The architectural diversity of near-term superconducting [3] platforms raises a basic question: how much does the choice of connectivity graph impact the efficiency of zero-level MSD? Idealized lattice architectures provide dense local coupling and short paths, while heavy-hex style layouts restrict degree to mitigate crosstalk but introduce longer and more irregular routes between logical neighbors. These structural differences change the space–time volume of realistic distillation circuits and hence the balance between logical error suppression and acceptance probability across physical noise levels.

This work studies zero-level Steane-based magic-state distillation [2] under lattice and heavy-hex connectivity using a routing-aware, analytic framework. A 15-qubit Steane block is constructed with a parameter r that controls the number of verification rounds. For each topology and $r \in \{3, 5\}$, the circuit is transpiled under strict connectivity constraints, and the resulting depth, two-qubit gate count, and space–time volume feed into an analytic zero-level model that produces logical fidelity and acceptance curves as functions of the physical error rate. This isolates how connectivity and verification depth jointly limit the throughput of early-stage distillation pipelines on realistic architectures.

II. Methodology

Fig. 1 summarizes the Steane-based zero-level distillation block and its use in the topology-aware benchmark. Fig 1(a) shows the conceptual structure of the zero-level distillation procedure: an encoded magic-state register, a cat-state ancilla register, and a sequence of verification rounds that implement a Hadamard-test style distillation. In our implementation this structure is instantiated as a 15-qubit Steane-style circuit composed of three modules (Fig. 1(b–c))::

- Encode (blue block): Seven data qubits are initialized and transformed into a Steane/color-code stabilizer [4] state using a CZ-graph construction followed by additional Hadamards. A transversal H and T on the data qubits then approximate a logical magic state $|A_L\rangle$.
- Cat (red block): A 7-qubit cat state is prepared by applying H to one ancilla and fanning out with CNOTs, producing a GHZ-type state across the ancilla lines.
- The green Distill block consists of r repeated verification rounds, each performing the same interaction pattern between the cat register and the seven data qubits. In every round, each cat qubit connects to one data qubit through a short sequence of gates: a Hadamard is applied to the cat qubit, followed by two CNOTs to and from the corresponding data qubit, and then a final Hadamard. The only difference between rounds is which data

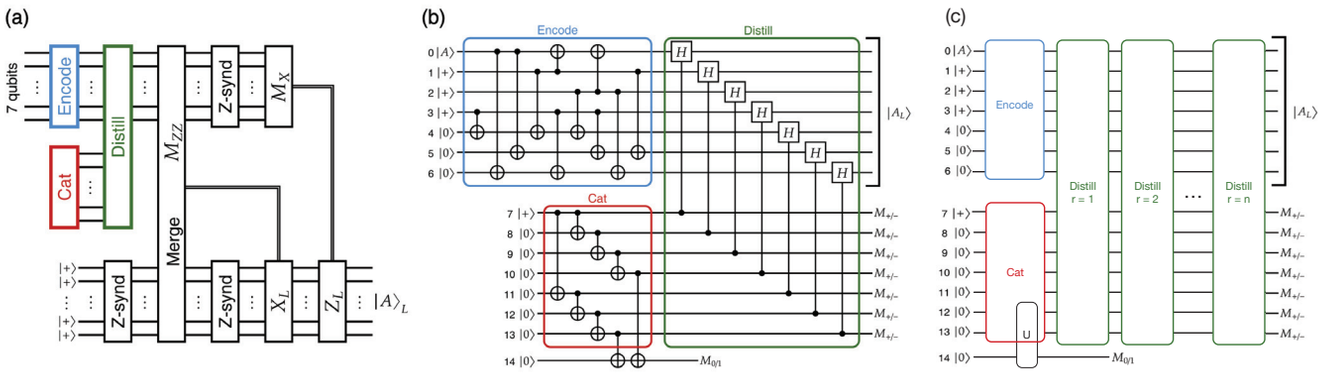


Figure 1: Overview of the Steane zero-level distillation experimentation. (a) High-level layout of zero-level distillation using 7 data qubits and a 7-qubit cat state. (b) Gate-level circuit showing Steane encoding (blue), cat preparation (red), and one full distillation sequence (green), where cat qubits interact with data qubits in a shifting pattern. (c) Abstract view of the protocol, highlighting that the same distillation block is repeated for r verification rounds.

qubit each cat qubit touches the targets shift by one position per round, cycling through all seven data qubits. This creates the diagonal pattern seen in the figure, where each verification round checks a different pairing between cat and data qubits. Together, these repeated interactions implement the simplified parity-check mechanism that determines whether the encoded state is accepted or rejected.

This circuit follows the structure shown in Fig. 1(b–c): seven data qubits form the encoded Steane block, seven ancillas form the cat register, and r identical distillation rounds act on all fourteen qubits. To study hardware effects, the full circuit is routed on two constrained connectivity graphs: a compact 2D lattice and a heavy-hex tiling with sparse links. A deterministic placement is used so that the lattice packs all qubits into a tight cluster, while the heavy-hex layout deliberately spreads them across multiple unit cells, forcing longer routes. Transpilation uses SABRE [5] with a fixed basis $\{rz, sx, X, CX, CZ\}$ to ensure that differences arise solely from topology. From the routed circuits we extract SWAP count, two-qubit gate count, depth, width, and the resulting space–time volume, which together reflect the architectural cost of running the ZLD block.

Logical performance is then computed using an analytic zero-level model tied directly to these routed metrics. A depolarizing noise rate p is assumed, with two-qubit gates assigned $20p$ to reflect their higher error. For each topology and verification depth r, the space–time volume (STV) determines an effective logical error $P_L \propto p^2 STV$, yielding a logical fidelity $F(p) = 1 - p_L$. Acceptance probability is modeled as an exponential decay that depends on both topology and verification rounds, causing heavy-hex layouts and larger r to reject more often as p increases. By sweeping p from 10^{-4} to 10^{-2}, the model provides paired fidelity and acceptance curves for each configuration, along with the corresponding routed resource metrics used in the evaluation.

III. Result

The transpilation results show a clear dependence of routing cost on hardware topology. Heavy-hex layouts consistently produce deeper circuits and substantially more two-qubit gates than lattice layouts, despite both requiring zero SWAP gates due to deterministic placements. For r = 3,

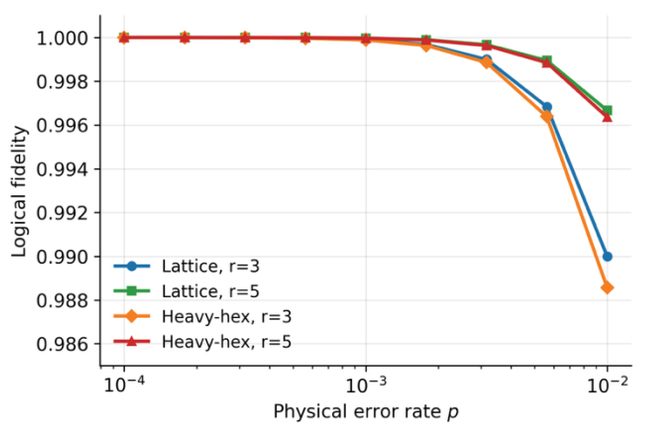


Figure 2: Logical Fidelity vs Physical Error Rate. Logical fidelity remains near unity at low noise and gradually decreases as the physical error rate increases. Heavy-hex layouts retain slightly higher fidelity than lattice layouts in the high-noise regime due to reduced correlated-error exposure, and increasing the verification rounds from r=3 to r=5 improves fidelity for both architectures.

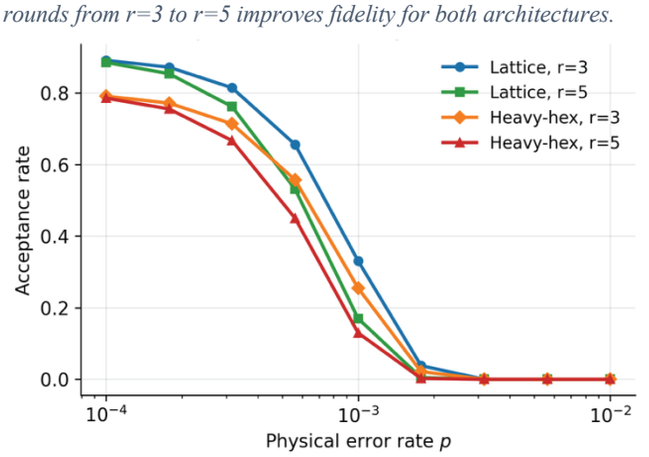


Figure 3: Acceptance Rate vs Physical Error Rate. Acceptance probability drops sharply as noise increases, with lattice architectures consistently achieving higher acceptance than heavy-hex layout at all p. Larger verification depth (r=5) reduces acceptance relative to r=3, reflecting the added cost of additional parity checks.

the lattice circuit compiles to a depth of 105 with 130 entangling gates, whereas the heavy-hex version reaches a depth of 120 with 243 entangling gates. At r = 5, these values rise to 131 and 182 on the lattice and to 143 and 334 on heavy-

hex, respectively. This increases space–time volume by roughly 15–20% on heavy-hex, reflecting the penalty imposed by sparse connectivity.

Logical fidelity trends, shown in Fig. 2, exhibit the complementary pattern characteristic of post-selected distillation. Since fidelity is computed only on accepted outcomes, both topologies achieve near-ideal values at low physical error rates. As p increases, heavy-hex circuits retain slightly higher conditional fidelity than lattice circuits despite their greater depth. Around $p \approx 3 \times 10^{-3}$, the lattice layout with r = 3 yields $F \approx 0.999$, while heavy-hex maintains $F \approx 0.998$, and the same relative ordering appears for $r = 5$. This reflects the reduced correlated-error propagation in degree-limited heavy-hex connectivity.

Fig. 3 shows the corresponding acceptance behavior under depolarizing noise. Acceptance decreases monotonically with increasing physical error rate p, and the lattice layout maintains a consistently higher acceptance probability across all tested values. At $p = 10^{-3}$, the lattice configuration achieves an acceptance of about 0.66 for $r = 3$, compared to 0.56 on heavy-hex. With more verification rounds, acceptance falls much more steeply: for $r = 5$, acceptance at the same noise level drops to 0.53 on the lattice and 0.45 on heavy-hex. This demonstrates the compounded overhead of deeper verification.

IV. Conclusion

This work examined zero-level magic-state distillation under two representative hardware connectivity models, lattice and heavy-hex, and quantified how topology shapes routing cost, acceptance probability, and conditional logical fidelity. By compiling a Steane-based ZLD block with deterministic placement and extracting routing-aware resource metrics, we exposed a clear architectural trade-off. Lattice layouts benefit from compact connectivity, producing shallower circuits and consistently higher acceptance across all tested noise levels. Heavy-hex layouts, despite incurring higher depth and CNOT counts, demonstrate slightly stronger fidelity retention due to reduced correlated-error propagation. The analytic model tied to the routed circuits further confirms that verification overhead dominates throughput, while connectivity governs how physical noise translates into logical performance. Together, these results highlight that zero-level distillation already exhibits strong topology dependence and that future distillation pipelines will require hardware-aware optimization to balance acceptance, fidelity, and resource cost on emerging fault-tolerant quantum platforms.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) [No. RS-2025-25434537].

References

[1] S. Bravyi, A. Kitaev. "Universal quantum computation with ideal Clifford gates and noisy ancillas," in Phys. Rev. A, vol. 71, pp. 022316, 2005.

[2] Itogawa, T., et al. "Efficient Magic State Distillation by Zero-Level Distillation," in PRX Quantum, vol. 6, pp. 020356, 2025.

[3] M. AbuGhanem. "Superconducting quantum computers: who is leading the future?," in EPJ Quantum Technology, vol. 12, no. 1, pp. 102, 2025.

[4] A. Steane. "Simple quantum error-correcting codes," in Physical Review A, vol. 54, no. 6, pp. 4741, 1996.

[5] G. Li, Y. Ding, Y. Xie, "Tackling the qubit mapping problem for NISQ-era quantum devices," in Proceedings of the twenty-fourth international conference on architectural support for programming languages and operating systems, 2019, pp. 1001–1014.

Wireless Power Transfer System Primary Current Detector Based for Wide Range Resonant Frequency Tracking

Youngseok Ko *, Eunho Choi †, Hyunkyeong Jo §, Heeyoung Chae ‡, Shihong Park *
* *Department of Electronic and Electrical Engineering, DanKook University, Yongin-si, South Korea*
† *Newhealthvision Co., Ltd., Jangan-gu, Suwon-si, South Korea*
§ *BWRIGHT Inc., 3F 311, 27, Sampung-ro, Gyeongsan-si, South Korea*
‡ *Department of Electrical Engineering, Dong-A University, Busan, South Korea*
Emails: *kosu2001@gmail.com, choi46337@gmail.com, hk.jo@b-wright.com, hchae@dau.ac.kr, shihong@dankook.ac.kr*

***Abstract*— This study introduces a frequency-adaptive tracking technique for wireless power transfer (WPT) systems that relies solely on primary-side current observation. The proposed controller integrates a peak-sensing process with a dual-phase search routine consisting of an initial coarse scan and a successive refinement stage. Without requiring feedback from the receiver, the algorithm rapidly identifies the resonant point through iterative peak comparison, allowing stable operation across a broad frequency range from 80 kHz to 300 kHz. Hardware verification demonstrates convergence within 2.31ms and frequency error below 125 Hz, achieving a maximum power transfer efficiency of 92%, which is roughly 10% higher than fixed-frequency operation.**

Keywords*— *wireless power transfer, resonance tracking, primary current detection, adaptive frequency control

I. INTRODUCTION

Wireless power transfer (WPT) has become increasingly significant in modern electronics ranging from small wearable devices to high-power EV charging systems. The transfer efficiency of such systems depends heavily on maintaining resonance between the transmitter (Tx) and receiver (Rx) circuits. Minor variations in component parameters—such as coil inductance, capacitance, or mutual coupling—can shift the resonant frequency, severely reducing power transfer efficiency.

As shown in Fig. 1, when component tolerances reach ±20%, the narrow resonant bandwidth can cause efficiency degradation of up to 44%. Maintaining precise resonance is thus critical for stable operation. Previous frequency-tracking strategies have used proportional–integral (PI) control and adaptive schemes such as fuzzy PI [1]–[3], whale optimization [2], and neural network–based adaptation [3]. Although effective, these methods often require complex tuning or secondary-side communication links. Perturb-and-observe (P&O) [4] and impedance-based dual-side control [5] achieve faster convergence but remain limited to narrow frequency bands.

To overcome these issues, this study proposes a primary-side-only resonance tracking method. The algorithm measures the transmitter current, identifies its peak, and adaptively modifies the PWM frequency to sustain resonance. The technique allows real-time operation across 80–300 kHz without receiver feedback, achieving sub-3 ms tracking and high efficiency. The conceptual flow and verification setup are detailed in later sections.

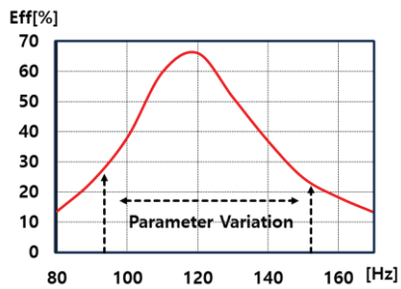


Fig. 1. The graph from Measured result of WPT System Efficiency versus operating frequency.

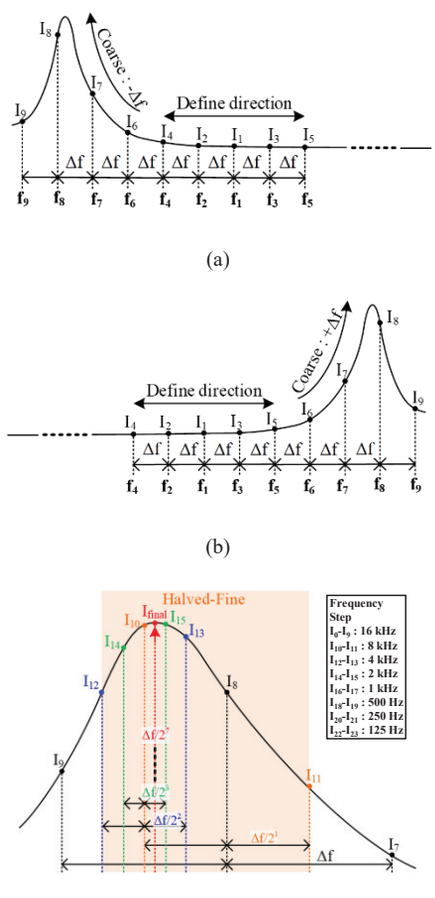


Fig. 2. Conceptual diagram of the proposed 3-point coarse and halved-fine frequency tracking algorithm. (a) illustrates the left direction coarse stage. (b) illustrates the right direction coarse stage (c) depicts the halved-fine stage

II. PROPOSED FREQUENCY TRACKING ALGORITHM

In this work, the resonant frequency tracking algorithm and its corresponding control strategy are proposed. First, the input signal of the transmitter inverter is modulated by a programmable PWM generator. The PWM frequency is then adaptively controlled based on the measured transmitter current. The proposed algorithm enables tracking of the desired resonant frequency of WPT system within less than 2.31 ms. The proposed method consists of two sequential stages: a 3-point defining direction and coarse stage followed by a halved-fine stage, as illustrated in Fig. 3. The proposed tracking algorithm begins from the center frequency (f_c) of the predefined search range and performs a spread exploration toward both sides (left and right directions). At each step, the algorithm measures three points (f_c, f_L, f_R) and determines the direction of monotonic increase in the peak amplitude I(f). If $I(f_R) > I(f_L)$, the frequency sweeps toward the right side; otherwise, it moves toward the left side.

$$\begin{cases} f_L = f_c - \Delta f,\ f_R = f_c + \Delta f \\ Direction = \begin{cases} Right = I(f_L) < I(f_R) \\ Left = I(f_L) > I(f_R) \end{cases} \end{cases} \quad (1)$$

Fig. 2 (a) and (b) illustrates coarse stages. The system evaluates the transmitter current at three consecutive frequency points using a predefined frequency step (Δf). Once the peak region (f_7 – f_9) is detected in Fig. 2 (a) and (b), the algorithm inserts a halved-fine stage.

Fig. 2 (c) depicts the halved-fine stage. In this stage, the algorithm continuously evaluates left-center-right points (I_L, I_C, I_R). The fine stage ends when a step size $\Delta f/2^n$ becomes smaller than 125 Hz.

After the coarse stage brackets the resonance peak, the halved-fine stage progressively reduces the search interval by half at each iteration. Let the current center frequency be $f_c^{(n)}$ with a search step s_n (= Δf). The algorithm measures three points:

$$f_L^{(n)} = f_C^{(n)} - \frac{s_n}{2}, f_C^{(n)}, f_R^{(n)} = f_C^{(n)} + \frac{s_n}{2} \quad (2)$$

The next center frequency is selected as:

$$f_C^{(n+1)} = \begin{cases} f_L^{(n)},\ if\ I(f_L^{(n)}) > I(f_C^{(n)}), I(f_L^{(n)}) > I(f_R^{(n)}), \\ f_R^{(n)},\ if\ I(f_R^{(n)}) > I(f_C^{(n)}), I(f_R^{(n)}) > I(f_L^{(n)}), \\ f_C^{(n)},\ otherwise \end{cases} \quad (3)$$

and the step size is halved at each iteration:

$$s_{n+1} = s_n/2 \quad (4)$$

The iteration continues until $s_n \leq s_{min}$ (125 Hz in this design), at which point the central frequency is considered the optimal resonant frequency f_{res}.

Its convergence rate is logarithmic, with a complexity of n:

$$n = log_2(\Delta f/s_{min}) = log_2(16kHz/125Hz) = 7 \quad (5)$$

For an initial range of 16 kHz and a final step of 125 Hz, only about 7-iterations are required. Additionally, the peak value within each measurement window is refined by parabolic interpolation among the three neighboring ADC samples.

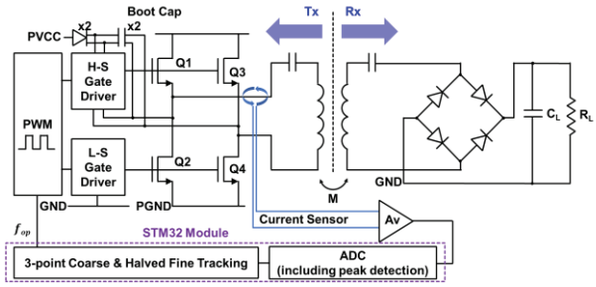


Fig. 3. Proposed WPT System Block Diagram

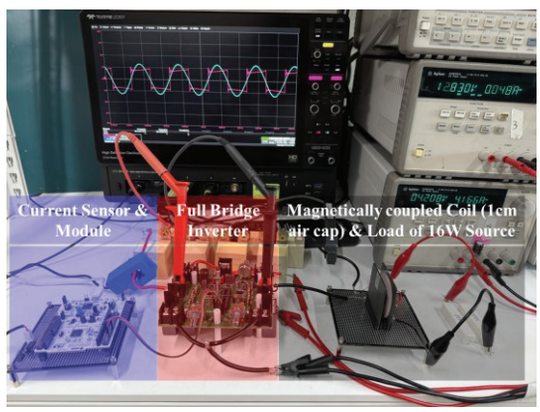


Fig. 4. Environment setup for proposed WPT System

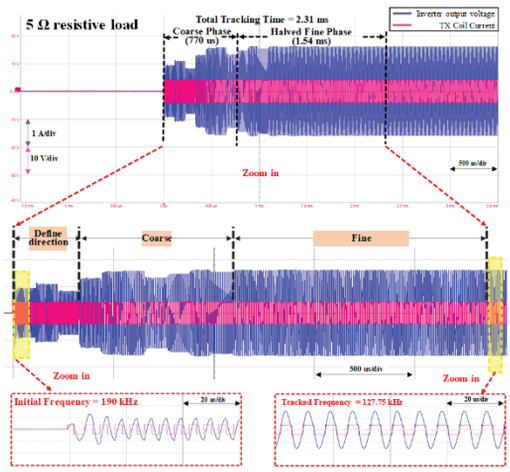
Each measurement cycle consists of a settling period and a sampling period. The settling time (60 µs) allows the resonant waveform to stabilize after frequency switching, and the following 50 µs sampling window captures the steady-state response. By using the 3-point coarse and halved-fine frequency tracking, the algorithm can achieve a minimum tracking time of 2.31 ms to find resonant frequency.

III. SYSTEM ARCHITECTURE

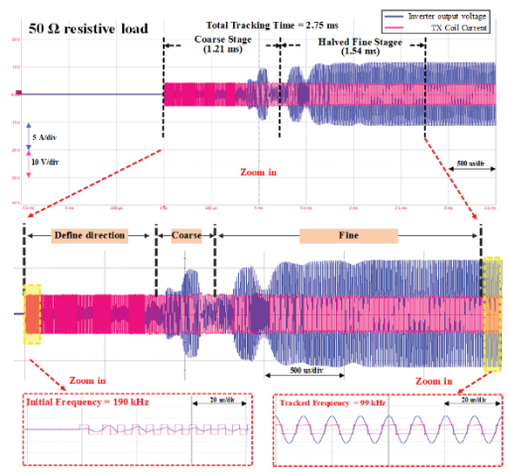
Fig. 3 illustrates the overall architecture of the proposed wireless power transfer (WPT) system implementing the 3-point coarse and halved-fine tracking algorithm. The system is composed of a full-bridge inverter, high-speed gate drivers, a current sensing and amplification stage, and the STM32G4-based digital control module. The inverter employs FDP2532 MOSFETs for efficient high-frequency operation, and Si8244 gate drivers provide up to 4 A peak output current with built-in dead-time control for reliable switching.

The current sensing stage in Fig. 3 detects the primary coil current using an SCT-013-020 Hall-effect sensor. Because the sensor output is relatively small, a half-wave amplifier with a gain of four is used to boost the signal amplitude. This amplified signal is then digitized through the 12-bit ADC of the STM32G4 MCU, ensuring accurate peak detection for frequency tracking.

Within the control block, the MCU continuously compares current peak values at different frequencies and dynamically adjusts the PWM frequency until the resonance condition is met. This primary-side-only feedback loop enables sub-3 ms convergence without relying on any receiver-side communication, resulting in improved system efficiency and stable operation even under varying load and coupling conditions.



(a)



(b)

Fig. 5. Measurement results of proposed WPT system with 3-point coarse and halved-fine algorithm and peak detector algorithm. (a) 5 Ω resistive load. (b) 50 Ω resistive load.

IV. Experimental Results

The proposed system was implemented on a discrete PCB using FDP2532 MOSFETs for the full-bridge inverter, Si8244 drivers, and STM32G4 as the control unit, as shown in Fig. 4. The input voltage was 12 V for the gate driver and 5 V for control circuits. Under a 5 Ω resistive load, the coarse stage completed in 770 µs and the halved-fine stage in 1.54 ms, resulting in a total tracking time of 2.31 ms at a resonant frequency of 127.75 kHz. Under the lighter load of 50 Ω, the system achieved convergence in 2.75 ms at 99 kHz, verifying the robustness of the algorithm under different operating conditions.

As shown in Fig. 5, the voltage and current waveforms at the transmitter output exhibited nearly zero phase difference when the system reached resonance. This

confirms that the proposed frequency-tracking algorithm effectively aligns with the true resonant frequency without relying on receiver feedback. In Fig. 6, the measured power transfer efficiency is compared with conventional fixed-frequency operation. The proposed algorithm achieved a maximum efficiency of 92%, improving efficiency by up to 10% compared with the fixed-frequency case across all load conditions.

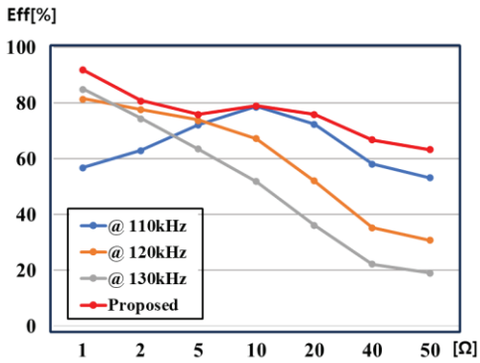


Fig. 6. Measured Efficiency Graph versus resistive load with proposed WPT system

These experimental results validate that the 3-point coarse and halved-fine frequency tracking algorithm provides rapid and stable frequency convergence. The sub-3 ms tracking performance makes it suitable for fast-dynamic wireless charging environments, while maintaining high efficiency and wide-range adaptability.

V. Conclusion

A novel frequency-tracking algorithm using peak detection with 3-point coarse and halved-fine strategy has been proposed for WPT systems. The method achieves fast, feedback-free resonance tracking within 2.31 ms, maintaining high efficiency (92%) across wide frequency variations. The design offers a simple and scalable solution for compact and high-speed WPT applications.

In addition, the proposed algorithm's adaptability to a wide frequency range (80–300 kHz) demonstrates its suitability for various power levels and coupling conditions. By relying solely on primary-side current monitoring, it reduces system complexity and cost while preserving precise resonance control. Future work will focus on extending this method to multi-coil or dynamic WPT systems, further enhancing its applicability to real-time wireless charging environments.

Acknowledgment

Youngseok Ko and Eunho Choi contributed equally to this work.

References

[1] Zheng, "Adaptive frequency tracking control with fuzzy PI compound controller for magnetically coupled resonant wireless power transfer." International Journal of Fuzzy Systems 23.6 (2021): 1890-1903.

[2] Yang, "PI parameters tuning for frequency tracking control of wireless power transfer system based on improved whale optimization algorithm." IEEE Access 12 (2024): 13055-13069.

[3] Liu, "Frequency tracking control of the WPT system based on fuzzy RBF neural network." International Journal of Intelligent Systems 37.7 (2022): 3881-3899.

[4] Liu, Chang, "Receiver Resonant Frequency Adaptive-Tracking in Wireless Power Transfer Systems Using Primary Variable Capacitors." IEEE Journal of Emerging and Selected Topics in Industrial Electronics (2025).

[5] Song, Kai,. "An impedance decoupling-based tuning scheme for wireless power transfer system under dual-side capacitance drift." IEEE Transactions on Power Electronics 36.7 (2021): 7526-753

Design Enhancement of a Lifecycle-based Verification Simulator for Digital Medical Products

Sung-Min Jeon
Dept. of Computer Engineering
Sangji-Ji University
Wonju, Republic of Korea
2025015104@sj.sangji.ac.kr

Jung-Hyun Woo
Dept. of Computer Engineering
Sangji-Ji University
Wonju, Republic of Korea
2025015001@sj.sangji.ac.kr

Kwang-Man Ko
Dept. of Computer Engineering
Sangji-Ji University
Wonju, Republic of Korea
kkman@sangji.ac.kr

Jin-Hyong Lee
Dept. of Computer Engineering
Sangji-Ji University
Wonju, Republic of Korea
2025015002@sj.sangji.ac.kr

Byung-Suk Seo
Dept. of Computer Engineering
Sangji-Ji University
Wonju, Republic of Korea
seobs@sangji.ac.kr

Han-Yong Choi
Dept. of Software Convergence
Shin-Han University
Seoul, Republic of Korea
hychoi@shinhan.ac.kr

***Abstract*—Digital medical products, particularly software-based medical devices, are characterized by frequent modifications and the need to simultaneously address complex standard and regulatory requirements. However, conventional compliance verification practices remain largely document-driven and manual, resulting in limited repeatability and reproducibility of verification outcomes, as well as weak integration with the product development lifecycle. These limitations often lead to the late discovery of regulatory nonconformities, causing schedule delays and increased development costs. To address these challenges, this paper proposes a simulator-based verification framework designed to support standard and regulatory compliance for digital medical products. The proposed framework is built upon two core components: a checklist-based verification structure that decomposes standard requirements into verifiable minimal units, and a lifecycle-oriented verification workflow that reflects product development and change management processes. This design enables systematic and repeatable compliance evaluation through explicit mapping between standard requirements and product-specific information. The proposed simulator integrates multiple international standards and regulatory guidelines within a unified management environment and supports re-verification in response to product changes, thereby facilitating continuous maintenance of regulatory compliance. The results of this study demonstrate the feasibility of embedding regulatory compliance activities throughout the development lifecycle of digital medical products, providing a practical verification framework that supports quality assurance and safety.**

***Index Terms*—Digital medical products, standard and regulatory compliance verification, simulator-based verification, checklist-based evaluation, lifecycle management**

I. Introduction

The medical device industry has rapidly shifted from hardware-centered systems to software-oriented digital medical products. Advances in artificial intelligence, cloud computing, and mobile technologies have expanded the use of medical device software for diagnostic, analytical, and therapeutic functions, improving healthcare efficiency and accessibility while introducing new safety and regulatory challenges [1,2]. Unlike conventional medical devices, software-based medical products undergo frequent updates, functional changes, and network-based operation, requiring quality management and regulatory compliance strategies distinct from those of hardware-centric devices. In response, international standards and regulatory frameworks addressing the safety, effectiveness, and security of medical device software have been continuously established and revised. These frameworks generally define lifecycle-wide requirements and form the basis for approval, review, and post-market surveillance systems in different countries. However, their practical application remains challenging due to the complexity and volume of standard documents, overlapping requirements, and evaluation processes that rely heavily on qualitative judgment. As a result, compliance verification is often document-centered and manual, leading to inconsistent and non-reproducible evaluation outcomes, increased development effort, and delays in regulatory approval. To address these limitations, there is a growing need for approaches that systematically structure standard and regulatory requirements and enable repeatable and consistent compliance verification. In this context, simulator-based verification has emerged as a promising alternative to conventional methods by linking regulatory requirements with product-specific information and supporting automated evaluation. This paper proposes a simulator-based framework for verifying international standards and domestic and international regulatory requirements for digital medical products. The proposed framework organizes requirements into checklist-based units, performs automated mapping and evaluation, and incorporates a lifecycle-oriented verification flow to support development, change management, and re-evaluation.

II. STANDARDS AND REGULATORY LANDSCAPE

A. International Standards and Regulatory Frameworks

With the rapid expansion of digital medical products, international standards and regulatory frameworks addressing the

safety, reliability, and security of medical device software have become increasingly important. Unlike traditional medical devices, software-centered products undergo frequent updates, functional modifications, and network-based operation, which require systematic management across the entire lifecycle. In response, international standardization organizations have established frameworks covering software development, verification, maintenance, and security. IEC 62304 defines software lifecycle processes and adopts a risk-based approach through software safety classification, providing a structured foundation for software development and maintenance. However, its practical application requires substantial expertise to interpret requirements and generate appropriate compliance documentation [3]. From a quality management perspective, ISO 13485 specifies organization-level requirements applicable to medical devices, including software, but provides limited guidance for verifying software-specific technical or security characteristics [4]. As cybersecurity has emerged as a critical concern, ISO/IEC 81001-5-1 introduces requirements for cybersecurity risk management and secure operation of medical device software [5]. Nevertheless, security standards are often applied separately from quality and safety standards, increasing overall compliance complexity. In addition, IEC 82304-1 addresses safe design and user information for medical software products, including standalone software medical devices, but must be applied in conjunction with other standards to support comprehensive regulatory compliance [6]. At the regulatory level, international frameworks adopt risk-based and staged approaches to manage software-intensive medical products. The World Health Organization Global Model Regulatory Framework for Medical Devices provides overarching guidance for establishing national regulatory systems, presenting a phased regulatory model that distinguishes between basic-level and expanded-level controls and defining regulation as a continuous process encompassing premarket evaluation, market entry control, and postmarket surveillance [1]. While this framework effectively defines regulatory principles and system-level structures, it does not offer concrete evaluation procedures or execution tools tailored to individual software medical devices. International regulatory authorities operate national systems based on WHO guidance and international harmonization initiatives. In parallel, the International Medical Device Regulators Forum provides common references for regulatory systems through guidance documents on topics such as software as a medical device and cybersecurity. However, documents such as IMDRF N60 remain principle-based and leave implementation and verification methods to individual interpretation, resulting in continued reliance on document-based and manual evaluation. A common characteristic of international standards and regulatory frameworks is their emphasis on defining what must be satisfied rather than how compliance should be systematically verified. As a result, consistent and efficient assessment across products and development stages remains difficult. Overall, international standards and regulatory frameworks define essential quality, safety, and security requirements from multiple perspectives, but these requirements are distributed across separate standards and lack integrated verification mechanisms. This limitation constrains repeated pre-verification and rapid response to software changes during development, highlighting the need for structured verification frameworks that integrate multiple standards and regulatory requirements and support repeatable and consistent compliance verification. These challenges motivate simulator-based approaches that extend beyond traditional document-centered compliance checks and provide a foundation for addressing domestic regulatory environments discussed in the following section.

B. Domestic Regulatory Environment and Limitations

In Korea, the regulatory environment for digital medical products and medical device software is primarily governed by the Medical Device Act, with regulatory systems and guidelines continuously refined in response to the rapid expansion of digital health technologies. The regulatory scope clearly distinguishes between standalone software medical devices and software embedded in or interoperating with hardware-based medical devices, enabling a structured approval and review framework. The Ministry of Food and Drug Safety serves as the primary regulatory authority responsible for approval and review processes and has issued dedicated guidelines reflecting international regulatory trends and harmonization principles [7]. A defining characteristic of the domestic regulatory framework is its strong reliance on document-centered review processes. Manufacturers are required to submit extensive documentation, including software requirements specifications, verification and validation reports, and risk management materials, which are predominantly assessed through qualitative and expert-based review [8]. As a result, a significant portion of regulatory effort is devoted to document preparation rather than continuous compliance assessment. Furthermore, regulatory outcomes may vary depending on interpretations of software functionality, intended use, and operational environments, reducing the predictability of regulatory responses, particularly for products subject to frequent updates or functional changes. Although domestic guidelines provide relatively detailed classifications for digital medical device software, these classifications are not directly linked to development processes or verification activities. Regulatory guidance primarily specifies what materials must be submitted for approval rather than how compliance should be systematically verified throughout the software lifecycle. Consequently, developers often rely on ad hoc checklists, prior approval cases, or manual interpretation of standards to demonstrate compliance. In addition, mechanisms for early-stage or iterative compliance assessment during development are limited, as verification activities are largely concentrated at the regulatory submission stage. This post hoc and document-driven verification structure increases the risk of late identification of regulatory nonconformities, potentially resulting in schedule delays, rework, and increased development costs. While the Korean regulatory framework aligns closely with international standards and principle-based regulatory models,

it provides limited practical support for integrated management of standard and regulatory requirements or for repeatable verification across development phases. These limitations underscore the need for tools that enable structured, automated, and lifecycle-oriented compliance verification, reinforcing the motivation for a simulator-based framework capable of supporting both international standards and domestic regulatory requirements.

III. PROBLEM DEFINITION AND DESIGN REQUIREMENTS

A. Limitations of Conventional Verification Approaches

Verification of digital medical products and medical device software is predominantly conducted through document-centered and manual processes. Although international standards and domestic regulatory guidelines define detailed requirements at the clause level, practical verification largely relies on developers' interpretation and written documentation rather than objective evaluation of product-specific technical characteristics. Consequently, compliance outcomes often depend on the completeness and presentation of submitted documents rather than on systematic technical assessment. One of the most significant limitations of conventional verification approaches is the lack of repeatability and reproducibility. Even when identical standard requirements are applied, checklist composition and evaluation results may vary depending on the product type, organization, or evaluator. In particular, requirements involving qualitative judgment frequently yield inconsistent outcomes due to differences in evaluator experience and perspective [9], making objective comparison across products or software versions difficult. Another critical limitation arises from the fragmented application of standards. Software lifecycle standards, quality management system standards, and cybersecurity guidelines are often managed independently without systematic integration. This fragmented approach leads to redundant documentation efforts and increases the risk of conflicting or omitted requirements, thereby reducing verification efficiency. Conventional verification practices also exhibit limited integration with the development lifecycle. Verification activities are typically concentrated at the regulatory submission stage, restricting continuous assessment during early development phases. When software modifications or functional changes occur, regulatory impacts are addressed retrospectively, which can result in late identification of nonconformities, schedule delays, and increased development costs. Furthermore, the absence of automation and tool-based support makes it difficult to systematically track evaluation results, change histories, and overall compliance status. As a result, regulatory compliance activities are often perceived as post hoc, approval-oriented tasks rather than as integral components of the development process. Table I summarizes the key limitations of conventional standard and regulatory compliance verification approaches, highlighting structural issues related to repeatability, integration, lifecycle alignment, and automation.

Table I
LIMITATIONS OF CONVENTIONAL STANDARD AND REGULATORY COMPLIANCE VERIFICATION APPROACHES

Aspect	**Conventional Approaches**	**Limitations**
Repeatability and Reproducibility	Interpretation depends on individual evaluators	Inconsistent evaluation results
Application of Standards	Standards applied independently	Redundant documentation and potential conflicts
Lifecycle Integration	Verification focused on pre-market approval stage	Limited support for change management and early validation
Level of Automation	Manual checklist preparation and document linkage	Low efficiency and poor traceability
Regulatory Response	Reactive, approval-oriented process	Increased risk of approval delays and rework

These limitations become increasingly pronounced when considering the characteristics of digital medical products and the growing complexity of standard and regulatory environments. Accordingly, there is a clear need for new approaches that can structurally manage standard requirements and enable repeatable and consistent verification throughout the entire development lifecycle. This problem recognition serves as the starting point for discussing the necessity of simulator-based verification approaches in the following subsection.

B. Necessity of a Simulator-Based Approach

To address the limitations of conventional document-centered verification practices, a structured and proactive compliance verification approach is required. Digital medical products are characterized by frequent changes and short development cycles, making post hoc and approval-centered verification insufficient for ensuring consistent regulatory compliance. A simulator-based approach enables standard and regulatory requirements to be represented as structured and verifiable models, supporting repeatable and reproducible evaluation results. By explicitly defining relationships between product-specific information and applicable requirements, compliance assessment can be performed using rule-based mechanisms rather than subjective interpretation, thereby reducing uncertainty inherent in qualitative document reviews. In addition, managing multiple international standards and regulatory guidelines within a unified verification environment allows overlapping requirements to be handled efficiently and minimizes the risk of conflicts or omissions. Furthermore, a simulator-based approach supports continuous compliance assessment throughout the development lifecycle. Functional changes or software updates can be promptly re-evaluated, enabling early identification of regulatory impacts and shifting verification from a reactive, submission-stage activity to a lifecycle-oriented process. Automation reduces manual checklist preparation and document linkage efforts while enabling systematic recording and management of evaluation results [10]. Through these capabilities, simulator-based

Table II
MAPPING BETWEEN IDENTIFIED LIMITATIONS AND SIMULATOR DESIGN REQUIREMENTS

Identified Limitation	Corresponding Design Requirement
Frequent revisions and fragmented standards management	Modular and standard-agnostic architecture
Subjective interpretation of requirements	Checklist-based structure with explicit evaluation criteria
Manual identification of applicable requirements	Automatic mapping between product information and requirements
Lack of lifecycle integration	Lifecycle-based verification and version management
Binary pass/fail evaluation	Detailed, item-level compliance status and justification
Limited support for regulatory documentation	Evidence generation and traceable evaluation records
Poor traceability of evaluation decisions	Explicit recording of decision rationale and change history

verification provides a practical foundation for consistent, lifecycle-integrated compliance management of digital medical products.

C. Derivation of Design Requirements

Based on the identified limitations of conventional verification approaches, essential design requirements for simulator-based compliance verification can be derived. The simulator should adopt a modular and standard-agnostic architecture to accommodate revisions to existing standards and the introduction of new regulatory requirements. Standard and regulatory requirements must be decomposed into verifiable units with explicit evaluation criteria to ensure consistent and reproducible verification results. The simulator should automatically identify and map applicable requirements based on product-specific information, such as intended use and functional characteristics, thereby improving evaluation efficiency and reducing unnecessary verification efforts. Verification workflows must reflect the entire software lifecycle, supporting version management, change tracking, and re-verification to maintain continuous compliance. In addition, verification outcomes should be provided at an item level rather than as binary pass/fail results, enabling clear identification of nonconformities and corrective actions. The simulator should also support regulatory activities by generating evidence and maintaining traceable evaluation histories for regulatory review and internal quality management. Table II summarizes the relationship between the identified limitations and the derived simulator design requirements.

Table II summarizes the relationship between the identified limitations and the derived simulator design requirements.

IV. SIMULATOR-BASED VERIFICATION FRAMEWORK

A. Overall System Architecture

The proposed simulator adopts a layered system architecture to systematically manage digital medical products and complex standard and regulatory requirements. The system

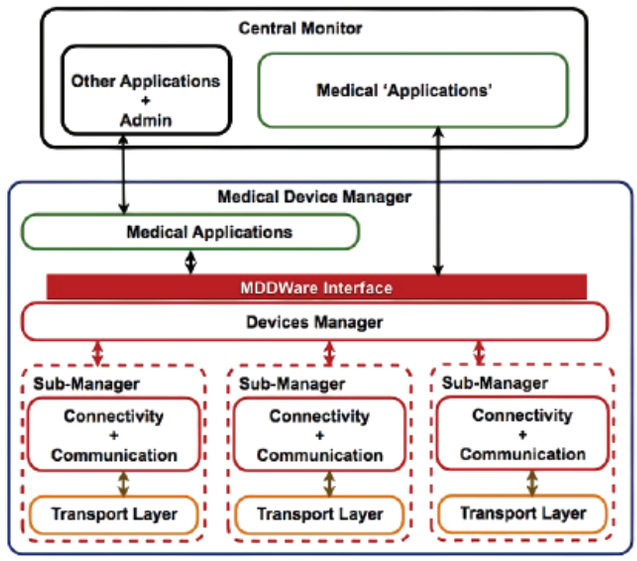


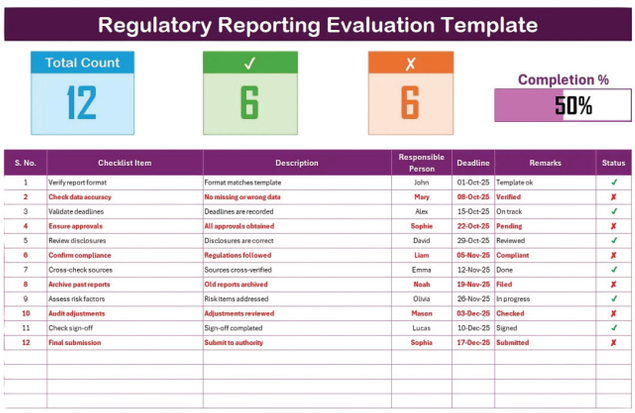
Figure 1. System architecture of a simulator for standard and regulatory compliance verification of digital medical products

supports an end-to-end verification workflow encompassing product information input, requirement mapping, verification execution, and result management. Each function is implemented as a modular component to ensure scalability and extensibility.

As illustrated in Fig. 1, the simulator consists of a user interface layer, a product information management layer, a standard and regulatory management layer, a verification engine layer, and a result management and output layer. The user interface layer provides functions for product registration, standard selection, verification execution, and result review. The product information management layer maintains structured product data, including intended use, functional characteristics, development stage, and change history, which are used to determine applicable verification scopes. In the standard and regulatory management layer, international standards and regulatory requirements are organized into modular checklist-based units. A product–standard mapping module automatically identifies applicable requirements and generates a product-specific verification checklist. The verification engine executes predefined evaluation rules in an automatic or semi-automatic manner, and verification results are recorded in a consistent format. The result management and output layer stores item-level compliance status together with version and change information and supports visualization and report generation for regulatory submission and review.

B. Checklist-Based Verification Structure

A core component of the proposed simulator is a checklist-based verification structure that transforms standard and regulatory requirements into explicit, verifiable units. Because medical device software standards and regulatory guidelines are typically described at the clause level, direct application can lead to ambiguity and subjective interpretation. To address



S. No.	Checklist Item	Description	Responsible Person	Deadline	Remarks	Status
1	Verify report format	Format matches template	John	01-Oct-25	Template ok	✓
2	Check data accuracy	No missing or wrong data	Mary	08-Oct-25	Verified	✗
3	Validate deadlines	Deadlines are recorded	Alex	15-Oct-25	On track	✓
4	Ensure approvals	All approvals obtained	Sophie	22-Oct-25	Pending	✗
5	Review disclosures	Disclosures are correct	David	29-Oct-25	Reviewed	✓
6	Confirm compliance	Regulations followed	Liam	05-Nov-25	Compliant	✗
7	Cross-check sources	Sources cross-verified	Emma	12-Nov-25	Done	✓
8	Archive past reports	Old reports archived	Noah	19-Nov-25	Filed	✗
9	Assess risk factors	Risk items addressed	Olivia	26-Nov-25	In progress	✓
10	Audit adjustments	Adjustments reviewed	Mason	03-Dec-25	Checked	✗
11	Check sign-off	Sign-off completed	Lucas	10-Dec-25	Signed	✓
12	Final submission	Submit to authority	Sophia	17-Dec-25	Submitted	✗

Figure 2. Checklist-based verification structure

Table III
STRUCTURE OF CHECKLIST-BASED COMPLIANCE VERIFICATION ITEMS

Component	Description
Standard Identifier	Identifier of the applicable international standard or regulatory guideline clause
Requirement Description	Summary of the compliance requirement to be verified
Applicability Condition	Criteria used to determine whether the requirement applies based on product characteristics
Evaluation Criteria	Specific criteria used to assess compliance with the requirement
Evaluation Method	Type of evaluation performed (automatic, semi-automatic, or manual)
Required Evidence	Documents or artifacts required to support the evaluation
Evaluation Result	Compliance status (compliant, non-compliant, or not applicable)

this issue, standard requirements are decomposed into minimal checklist items suitable for systematic evaluation.

As illustrated in Fig. 2, standard and regulatory clauses are decomposed into checklist items defined by structured attributes such as applicability conditions, evaluation criteria, evaluation methods, and required evidence. Applicability of checklist items is automatically determined based on product-specific information, including intended use, functional characteristics, risk level, and development stage, enabling verification to focus only on relevant requirements. Evaluation is performed through a combination of automated rule-based checks and user-assisted assessment for items requiring expert judgment, and evaluation results are recorded together with decision rationales to support traceability.

This checklist-based structure also enables integrated management of overlapping requirements across multiple standards by consolidating similar requirements into unified verification items, thereby reducing redundant evaluations and improving consistency across standards.

C. Lifecycle-Based Verification Flow

Digital medical products, particularly software-based medical devices, undergo frequent functional updates and modifications even after initial development. Accordingly, standard

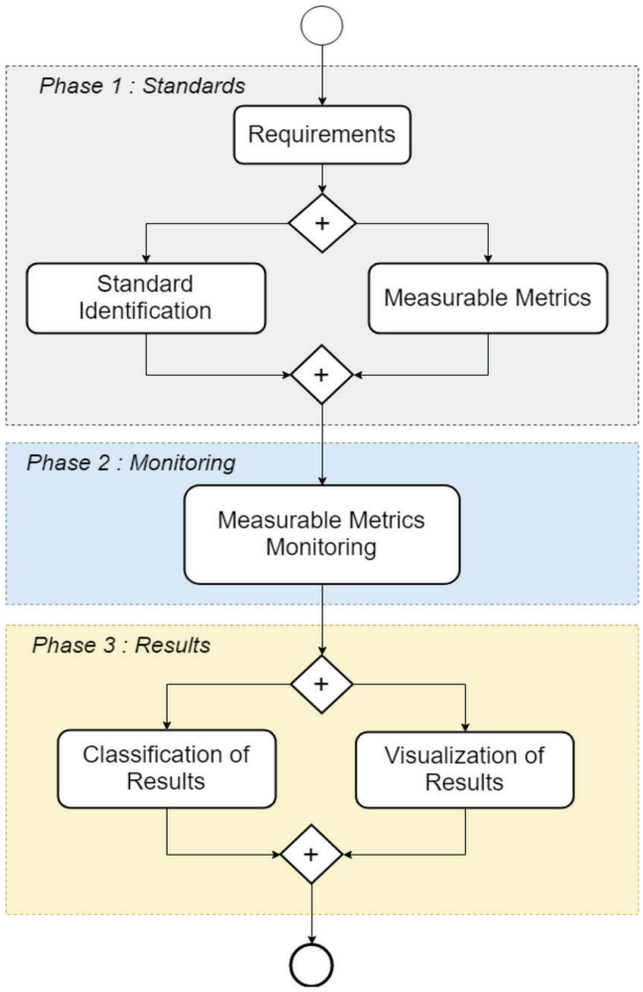


Figure 3. Lifecycle-based verification flow for standard and regulatory compliance

and regulatory compliance verification must be performed continuously throughout the product lifecycle rather than as a one-time activity. To support this requirement, a lifecycle-based verification flow is incorporated as a core element of the proposed simulator.

As illustrated in Fig. 3, the proposed verification flow follows a cyclic process consisting of requirement identification, checklist-based verification, result management, and re-verification upon change. Applicable international standards and regulatory requirements are identified at early development stages and structured into checklist-based verification items. Verification is performed based on product-specific information, and evaluation results are recorded at the item level together with version and change histories to support traceability. When functional changes or software updates occur, the simulator automatically adjusts the verification scope to focus on affected checklist items, reducing unnecessary re-evaluation. This cyclic flow enables continuous regulatory compliance and integrates verification activities into the product lifecycle rather than treating them as late-stage approval

tasks.

V. Applicability and Expected Effects

This section examines the practical applicability of the proposed simulator-based verification framework and its expected effects in digital medical product development and regulatory compliance. Owing to its standard-agnostic and modular architecture, the simulator can be flexibly applied to a wide range of digital medical products and development environments. From a development perspective, the proposed simulator supports systematic integration of standard and regulatory requirements from early development stages. Unlike conventional approaches that focus compliance verification primarily at the approval stage, the checklist-based structure and lifecycle-oriented verification flow enable continuous monitoring of regulatory conformity throughout development, facilitating early detection of nonconformities and reducing the likelihood of extensive revisions or rework in later stages.

For quality and regulatory management, the simulator enhances the consistency and traceability of verification activities by managing standard requirements, verification items, and evaluation results in a structured and unified manner. In addition, version-based management of verification results and change histories supports systematic preparation of evidence required for regulatory review, thereby improving the reliability and transparency of regulatory responses. From a regulatory approval perspective, the simulator can function as a pre-assessment tool prior to formal submission, enabling early identification of potential deficiencies and reducing the risk of approval delays. Overall, the proposed simulator-based verification framework reduces inefficiencies in digital medical product development and regulatory compliance activities by embedding compliance verification into the development lifecycle and supporting effective regulatory responses.

VI. Conclusion and Future Work

This paper analyzed the limitations of document-centered and post hoc approaches commonly used for standard and regulatory compliance verification of digital medical products and proposed a simulator-based verification framework to address these challenges. Software-oriented digital medical products require systematic management of frequent changes and complex regulatory requirements, which are difficult to handle efficiently using conventional qualitative and manual compliance practices. To overcome these limitations, this study introduced a simulator concept that structures international standards and domestic and international regulatory requirements into checklist-based verification items and links them with product-specific information. The proposed framework consists of a layered system architecture, a checklist-based verification structure, and a lifecycle-oriented verification flow, enabling continuous compliance management from early development through change management and re-verification. By integrating distributed standard and regulatory requirements into structured verification units, the framework improves consistency and reproducibility of verification results and supports systematic tracking of compliance status across product versions. In future work, the proposed simulator should be applied to real-world digital medical product development cases to conduct empirical validation. Quantitative evaluation of verification accuracy, operational efficiency, and regulatory response effectiveness will be necessary to assess its practical applicability. Further research should also extend the framework to support country-specific regulatory requirements and expand automated evaluation capabilities. Incorporating intelligent techniques based on accumulated verification data, such as regulatory risk prediction and decision support, represents another promising direction. Through these extensions, the proposed simulator can evolve into an integrated and intelligent platform for regulatory compliance management throughout the digital medical product lifecycle.

Acknowledgment

This research was supported by the Regional Innovation System & Education(RISE) program through the Gangwon RISE Center, funded by the Ministry of Education(MOE) and the Gangwon State(G.S.), Republic of Korea.(2025-RISE-10-005). This work was supported by the National IT Industry Promotion Agency(NIPA) grant funded by the Korea government(MSIT) (H1201-24-1001, 2025 Digital Twin Convergence Medical Innovation Project).

References

[1] World Health Organization, *WHO Global Model Regulatory Framework for Medical Devices Including In Vitro Diagnostic Medical Devices*. Geneva, Switzerland: WHO Press, 2017.

[2] International Medical Device Regulators Forum, *Software as a Medical Device (SaMD): Clinical Evaluation*. IMDRF, 2017.

[3] International Electrotechnical Commission, *IEC 62304: Medical Device Software—Software Life Cycle Processes*. Geneva, Switzerland: IEC, 2015.

[4] International Organization for Standardization, *ISO 13485: Medical Devices—Quality Management Systems—Requirements for Regulatory Purposes*. Geneva, Switzerland: ISO, 2016.

[5] ISO/IEC, *ISO/IEC 81001-5-1: Health Software and Health IT Systems Safety, Effectiveness and Security—Part 5-1: Security*. Geneva, Switzerland: ISO/IEC, 2021.

[6] International Electrotechnical Commission, *IEC 82304-1: Health Software—Part 1: General Requirements for Product Safety*. Geneva, Switzerland: IEC, 2016.

[7] Y. Zhang, Y. Chen, and J. Wang, "A systematic approach to regulatory compliance verification for medical device software," *Computer Methods and Programs in Biomedicine*, vol. 177, pp. 199–210, 2019.

[8] Ministry of Food and Drug Safety (MFDS), *Guideline on Software Validation for Medical Devices*. Cheongju, Republic of Korea, latest ed.

[9] M. McHugh, F. McCaffery, and V. Casey, "Standards-based software process improvement for medical device software," *Journal of Systems and Software*, vol. 85, no. 3, pp. 595–611, 2012.

[10] D. Ayed, S. Abdelhak, and A. Trabelsi, "A model-driven approach for compliance management of medical device software," *IEEE Access*, vol. 8, pp. 189012–189025, 2020.

SWKD: Scalable Wavelet Knowledge Distillation for Vision Transformers

Jungmin Kim*, Changhee Kang†, Sang-ug Kang‡§
* *Department of Computer Science, Sangmyung University, Seoul, Korea*
† *Department of Computer Science, Sangmyung University, Seoul, Korea*
‡ *Department of Computer Science, Sangmyung University, Seoul, Korea*
Emails: *202532010@sangmyung.kr, 202434012@sangmyung.kr, sukang@smu.ac.kr*
§ *Corresponding author*

***Abstract*—Vision Transformers (ViTs) provide superior performance but face deployment challenges due to massive computational demands. While Knowledge Distillation (KD) enables compression, current Discrete Cosine Transform-based methods like ScaleKD lose spatial localization. We propose the Wavelet Transform Module (WTM), employing Haar Discrete Wavelet Transform (DWT) and learnable weights to preserve simultaneous spatial and spectral consistency. Specifically, WTM decomposes feature maps into LL, LH, HL, and HH sub-bands to extract directional textures and edges while strictly maintaining their spatial coordinates. These sub-bands are then adaptively calibrated and reconstructed via Inverse DWT (iDWT) to provide spatially consistent supervision through a reconstruction-based loss. Our method achieves 82.3% Top-1 accuracy on ImageNet-1K.**

Keywords—vision transformer, knowledge distillation, wavelet transform

I. Introduction

ViTs [1] have established a new paradigm in computer vision by effectively modeling global contextual dependencies, outperforming traditional Convolutional Neural Networks (CNNs) across various tasks. However, their massive parameter scale and computational intensity hinder deployment in resource-constrained environments. KD [2] has been extensively explored to address this by transferring expertise from large teachers to compact students. Yet, early KD methods designed for CNNs often fail to capture the unique patch-based operations and complex frequency characteristics inherent in ViT architectures.

To overcome these limitations, several ViT-specific KD frameworks have been proposed. Manifold KD attempts fine-grained knowledge transfer by aligning patch-level manifold spaces, while ScaleKD introduces frequency-domain mimicking based on the DCT to resolve differences in computing paradigms. Furthermore, SpectralKD optimizes distillation efficiency by identifying information-rich layers through spectral analysis. However, existing frequency-based approaches, such as those utilizing DCT, suffer from a critical drawback: the loss of spatial localization. By focusing on global frequency components, these methods struggle to preserve the local spatial details that are essential for high-fidelity feature representation.

In this paper, we argue that existing methods suffer from three critical shortcomings that our proposed framework aims to solve. First, while Manifold KD captures patch relationships, it remains "spectrally blind," overlooking the intrinsic frequency distributions (e.g., textures and edges) that define those patches. Second, although ScaleKD considers frequency, its reliance on DCT causes a loss of spatial localization, making it impossible to determine where specific frequency components occur. Third, SpectralKD focuses on global spectral patterns but lacks adaptive, fine-grained calibration due to its parameter-free nature.

To bridge these gaps, we propose a novel framework utilizing the DWT and learnable weights. Unlike previous approaches, DWT enables multi-resolution analysis that simultaneously preserves spatial and frequency information by decomposing features into LL, LH, HL, and HH sub-bands. We introduce learnable weights tailored to feature resolution to adaptively calibrate sub-band importance, surpassing the rigid alignment of SpectralKD. Finally, features are reconstructed via Inverse DWT (iDWT) for supervision, ensuring a stable and spatially consistent transfer of both global context and localized details.

II. Proposed method

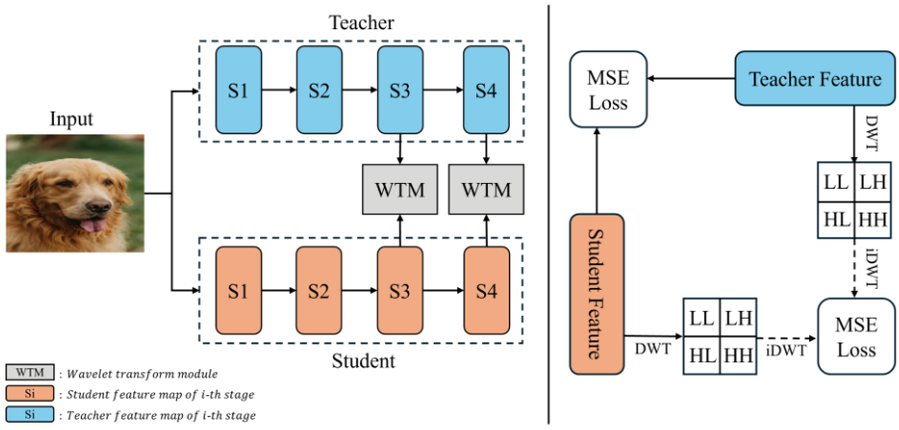
The proposed framework ensures hierarchical alignment between teacher and student models by maintaining both spatial and spectral consistency at Stages 3 and 4. We introduce a fundamental shift in the frequency-aware supervision path by replacing the conventional Dual-view Feature Mimicking (DFM) of ScaleKD with our Wavelet Transform Module (WTM).While ScaleKD utilizes the DCT domain and iDCT reconstruction—often losing precise spatial-frequency localization by zeroing out the DC component—our WTM leverages the Haar Discrete Wavelet Transform (DWT). By decomposing features into multi-resolution sub-bands, WTM allows the student to adaptively inherit directional textures and edges, ensuring superior feature alignment. As illustrated in Figure 1(b), the WTM first applies the Haar DWT to decompose a given feature map into four distinct sub-bands: LL (low-frequency approximation), LH (horizontal), HL (vertical), and HH (diagonal). This multi-resolution analysis enables the student to capture fine-grained spectral details while strictly preserving their original spatial coordinates.

To enable adaptive knowledge transfer, we introduce learnable weights W_k that match the resolution ratio of each sub-band. The student's sub-band features are then calibrated using these weights through an element-wise product:

$$\hat{F}_k^s = W_k \circ F_k^s, k \in \{LL, LH, HL, HH\} \quad (1)$$

After this adaptive calibration, the sub-bands are reconstructed into the spatial domain via the Inverse DWT (iDWT):

$$R = iDWT\left(\hat{F}_{LL}, \hat{F}_{LH}, \hat{F}_{HL}, \hat{F}_{HH}\right) \quad (2)$$



(a) SWKD Framework (b) Wavelet Transform Module

Figure 1. Block diagram of the proposed knowledge distillation method. (a) The overall proposed framework, where features from different stages of the teacher are projected and aligned with student features through the Wavelet Transform Module (WTM). (b) Dual-view supervision strategy that enforces feature consistency in both spatial and frequency domains. Frequency-aware supervision is achieved by decomposing features into wavelet sub-bands (LL, LH, HL, HH) using DWT and reconstructing them via iDWT.

The final supervision is performed by calculating the Mean Squared Error (MSE) on these reconstructed features. To match the implementation where the loss is calculated as the sum of squared errors divided by the batch size B, the frequency-aware loss L_{freq} is defined as:

$$L_{freq} = \frac{1}{B}\sum_{b,c,h,w} \left\| \hat{R}^t - \hat{R}^s \right\|^2 \quad (3)$$

where $\hat{R}$ represents the normalized reconstructed feature maps. The total loss function L_{total} is defined as a combination of the standard spatial mimicking loss and our frequency-aware reconstruction loss:

$$L_{total} = L_{spatial}(F^t, F^s) + \beta L_{freq}(R^t, R^s) \quad (4)$$

III. Experiments

A. Data setup and Experimental details

All experiments are conducted on the ImageNet-1K dataset, which contains 1.28 million training and 50,000 validation images. We employ Swin-Small as the teacher and Swin-Tiny as the student. To facilitate effective knowledge transfer, our proposed WTM is integrated into the output of Stages 3 and 4. We train the student for 300 epochs with a total batch size of 1024. We use the AdamW optimizer with a base learning rate of 1e-3 and a 20-epoch linear warmup, followed by a Cosine Decay schedule.

B. Experimental results

The proposed method achieves 82.3% on ImageNet-1K and surpassing Manifold KD, as summarized in Table I. While the performance is slightly lower than SpectralKD and ScaleKD, our framework demonstrates superior efficiency; specifically, ScaleKD utilizes a significantly larger Swin-Large teacher. These results are driven by the WTM, which preserves critical spatial localization through Haar DWT and learnable weights. Our iDWT-based supervision ensures a spatially consistent transfer of fine-grained textures and edges.

IV. Conclusion and Future Plan

The proposed method surpasses Manifold KD and demonstrates superior efficiency compared to ScaleKD by utilizing a teacher model that is approximately 4x smaller. Building on these results, our future research will explore removing the low component to focus exclusively on high-frequency details, aiming to eliminate informational redundancy and further optimize the distillation path.

TABLE I. Comparison with existing methods on ImageNet 1K

Method	ImageNet 1K		
	Model (Teacher/Student)	***Params***	***Top-1 Acc(%)***
Manifold KD [3]	Swin-Small/Swin-Tiny	50M/29M	82.2
Spectral KD [5]	Swin-Small/Swin-Tiny	50M/29M	82.7
Scale KD [4]	Swin-Large/Swin-Tiny	197M/29M	83.8
Proposed	Swin-Small/Swin-Tiny	50M/29M	82.3

Acknowledgment

This work is supported by the National Research Foundation of Korea grant funded by Korea government (NRF- 2022R1A2C1004674).

References

[1] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in Proc. IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10012-10022, 2021.

[2] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv:1503.02531, 2015.

[3] Z. Hao, J. Guo, D. Jia, K. Han, Y. Tang, C. Zhang, H. Hu, and Y. Wang, "Learning efficient vision transformers via fine-grained manifold distillation," Advances in Neural Information Processing Systems, 2022, pp. 9164-9175.

[4] J. Fan, C. Li, X. Liu, and A. Yao, "Scalekd: Strong vision transformers could be excellent teachers," Advances in Neural Information Processing Systems, 2024, pp. 63290-63315.

[5] H. Tian, B. Xu, S. Li, and G. Pan, "SpectralKD: A unified framework for interpreting and distilling vision transformers via spectral analysis," arXiv preprint arXiv:

An Initial Study Toward a Protocol-Level Cybersecurity Maturity Model: The Evolution of MQTT Authentication

Arfan Ridwan Hartawan*, Muhammad Salman*
*Department of Electrical Engineering, University of Indonesia, Depok, Indonesia
arfan.ridwan@ui.ac.id, muhammad.salman@ui.ac.id

***Abstract*—The increasing adoption of the Message Queuing Telemetry Transport (MQTT) protocol in Internet of Things (IoT) systems has raised significant concerns regarding its security maturity, as existing cybersecurity frameworks fail to evaluate security evolution at the protocol layer. Conventional models, such as the NIST Cybersecurity Framework (CSF) and Capability Maturity Model (CMM), primarily emphasize organizational governance, leaving a methodological gap in assessing protocol-level protection mechanisms. This study addresses that gap by proposing the MQTT Security Maturity Model (MQTT-SMM), a protocol-oriented framework derived from a systematic mapping study of MQTT security research published between 2021 and 2025. Four major security domains were identified authentication, authorization, secure communication, and resilience represent the progressive evolution of MQTT security. Focusing on authentication as a controlled case study, this study formalizes the advancement of authentication mechanisms into four maturity tiers aligned with the NIST CSF 2.0 (Partial, Informed, Repeatable, Adaptive). The proposed model provides a conceptual foundation for assessing MQTT implementation maturity and establishes a methodological baseline for future multi-domain expansion and validation within real-world IoT environments.**

Keywords—MQTT Security, Security Maturity Model, NIST Cybersecurity Framework, Protocol-Level Security, Systematic Mapping Study.

I. INTRODUCTION

The rapid growth of the Internet of Things (IoT) has transformed industries ranging from smart homes and healthcare to transportation and industrial automation. Billions of devices are now interconnected, generating massive volumes of data and requiring lightweight yet reliable communication protocols. Among the various protocols available, MQTT has become one of the most widely adopted standards due to its lightweight design, efficient bandwidth usage, and publish/subscribe communication model [1]. MQTT is used extensively across different domains, including sensor networks, supervisory control and data acquisition (SCADA) systems, and industrial IoT (IIoT) deployments [2] [3] [4]. Several surveys have consistently highlighted MQTT as one of the most popular IoT protocols in both academia and industry [5] [6]. Figure 1 shows the frequently used protocols.

However, this popularity has also expanded the attack surface, making MQTT implementations increasingly attractive to adversaries and exposing significant security challenges. By default, MQTT brokers often allow anonymous connections, rely on simple username–password authentication, and treat TLS as optional rather than mandatory. Previous studies have shown that MQTT implementations are vulnerable to unauthorized access, man-in-the-middle attacks, denial-of-service, and replay attacks [7].

Despite these efforts, most solutions are ad hoc, addressing specific vulnerabilities rather than providing a systematic way to evaluate overall security readiness. A large body of literature has focused on technical improvements, including authentication and credential management [8], authorization and access control [9], secure communication [10], and monitoring and resilience. According to internet-wide measurements using platforms such as Shodan—a cyberspace search engine that indexes publicly accessible devices and services—only a small fraction of MQTT deployments enforce TLS, and many instances remain misconfigured. Figure 2 illustrates this condition, showing that unsecured or improperly configured MQTT brokers continue to represent a significant portion of publicly accessible deployments.

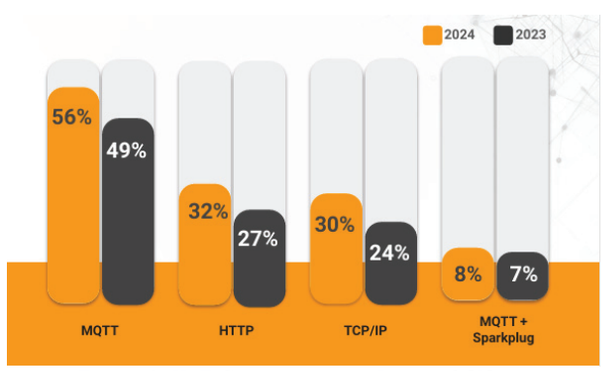


Figure 1. IIoT Communication Protocol [5]

In the context of communication technologies, the security of protocols such as MQTT has evolved progressively through continuous enhancements, including stronger authentication, encryption, and access control mechanisms. These improvements illustrate a gradual and adaptive

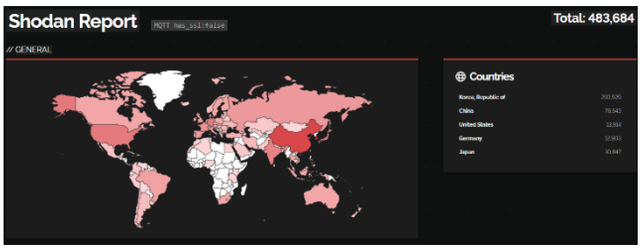


Figure 2. MQTT Server Has no SSL detected in Shodan [9]

response to emerging cybersecurity challenges, reflecting an evolutionary trajectory similar to process maturity in organizational settings. By systematically examining the progression of MQTT security features over time, this study seeks to formalize that evolution into distinct maturity levels, providing a structured understanding of how the protocol's security capabilities develop and strengthen in response to an increasingly complex threat landscape.

Existing cybersecurity maturity and assessment frameworks—such as ISO/IEC 27001 and the Secure Control Framework (SCF)—have been widely adopted to evaluate institutional readiness, risk governance, and organizational resilience [11], [12]. While these models effectively structure cybersecurity management at the strategic and operational levels, they inherently overlook the protocol layer, where vulnerabilities often emerge in interconnected and resource-constrained systems such as the Internet of Things (IoT). Existing hierarchical models tends to emphasize governance and process maturity rather than the technical evolution of protective mechanisms within communication protocols. Consequently, no current framework adequately captures how protocol-level security, such as authentication or encryption mechanisms, evolves and matures in response to emerging threats and implementation practices.

Building on earlier works that have developed cybersecurity maturity models [13] , [14], this study extends the conceptual reasoning behind their methodological construction. Prior frameworks generally define maturity by translating empirical observations of capability development into structured, incremental readiness levels. Adapting this line of thinking, our work applies the same analytical logic to the technical evolution of protocol security—specifically, how authentication mechanisms in MQTT have progressively advanced in response to changing threat landscapes and implementation practices. Through a controlled case study, MQTT is employed not as an isolated example but as a validation environment to demonstrate the generalizability of this maturity assessment approach across diverse communication systems and protocol architectures. Beyond formalizing this evolution, the study also introduces a theoretical framework for prioritizing future security enhancements derived from the assessed maturity levels. Collectively, this proposed model offers an integrative contribution to the IoT security domain: it establishes a replicable methodology for constructing protocol-level maturity frameworks while simultaneously reinforcing the vision of a secure and resilient MQTT ecosystem through the balanced strengthening of four essential security dimensions—authentication, authorization, secure communication, and resilience.

II. Research Methodology

This study begins by applying the *Systematic Mapping Study* (SMS) approach, which provides a structured method for identifying, classifying, and analyzing existing research related to MQTT security [15]. The process covers the entire sequence from the formulation of research questions to the construction of a systematic map, ensuring that the collected evidence is comprehensive, traceable, and aligned with the study's objective of identifying key domains of MQTT security enhancement.

A. Data Collection

At the initial stage of this study, data collection was conducted using two major academic databases: Scopus and IEEEXplore. The keyword "MQTT Secur*" was used to conduct the search, with the query restricted to the title field only to ensure that the retrieved studies are directly focused on MQTT security. The search results were filtered to include only conference papers and journal articles published between 2021 and 2025 to maintain relevance and focus on recent advancements. This filtering strategy ensures that the dataset reflects the most up-to-date research trends and technical developments in MQTT security within the IoT communication context.

Table I. Search Query

Source	Keyword	Limitation	Result
Scopus	MQTT Secur*	Document Title, 2021-2025, Conference Paper and Article	126
IEEEXplore		Document Title, 2021-2025, Conferences and Journal	101

B. Data Analysis

To systematically identify the dominant research themes within MQTT security studies, VOSViewer version 1.6.20 was used to perform text mining–based term mapping. The analysis was conducted by creating a map based on textual data extracted directly from the title and abstract fields of the selected publications. Data are imported from reference manager files, and the binary counting method is applied to ensure that each term is counted only once per document, thereby preventing frequency bias. This approach allows for the visualization of the most recurrent terms that characterize the MQTT Security research landscape.

From the combined dataset obtained from Scopus and IEEEXplore, 76 terms were extracted through text mining using VOSViewer. These terms were automatically grouped into four clusters to reflect the distribution of topics across the selected publications. To ensure analytical relevance, only terms directly related to MQTT security—such as *TLS*, *authentication*, and *encryption*—were retained, whereas unrelated or overly generic terms such as *industry*, *default*, and *machine* were excluded. After filtering, 37 relevant terms remained, which were then categorized into four primary domains representing the key areas of MQTT security enhancement: (1) authentication and credential management, (2) authorization and access control, (3) secure communication and encryption, and (4) monitoring, detection, and resilience. Figure 3 shows the resulting term map, highlighting frequently mentioned concepts, that serve as clear indicators of the major domains explored in MQTT security research and illustrate how current studies are distributed across these core areas of improvement.

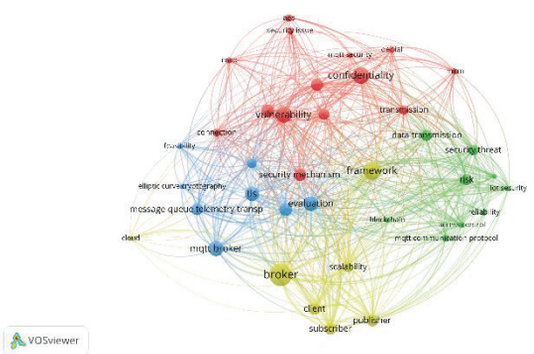


Figure 3. Result of Term Map.

C. Domain-Specific Literature Analysis

Following the identification of the four principal domains of MQTT security enhancement, a focused literature review was conducted exclusively through the IEEE Xplore digital library (2021–2025) to examine the mechanisms, techniques, and strategies proposed within each domain. From the 71 relevant papers, the studies were classified into authentication and credential management, authorization and access control, secure communication and encryption, and monitoring, detection, and resilience. The resulting keywords from the text-mining and filtering stages were then organized by domain to ensure clarity and reproducibility. This distribution highlights the diverse research focus across MQTT security and serves as the analytical basis for the subsequent synthesis presented in Table 3.

Table II. Distribution of Keywords and Papers Across MQTT Security Domains.

Domain	Keyword	Result
Authentication and Credential Management	MQTT Authentication	12
Authorization and Access Control	MQTT Authorization	6
Secure Communication and Encryption	MQTT Secure Communication Encryption	18
Monitoring, Detection, and Resilience	Resilience MQTT, Detection MQTT	34

This study identified four core areas reflecting the progressive evolution of MQTT security mechanisms through systematic mapping, text mining, and domain-specific classification, forming the conceptual foundation for the proposed MQTT Security Maturity Model. While the broader analysis revealed multiple security domains, this study focuses solely on authentication as the essential entry point of secure communication, allowing a more concentrated exploration of its evolutionary progression within scope and space limits.

III. Enhancement of Authentication Security and Mapping to NIST CSF

A. Authentication and Credential Management

Early studies on MQTT authentication revealed a shift from centralized to more adaptive and distributed mechanisms. Chien [16] employed an Edge Server (ES) to generate anonymous device identities via Composite Hashing, enabling only registered devices to communicate and mitigating DoS attacks. Fathan et al. [17] enhanced security by replacing static credentials with JSON Web Tokens (JWT) issued by a third-party server, reducing the risk of brute-force and sniffing risks. Nicolo et al. [18] advanced this with blockchain-based authentication using Algorand, where client identities are verified through edge nodes. Meanwhile, Rizka et al. [19] strengthened identity validation by combining RFID and fingerprint authentication with two-factor authentication. Together, these studies mark an evolution from basic credential systems to adaptive, decentralized, and layered authentication models in MQTT.

Subsequent studies on MQTT authentication reveal a move toward dynamic, cryptography-driven schemes. Virginia et al. [20] used QR codes as initial authentication media, embedding connection credentials for MQTT brokers and redirecting users to encrypted logins to prevent credential interception. Erukala et al. [21] combined blockchain and smart contracts to automate edge-device registration and key management via secp256k1, mitigating unauthenticated flooding, replay, and MitM attacks. Hasan et al. [22] integrated blockchain with the InterPlanetary File System (IPFS) to ensure decentralized identity integrity through unique client IDs. Yin et al [23] proposed a two-way authentication using Pre-Shared Keys (PSK) and Elliptic Curve Diffie–Hellman (ECDH) to generate unique session keys per communication. Collectively, these works demonstrate the

evolution of MQTT from static credential systems to adaptive, cryptography-based authentication mechanisms suited to modern IoT security demands.

In the latest phase of MQTT authentication development, research has advanced toward multi-layered architectures that combine cryptographic methods, Trusted Execution Environments (TEE), and blockchain-based trust management. Wu et al. [24] implemented full blockchain-based authentication, where all MQTT components—clients, brokers, and gateways—operate as nodes under a Practical Byzantine Fault Tolerance (PBFT) consensus, recording each message as a block to ensure integrity and resistance to impersonation, unauthorized access, replay, and brute-force attacks. Zhang et al. [25] strengthened broker-side security using TEE-based protection with mutual authentication via Identity-Based Encryption (IBE) between the Rich Execution Environment (REE) and TEE, producing temporary session keys for secure exchanges. Madiyar et al. [26] introduced a blockchain scheme that incorporate physical location as an authentication factor through an Adaptive Distance (AD) algorithm, validating devices based on proximity to authorized nodes to prevent external intrusion. Finally, Samad et al. [27] proposed an Authenticated Encryption with Associated Data (AEAD) mechanism that activates session keys only after mutual verification, enhancing confidentiality and replay attack resilience.

The evolution of MQTT authentication mechanisms appears to follow a chronological progression rather than a linear increase in technical robustness. Each approach reflects its period's research priorities and technological context. This study interprets these developments through a maturity-oriented lens aligned with the NIST Cybersecurity Framework (CSF) 2.0—namely Partial, Informed, Repeatable, and Adaptive—serving as the foundation for the proposed model in the next section to ensure consistent evaluation..

B. Mapping MQTT Authentication to NIST CSF

The NIST Cybersecurity Framework (CSF) 2.0 provides a structured yet flexible approach for managing cybersecurity risks through a set of high-level, outcome-oriented categories. Designed to be applicable across organizations of any size, sector, or maturity, it emphasizes adaptability and progressive improvement over prescriptive technical controls. This outcome-based nature makes the framework suitable for modeling the security evolution of technical systems, including communication protocols such as MQTT. By adopting the principles of the NIST CSF—particularly its maturity-oriented structure—this study aligns the proposed security model with an internationally recognized reference that emphasizes measurable progress and adaptive resilience [28].

The NIST CSF defines four implementation tiers—Partial, Risk-Informed, Repeatable, and Adaptive—that describe an organization's progression from ad hoc to continuously improving cybersecurity practices. Translating this concept to the protocol layer, the evolution of MQTT authentication observed in the literature can be aligned with these maturity levels. Early implementations relying solely on static credentials reflect the Partial tier, whereas the Risk-Informed stage is characterized by the introduction of token-based or external authentication. Subsequent approaches integrating

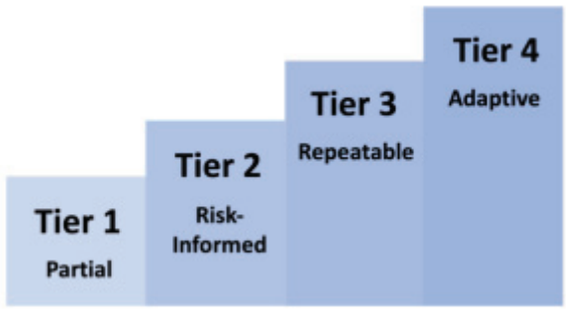


Figure 4. CSF Tiers for cybersecurity risk governance and management. [28]

cryptographic handshakes or mutual authentication mechanisms demonstrate Repeatable maturity, and recent works employing blockchain, trusted execution environments (TEE), or context-aware verification demonstrate Adaptive maturity. This alignment enables the proposed framework to express technical improvements in MQTT authentication as a structured maturity progression consistent with the NIST CSF philosophy. Accordingly, the next chapter builds upon this conceptual alignment to formalize the MQTT Authentication Security Maturity Model (MQTT-ASMM). The proposed model translates the observed evolutionary trends and their correspondence with NIST CSF tiers into a structured, progressive framework that defines measurable authentication maturity levels.

IV. PROPOSED MODEL

Building on the evolutionary analysis of authentication mechanisms presented in the previous chapter, this section introduces a conceptual maturity model designed to illustrate how MQTT authentication practices can be structured and assessed based on their level of sophistication. The model serves as a proof-of-concept, demonstrating how empirical findings can be translated into progressive maturity tiers aligned with the NIST Cybersecurity Framework (CSF) 2.0 — namely Partial, Risk-Informed, Repeatable, and Adaptive. The model reflects an incremental progression of capability rather than representing a chronological sequence, where each level embodies a higher degree of robustness, adaptability, and integration in securing MQTT-based communication. Although this model focuses exclusively on the authentication domain, it establishes the conceptual foundation for future expansion into a comprehensive, multi-domain MQTT Security Maturity Framework.

Table III. MQTT Authentication to NIST CSF Tier

Level	NIST CSF Tier Description	Mapping to the MQTT Authentication Context	Example of a Security Mechanism
Level 1 – Partial	Security practices are ad hoc and are-not formally based on objectives or the threat environment.	Authentication mechanisms are simple and isolated and rely on static credentials without	Basic username–password, plain MQTT connections without TLS, and no device

		verification or encryption. No centralized control or audit.	identity validation.
Level 2 – Risk-Informed	Organization or system begins to understand and manage cybersecurity risks; practices are documented but not standardized.	Authentication involves third-party or structured mechanisms, that provide initial mitigation against brute-force and replay attacks.	JWT-based authentication, token verification via edge or external authentication server.
Level 3 – Repeatable	Policies and controls are established and implemented consistently; processes are measurable and repeatable across environments.	Authentication adopts decentralized or distributed verification; identity is consistently validated using secure cryptographic methods.	Blockchain-based identity registration, composite hashing, and two-factor authentication (RFID + biometric).
Level 4 – Adaptive	Security processes are continuously improved using advanced intelligence and automation; the system adapts to evolving threats.	Authentication mechanisms integrate adaptive encryption, dynamic key management, or trusted hardware to enhance resilience.	Blockchain–TEE hybrid models, AEAD-based mutual authentication, and location-aware or context-sensitive validation.

The proposed maturity mapping demonstrates how the evolution of MQTT authentication mechanisms can be systematically aligned with the NIST CSF tiers, offering a structured representation of protocol-level security readiness. Although this study focuses solely on the authentication domain, it effectively serves as a proof of concept that the protocol layer can formalize cybersecurity maturity. Although the current scope does not yet capture the organizational breadth envisioned by the NIST CSF, it establishes a solid foundation for future expansion toward a multi-domain model encompassing authorization, secure communication, and resilience. The following section concludes reflecting on its implications, limitations, and prospective directions for advancing a comprehensive MQTT Security Maturity Model.

V. Conclusion

This study examines the progressive evolution of MQTT security mechanisms and formalize these advancements into a structured maturity perspective. Four major domains of MQTT security enhancement were identified through a systematic mapping approach. Focusing on the authentication domain as a representative case, twelve research papers were analyzed to trace how MQTT authentication evolved from simple static credentials to advanced mechanisms incorporating cryptography, blockchain, and trusted execution environments. The insights gained from this analysis provide a conceptual foundation for developing a maturity-based framework that captures the incremental improvement of MQTT security in alignment with the NIST Cybersecurity Framework (CSF).

Despite offering a structured perspective on the evolution of MQTT authentication mechanisms, this study is limited in scope, because it focuses solely on the authentication domain and lacks empirical validation. Consequently, its applicability is confined to a conceptual level. Future work will extend the framework to all four security domains while maintaining alignment with the NIST Cybersecurity Framework (CSF) 2.0. Using the Analytic Hierarchy Process (AHP), relative weights will be assigned to each domain to support a more balanced assessment. The enhanced model will be validated through expert review and testing on existing MQTT implementations establish it as both an evaluation tool and practical guidance for IoT engineers to systematically assess and improve the security maturity of MQTT-based systems.

Acknowledgment

This research was supported by Beasiswa Komdigi, Ministry of Communication and Digital of Indonesia (Kemkomdigi). Arfan Ridwan Hartawan is a researcher at the National Research and Innovation Agency (BRIN) Indonesia and is currently pursuing a Master's Program at the Universitas Indonesia funded by Kemkomdigi. The author acknowledges the use of artificial intelligence (AI)-based tools, such as grammar and sentence correction software, to improve the clarity and readability of the manuscript.

References

[1] OASIS, "MQTT Version 5.0," OASIS Standard, 2019.

[2] J. A.-F. Nael Radwan, "MQTT in Focus: Understanding the Protocol and Its Recent Advancements," *International Journal of Computer Science&Security (IJCSS),* vol. 18, no. 1, 2024.

[3] S. R. d. A. J. H. S. Ana Vitoria Freitas, "MQTT-VET: Exploring MQTT Protocol Vulnerabilities," in *Association for Computing Machinery*, New York, 2024.

[4] E. F. Pietro Colombo, "Access Control Integration in Sparkplug-Based Industrial Internet of Things Systems: Requirements and Open Challenges," in *Proceedings of the 20th International Conference on Web Information Systems and Technologies*, Marbella, 2024.

[5] E. Foundation, "2024 IoT & Embedded Developer Survey Report," Eclipse Foundation, 2024.

[6] I. o. p. M. Poonam Gupta, "A Survey of Application Layer Protocols for Internet of Things," in *IEEE International Conference on Communication information and ComputingTechnology (ICCICT)*, Mumbai, 2021.

[7] W. L. S. M. P. N. A. K. A.-A. S. K. SHAMS UL ARFEEN LAGHARI, "Securing MQTT Ecosystem: Exploring Vulnerabilities, Mitigations, and Future Trajectories," *IEEE Access,* vol. 12, pp. 139273-139289, 2024.

[8] D. M. A. V. M. N. S. R. K. Narasimha Swamy S, "Enabling Lightweight Device Authentication in Message Queuing Telemetry Transport Protocol,"

IEEE Internet of Things Journal, vol. 11, no. 9, pp. 15792-15807, 2024.

[9] G. D. U. C. C. Ozlem Seker, "MARAS:MutualAuthentication and Role-Based Authorization Scheme for Lightweight Internet of Things Applications," *MDPI,* vol. 23, no. 12, 2023.

[10] Z. Y. Shiyi Zhang, "Secure Communication Scheme of Internet of Things based on MQTT Protocol," in *5th International Conference on Neural Networks, Information and Communication Engineering (NNICE)*, 2025.

[11] ISO/IEC, "Information security, cybersecurity and privacy protection — Information security management systems — Requirements," ISO/IEC, Switzerland, 2022.

[12] S. C. Framework, "SCF OVERVIEW & PRACTITIONER GUIDEBOOK," SCF, 2025.

[13] M. G. Gulcin Buyukozkan, "Cybersecurity maturity model: Systematic literature review and a proposed model," *Technological Forecasting & Social Change,* vol. 213, 2025.

[14] E. B. M. C. Alessandro Brighente, "Are you Sure That is Secure? Assessing Organizations Security Readiness via a DevSecOps Maturity Model," in *2025 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, Venice, 2025.

[15] R. F. S. M. M. M. Kai Petersen, "Systematic mapping studies in software engineering," in *EASE'08: Proceedings of the 12th international conference on Evaluation and Assessment in Software Engineering*, Swindon, United Kingdom, 2008.

[16] H.-Y. Chien, "Highly Efficient Anonymous IoT Authentication Using Composite Hashing," in *IEEE Conference on Dependable and Secure Computing (DSC)* , 2021.

[17] R. R. P. P. S. Fathan Abdul Shodiq, "Secure MQTT Authentication and Message Exchange Methods for IoT Constrained Device," in *2021 International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA)*, 2021.

[18] V. D. A. M. M. S. G. V. A. V. Nicolo Cardamone, "Blockchain Based Public Key Authentication of IoT Devices for Electrical Energy Systems," in *2022 AEIT International Annual Conference (AEIT)*, Rome, 2022.

[19] V. S. H. H. N. R. Y. Rizka Reza Pahlevi, "Secure Two Factor Authentication for IoT Device," in *2022 10th International Conference on Information and Communication Technology (ICoICT)*, Bandung, 2022.

[20] V. M. F. D. H. N. Virginia Iskandar, "An Authentication Scheme for IoT Based Mechanical Relay Utilizing QR Code and MQTT," in *2023 IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob)*, Bali, 2023.

[21] A. A. D. I. K. B. S. Erukala Suresh Babu, "Blockchain based Authentication Mechanism for Edge Devices in Fog enabled IoT Networks," in *TENCON 2023 - 2023 IEEE Region 10 Conference (TENCON)*, Chiang Mai, 2023.

[22] K. K. Hasan Bahadir Karadas, "IBAM IPFS and Blockchain based Authentication for MQTT protocol in IoT," in *2023 IEEE Symposium on Computers and Communications (ISCC)*, Tunis, 2023.

[23] Y. L. Z. G. B. L. Hao Yin, "Lightweight Secure Identity Authentication Scheme for Restricted IoT Devices," in *2023 8th International Conference on Signal and Image Processing*, Wuxi, 2023.

[24] T. Z. Z. Y. Qiuhan Wu, "A Secure Communication Protocol Based on Blockchain and MQTT for IoT," in *2024 IEEE 7th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, Chongqing, 2024.

[25] D. L. F. Z. X. L. Y. C. R. W. H. Z. L. S. S. X. Fangzhe Zang, "Security Enhancement Method for MQTT Based on TEE," in *2024 International Conference on Networking and Network Applications (NaNA)*, Yinchuan City, 2024.

[26] Z. K. N. Y. M. M. Madiyar Seidullayev, "Improvement of the Authentication Method for IoT Technologies Based on Blockchain," in *2025 International Conference on Innovations in Intelligent Systems: Advancements in Computing, Communication, and Cybersecurity (ISAC3)*, Bhubaneswar, 2025.

[27] A. J. S. S. Y. B. N. B. P. P. Samad Rostampour, "Securing Industrial IoT A Novel Approach with MQTT Authentication," in *2025 IEEE International Conference on Communications (ICC): IoT and Sensor Networks Symposium*, Montreal, 2025.

[28] N. I. o. S. a. Technology, "The NIST Cybersecurity Framework (CSF) 2.0," National Institute of Standards and Technology, Gaithersburg, 2024.

[29] Steve, "Introduction to MQTT Security Mechanisms," Steve Cope, 10 10 2024. [Online]. Available: http://www.steves-internet-guide.com/mqtt-security-mechanisms/. [Accessed 13 10 2025].

UVASS-FDD: A Multimodal Deepfake Detection Framework Integrating Audio-Visual Consistency to Enhance Content Integrity

1st Laode Alif Ma'sum*, 2nd Prima Dewi Purnamasari†,
*Departemen Teknik Elektro, Universitas Indonesia, Depok, Indonesia
†Departemen Teknik Elektro, Universitas Indonesia, Depok, Indonesia
Emails: laode.alif@ui.ac.id, prima.dp@ui.ac.id

***Abstract*—This paper presents UVASS-FDD, a hybrid deepfake detection model that combines visual and audio features to enhance detection accuracy under varying video conditions. The model is trained on a balanced mix of the DFDC and Celeb-DF datasets, with optimized hyperparameters achieving a 89.26% test accuracy using a learning rate of 5×10^{-5}. Evaluation on the unseen LAV-DF dataset yields a 65% F1-score, indicating fair cross-dataset generalization. The results highlight the effectiveness of multimodal learning while emphasizing the need for more diverse training data and further optimization for real-time deepfake detection.**

***Keywords*—Deepfake Detection, Video Processing,**

I. Introduction

Recent advancements in Artificial Intelligence (AI), particularly Generative Adversarial Networks (GANs) have led to the rise of very realistic synthetic content, including images, audio and video [1]. One notable application of this technology is deepfake, which leverages machine learning to generate convincing fake videos [2]. Although initially used for entertainment, deepfakes have raised serious concerns due to their potential misuse in spreading misinformation, manipulating public opinion, and compromising digital integrity [3]. As deepfake become more sophisticated, detecting them through unimodal analysis, only using visual or audio features have been proven insufficient [4].

This study proposes a hybrid approach for deepfake detection that integrates both visual and audio features to enhance classification accuracy, while maintaining the ability to detect deepfakes even when audio is absent through consistency analysis of visual features. The proposed system is designed to detect inconsistencies, such as mismatches between lip movement and speech that often appear in manipulated videos [7]. The research specifically targets face-audio based deepfakes and utilizes the Deepfake Detection Challenge (DFDC) & Celeb DF dataset. The output is binary classification between "Real" and "Fake" videos with analysis limited to english-language audio.

The main objectives are to implement a two-stage detection architecture, identify performance-affecting factors (e.g., pre-processing steps and hyperparameters), and evaluate the model's generalization on standard benchmarks. This methodology includes literature review, data pre-processing, system design, model training, fine-tuning, evaluation using various metrics, and comprehensive result analysis.

This paper contributes a robust hybrid deepfake detection framework with improved resilience to high-fidelity manipulations. It also highlights the importance of synchronization analysis between modalities and provides insights for future research in trustworthy AI systems. The results are expected to aid in safeguarding digital content authenticity and mitigating the negative impacts of deepfake technology.

II. Literature Review

A. Deepfake

Deepfakes are synthetic media generated using deep learning models to manipulate faces, voices, or both [1]. Techniques such as lip-syncing, face-swapping, and text-to-speech make fabricated content highly realistic [8] [9]. Although deepfakes support creative applications, they also pose risks involving misinformation and reputational harm.

Deepfake generation mainly relies on autoencoders and Generative Adversarial Networks (GANs) [7]. Autoencoders learn latent representations that allow cross-identity reconstruction for face-swapping [10], while GANs train a generator and discriminator adversarially to produce photorealistic results [5] [6] [11].

As multimodal deepfakes grow more sophisticated, unimodal detection becomes insufficient. Multimodal approaches detect temporal and cross-modal inconsistencies—e.g., mismatched lip motion and speech—by jointly analyzing visual frame features and audio representations such as Mel-spectrograms derived via STFT.

B. Artificial Neural Networks

Artificial Neural Networks (ANNs) support hierarchical representation learning, enabling applications in vision, speech, and generative modeling. Convolutional Neural Networks

XXX-X-XXXX-XXXX-X/XX/$XX.00 © 20XX IEEE

(CNNs) extract spatial patterns using convolution, pooling, and fully connected layers [12] [13].

Activation functions introduce essential non-linearity. The linear activation remains:

$$y = x, \tag{1}$$

Common nonlinear functions include sigmoid (2), tanh (3), ReLU (4), and Softplus (5):

$$A(x) = \frac{1}{1 + e^{-x}} \tag{2}$$

$$f(x) = \tanh(x) = \frac{2}{1 + e^{-2x}} - 1 \tag{3}$$

$$A(x) = \max(0, x) \tag{4}$$

$$A(x) = \log(1 + e^{x}) \tag{5}$$

Loss functions such as cross-entropy [7], contrastive loss, and auxiliary losses guide optimization and improve class separation and feature consistency across stages.

C. Predictive Visual-Audio Alignment Self-Supervision for Multimodal Deepfake Detection (PVASS-MDD)

PVASS-MDD combines Predictive Visual-Audio Alignment Self-Supervision (PVASS) with a multimodal classifier to detect forged videos by modeling audio-visual correspondence [4]. PVASS learns natural temporal alignment between lip motion and speech, which the MDD module later uses for deepfake classification.

1) Swin Transformer: The Swin Transformer functions as the visual encoder, applying local window attention and shifted-window interaction for efficient hierarchical modeling [11]. Input video ($T \times H \times W \times 3$) is divided into 3D patches and embedded into tokens [14]. Each block uses 3D shifted window multi-head self-attention and a feed-forward network with residual connections:

$$\begin{aligned} \hat{z}^l &= \text{3DW-MSA}(\text{LN}(z^{l-1})) + z^{l-1}, \\ z^l &= \text{FFN}(\text{LN}(\hat{z}^l)) + \hat{z}^l. \end{aligned} \tag{6}$$

Attention incorporates 3D relative position bias:

$$\text{Attention}(Q, K, V) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d}} + B\right)V. \tag{7}$$

Swin variants—T, S, B, and L—differ in channel size and layer depth.

2) TorchAudio: TorchAudio provides GPU-accelerated audio feature extraction, including models such as Wav2Vec 2.0, HuBERT, and WavLM [15]. In PVASS-MDD, it converts raw audio into Mel-spectrograms that capture spectral–temporal patterns and support alignment with visual features. This cross-modal alignment enhances detection of inconsistencies characteristic of deepfake forgeries.

III. System Overview and Implementation

The system architecture, illustrated in Fig.4, is conceptually inspired by the PVASS-MDD framework [7]. The model takes a video file as input and produces a classification output indicating whether the video is real or manipulated. It follows a two-stage design: in the first stage, the model learns audio-visual (AV) correspondence through a self-supervised training mechanism. Once trained, the UAVSS (Unified Audio-Visual Self-Supervision) module is frozen and subsequently utilized by the FDD (Flexible Deepfake Detection) module in the second stage to identify multimodal or unimodal inconsistencies for deepfake detection.

1) Unified Audio-Visual Self Supervision (UAVSS): The UAVSS pipeline begins with video augmentation, followed by the seperation of visual and audio components. Visual data is augmented using random flipping and face regions are extracted using MTCNN. If multiple faces are detected, the largest face will be selected. The cropped face is resized to 224x224 pixels and passed through the Swin Transformer, which has the classification head removed to only extract the visual features.

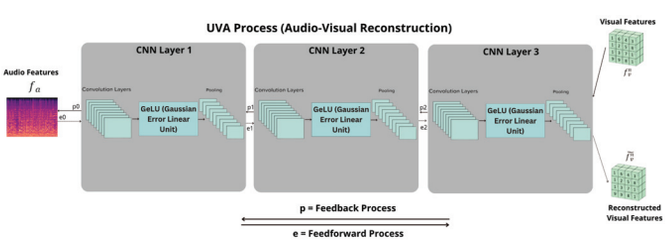


Fig. 1. UVA Block Architecture for Visual-audio Video.

The extracted features are processed in the Unified Visual-Audio (UVA) alignment block shown in Fig. **??** and Fig. 2. For videos containing both audio and visual streams, the UVA block consists of feedback and feedforward networks. The feedback networks reconstructs audio features from visual features, while the feedforward network reconstructs visual features from audio features. Each networks employ 1 to 3 convolutional layers with GELU activation and average pooling. The UVA process learns cross-modal alignment through back-and-forth reconstruction, gradient flow modulation, and a top-down prediction mechanism.

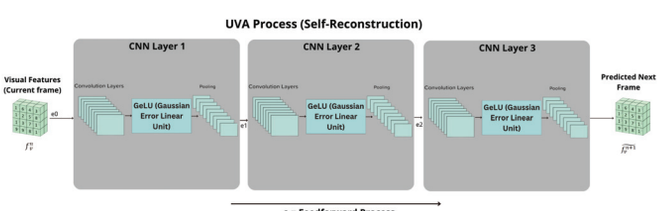


Fig. 2. UVA Block Architecture for Visual-Only Video.

In the case where the input video does not have audio, the UVA component switches to a unimodal prediction mode. In this setting, the feedforward network predicts the next

visual frame based on preceding frames, enabling the model to preserve temporal consistency and detect manipulation artifacts even in the absence of audio information.

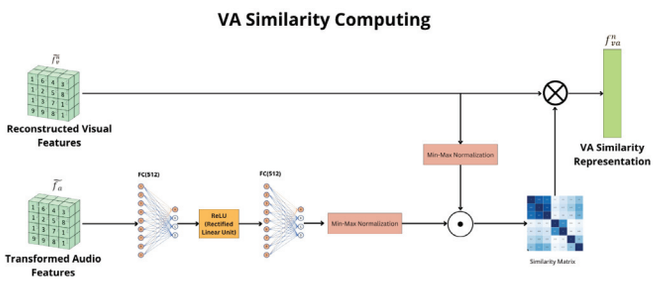


Fig. 3. VASC Block Architecture.

The VA Similarity Computing (VASC) block computes spatial similarity between modalities. Audio features are transformed non-linearly using two-layer MLP with ReLU activation. Then, pairwise similarity is calculated with visual features to obtain a similarity map, which is used to generate a VA representation. The contrastive objective designed to enforce strong cross-modal correspondence can be formally defined by the following expression:

$$\mathcal{L}pvass = \frac{1}{2}\mathcal{L}ncs(z^1, SG(z^2)) + \frac{1}{2}\mathcal{L}_{ncs}(z^2, SG(z^1)) \quad (8)$$

2) Flexible Deepfake Detection (FDD): The FDD module performs classification by using the frozen UVASS output to detect cross-modal inconsistencies. The input undergoes preprocessing steps identical to UVASS: visual augmentation, face detection and cropping, visual extraction via Swin Transformer, and audio extraction via TorchAudio.

The features are passed through the VASC module again to generate a similarity map. This map captures alignment discrepancies, which are critical to classify videos as real or fake. The classification head is trained using a composite loss function:

$$\mathcal{L}_{\text{total}} = \alpha\mathcal{L}_{ce} + \beta\mathcal{L}_{con} + \gamma\mathcal{L}_{aux} + \delta\mathcal{L}_{sr} \quad (9)$$

This module utilizes four key loss functions to optimize deepfake detection. The Cross-Entropy Loss ($\mathcal{L}CE$) serves as the primary objective function for binary classification by penalizing incorrect predictions, guiding the model to differentiate between real and manipulated videos. The Contrastive Loss ($\mathcal{L}Contrastive$) enhances the model's ability to learn discriminative multimodal features by encouraging aligned audio-visual pairs to have similar representations, while pushing apart mismatched pairs. Additionally, the Auxiliary Loss ($\mathcal{L}_{Auxiliary}$) provides a supplementary supervision signal by guiding intermediate representations during training, which help to stabilize learning and accelerates convergence.

To extend the framework capabilities for unimodal scenarios, the Self-Reconstruction Loss ($\mathcal{L}sr$) is introduced. This objective supervises the prediction of future visual frames from previous ones in cases where audio features are unavailable. By enforcing temporal consistency within the visual stream, $\mathcal{L}sr$ enables the model to detect subtle frame-level inconsistencies that are characteristic of deepfake manipulations, ensuring robust detection even in the absence of audio information.

3) Training and Evaluation: The model is trained using the Deepfake Detection Challenge (DFDC) and Celeb-DF dataset, split into 85% data for training and 15% for testing [16] [18]. The training process use Adam W optimizer and adjustable hyperparameters such as learning rate, batch size, and epoch count. To evaluate generalization capabilities, the model is tested on the Localized Audio Visual Deepfake Dataset (LAV-DF). LAV-DF includes three manipulation categories: Fake A/V, Fake A/Real V, and Real A/Fake V, challenging the model to detect partial deepfakes [17]. Metadata labelling was adjusted during the pre-processing stage.

The model was trained using a combination of the Deepfake Detection Challenge (DFDC) and Celeb-DF datasets, with an 85%-15% split for training and testing [16] [18]. The training used the AdamW optimizer with adjustable hyperparameters, including learning rate, batch size, and number of epochs. To assess the model's generalization capability, it was further evaluated on the Localized Audio-Visual Deepfake Dataset (LAV-DF), which contains three manipulation categories: Fake A/V, Fake A/Real V, and Real A/Fake V. These categories present a complex scenario by introducing partial deepfakes that manipulate either the audio or the video independently [17]. Metadata labels were refined during the pre-processing stage to ensure consistency across datasets.

Experimental results include intra-dataset and inter-dataset testing to assess robustness and generalization. The metrics used to evaluate the model performance are accuracy, loss, precision, recall, and F1-score.

IV. Implementation and Analysis

This chapter describes the implementation process, test results, and analysis of the deepfake detection system. The main focus is on evaluating the model's performance through a series of test scenarios to identify the most effective configuration.

A. Experimental Setup

Experiments were conducted on system with an AMD Ryzen 7 5700X CPU and NVIDIA GeForce RTX 4090 GPU. The software stack was built on PyTorch (v2.2.2) and CUDA (v12.1).

This research employs a balanced version of the DFDC (Deepfake Detection Challenge) dataset. Originally, the dataset exhibited a significant class imbalance, with a ratio of approximately 5.22:1 between "FAKE" and "REAL" videos. To mitigate this issue, the majority class was downsampled, resulting in an equal number of samples for both categories—19,154 videos each. The final dataset used for experimentation comprises a total of 38,309 videos, which were divided into training and testing sets with an 80:20 split, corresponding to 30,636 and 7,673 videos, respectively.

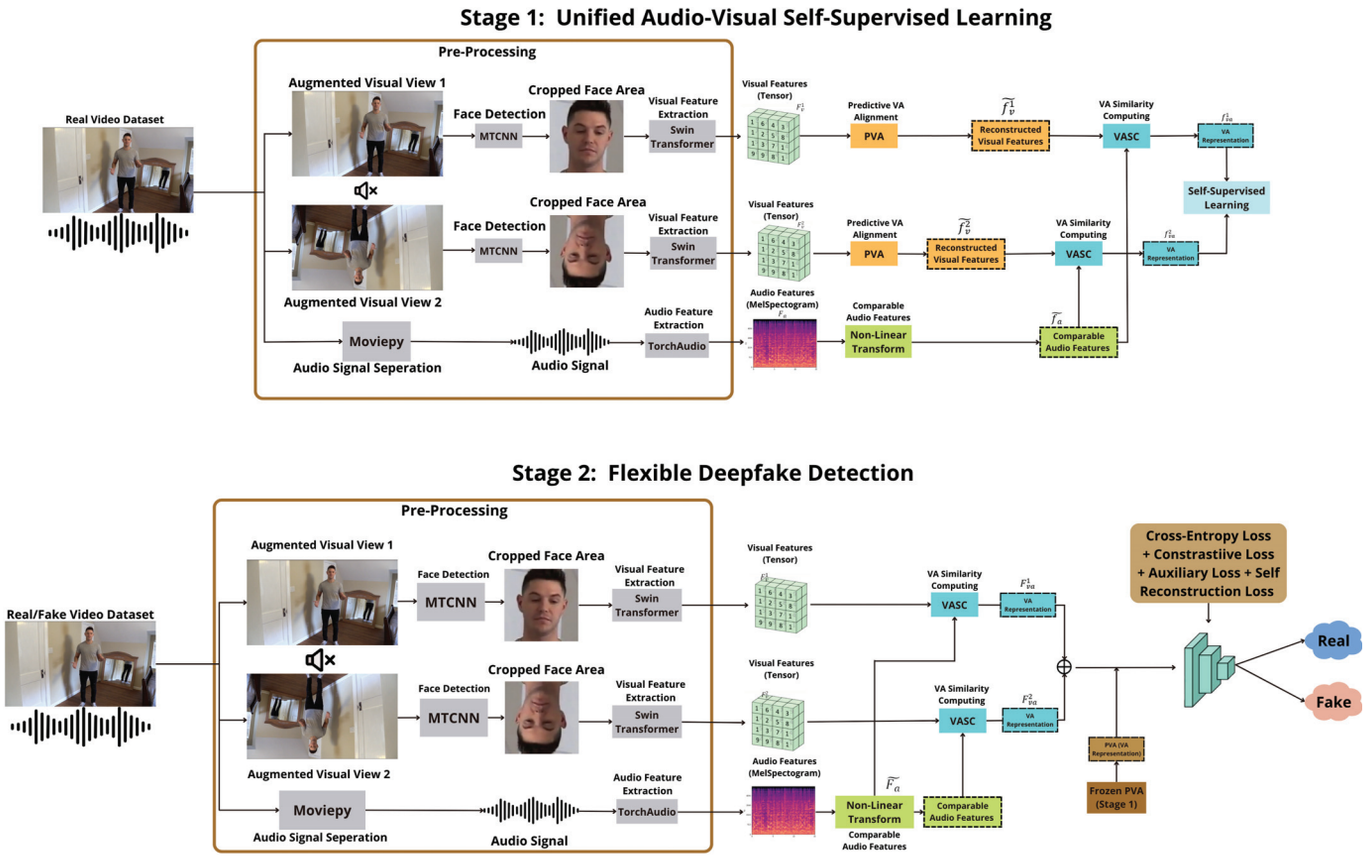


Fig. 4. UVASS-FDD system overview.

To enhance the model's generalization and enable evaluation without audio, the balanced DFDC dataset was combined with the Celeb-DF dataset, which consists of 5,608 videos that do not contain audio tracks. This combination allows the model to learn from both audio-visual and visual-only samples, strengthening its robustness in scenarios where audio information is absent.

The pre-processing stage involves the following steps:

1) **Face Detection and Cropping**: Using MTCNN, the main subject's face is detected, and the video is cropped to a 224x224 pixel resolution to focus on the lip area.
2) **Visual Feature Extraction**: The Swin Transformer model is used to extract visual features from cropped video frames.
3) **Audio Feature Extraction**: The VGGish model is used to extract audio features from the original sound signal.

B. Evaluation Metrics

The model's performance was assessed using standard classification metrics. In this context, the "positive" class refers to "FAKE" videos, which are the primary target for detection.

- **Accuracy**: Measures the overall proportion of correctly classified videos. It provides a general sens of the model's correctness.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Videos}}$$

- **Precision**: Measures the proportion of videos predicted as FAKE that were actually FAKE. High precision is crucial for minimizing false alarms (i.e., not misclassifying real videos).

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

- **Recall**: Measures the proportion of all actual FAKE videos that the model successfully identified. High recall is important to minimize missed detections.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

- **F1-Score**: The harmonic mean of precision and recall. It provides a single, balanced measure of a model's performance, especially useful when the cost of false positives and false negatives are both significant.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **Loss**: A value calculated during training that represents the model's error. A lower loss value indicates a model that is better fitted to the data.

C. Comparative Experiments and Analysis

Experiments were conducted in several scenarios to evaluate various aspects of the model.

TABLE I
MODEL PERFORMANCE ACROSS PARAMETER VARIATIONS

Learning Rate	Layers	Train Acc.	Test Acc.	Test Prec.
0.001	1	89.32	83.68	0
	2	90.49	83.68	0
	3	90.49	83.68	0
0.0002	1	94.33	87.19	86.96
	2	88.95	83.68	0
	3	89.21	83.68	0
0.00005	**1**	**97.86**	**89.26**	**87.9**
	2	98.8	89.06	87.9
	3	96.53	89.3	86.9

1) Hyperparameter Optimization: The initial experiment was conducted on a smaller dataset (2,418 videos) to determine the most effective combination of hyperparameters. Three learning rates (0.001, 0.0002, 0.00005) and three CNN layer configurations (1, 2, 3) were evaluated, as summarized in Table I. The results indicate that models that were trained with smaller learning rate (0.00005) consistently achieved superior performance across all evaluation metrics. In particular, the configuration with a single convolutional layer attained the highest test accuracy of 89.26% and a training accuracy of 97.86%, while maintaining a balanced precision value of 87.9%. On the contrary, larger learning rates exhibited noticeable performance degradation and instability, with minimal gains observed from increasing the number of convolution layers. These finding suggest that a lower learning rate facilitates more stable convergence and improved generalization, whereas deeper configurations provide limited benefits for this dataset size.

TABLE II
MODEL PERFORMANCE ACROSS NUMBER OF LAYERS VARIATIONS

Num Layer	Train Accuracy	Train Loss	Test Accuracy	Test Precision	Test Recall	Test F1	Training Time
1	97.86	0.109	89.262	0.879	0.882	0.880	06:19:58
2	98.806	0.067	89.065	0.879	0.877	0.878	06:54:22
3	96.532	0.176	89.301	0.869	0.893	0.880	07:53:54

Further analysis was conducted to evaluate the effect of increasing the number of layers on model performance using the best learning rate configuration (5×10^{-5}). As presented in the Table II, the test accuracies for the one, two, three layer configurations were 89.26%, 89.06%, and 89.3%, respectively, while the corresponding training time increased from approximately 6 hours 19 minutes to nearly 8 hours. Statistical significance testing revealed that the differences in test accuracy among these three configurations were within one standard deviation (approximately $\pm 0.12\%$), indicating that the variations of number of layers were not statistically significant ($p > 0.05$) to the model's performance. This suggest that adding more layer does not contribute to a meaningful improvement in model generalization and instead result in higher computational cost. Therefore, considering both accuracy stability and training efficiency, the single-layer configuration was identified as the optimal design choice for subsequent experiments.

2) Pre-Processing Method Comparison: This experiment was conducted to compare two different technique in handling video duration variations: *zero padding* to standardize length to 10 seconds and *data filtering* to remove videos shorter than 10 seconds.

TABLE III
PERFORMANCE COMPARISON OF PRE-PROCESSING METHODS

Method	Train Accuracy	Test Accuracy
Zero Padding	97.44%	88.23%
Data Filtering	**97.86%**	**89.26%**

The results presented in Table III demonstrate that the *data filtering* approach produced slightly higher and more consistent performance than the *zero padding* method, achieving a test accuracy of 89.26% compared to 88.23%. This improvement can be attributed to the fact that data filtering preserves the integrity of the original video content by removing samples that fall below target duration, whereas zero padding introduces artificial frames containing no meaningful information. Such non-informative segments may dilute temporal patterns and hinder the model's ability to learn discriminative features effectively. Based on these findings, data filtering was selected as the standard pre-processing technique for subsequent experiments.

3) Generalization Capability Test: To evaluate the model's generalization capability, the best-performing model configuration was tested on a completely unseen dataset, **LAV-DF**. Different from the training dataset (DFDC & Celeb-DF), LAV-DF features a different type of manipulation, where deepfake alterations are applied only to short portions of each video rather than the entire duration. To accommodate this characteristic, the detection process was modified to operate on a smaller video segments consisting of 60 frames (approximately two seconds). A video is classified as a deepfake if any of its segments are identified as manipulated.

TABLE IV
MODEL EVALUATION RESULTS ON THE LAV-DF DATASET

Testing Dataset	Accuracy	Precision	Recall	F1-Score
LAV-DF	55.4%	61.4%	69.0%	65.0%

The results presented in Table IV indicate that the model's generalization capability remains limited, achieving an F1-score of 65% when evaluated on the LAV-DF dataset. This performance represents a noticeable decrease compared to the result obtained on the DFDC and Celeb-DF dataset, indicating that the model struggles to adapt to unseen manipulation characteristics. One contributing factor to this decline may be the distinct characteristics of LAV-DF, in which deepfake are applied only to brief segments of each video rather than the entire duration. Because the model was primarily trained on datasets containing full-length manipulations, it may not have effectively learned to detect subtle, short-term inconsistencies. This discrepancy underscores the need for enhanced temporal sensitivity and mroe diverse training data that include localized or partial deepfakes to improve overall model generalization.

V. Conclusion and Future Works

A. Conclusion

This research proposed a hybrid deepfake detection framework that integrates both visual and audio modalities to enhance detection robustness, even when one modality is absent. Through extensive experiments on the balanced DFDC and Celeb-DF datasets, the model achieved optimal performance with a single-layer CNN and a learning rate of 5×10^{-5}, balancing accuracy and efficiency. Increasing the number of layers offered no significant improvement, while data filtering proved superior to zero padding as a pre-processing method.

When evaluated on the unseen dataset, LAV-DF, the model attained an F1-score of 65%, showing moderate generalization. However, the noticeable performance drop indicates difficulty in detecting manipulations of shorter duration, likely due to the training data being dominated by full-length deepfakes.

B. Future Works

Based on the findings of this study, several directions for future research can be identified:

1) **Improving Generalization with Diverse Data**: The most critical next step is to train the model on a larger and more diverse aggregation of datasets. This will expose the model to a wider variety of manipulation techniques from different generators, directly addressing the overfitting issue and improving its real-world robustness.
2) **Model and Pipeline Optimization**: Further work can be done on optimizing the unimodal components of the architecture and fine-tuning the model to enhance its ability to recognize modern deepfake patterns. Additionally, optimizing the data pipeline could reduce classification latency, making the system more viable for real-time applications.
3) **Real-World Integration and Evaluation**: Future efforts should include testing the model's performance on a continuous stream of new deepfake examples and exploring its integration into platforms like social media for automated content filtering and verification.

References

[1] Yogesh Patel, Sudeep Tanwar, Rajesh Gupta, Pronaya Bhattacharya, Innocent Ewean Davidson, Royi Nyameko, Srinivas Aluvala, and Vrince Vimal. Deepfake generation and detection: Case study and challenges. IEEE Access, 11:143296–143323, 2023.

[2] Gosse, Chandell, and Jacquelyn Burkell. "Politics and Porn: How News Media Characterizes Problems Presented by Deepfakes." Critical Studies in Media Communication, vol. 37, no. 5, 19 Oct. 2020, pp. 497–511, www.tandfonline.com/doi/full/10.1080/15295036.2020.1832697, https://doi.org/10.1080/15295036.2020.1832697.

[3] Andrei O. J. Kwok and Sharon G. M. Koh. Deepfake: a social construction of technology perspective. Current Issues in Tourism, 24:1798–1802, 7 2021.

[4] Hanyi Wang, Zihan Liu, and Shilin Wang. Exploiting complementary dynamic incoherence for deepfake video detection. IEEE Transactions on Circuits and Systems for Video Technology, 33:4027–4040, 8 2023.

[5] Dave Bergmann and Cole Stryker. What is an autoencoder?, 11 2023.

[6] Katarya, Rahul, and Anushka Lal. "A Study on Combating Emerging Threat of Deepfake Weaponization." 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 7 Oct. 2020, https://doi.org/10.1109/i-smac49090.2020.9243588.

[7] Yang Yu, Xiaolong Liu, Rongrong Ni, Siyuan Yang, Yao Zhao, and Alex C. Kot. PVASS-MDD: Predictive visual-audio alignment self-supervision for multimodal deepfake detection. IEEE Transactions on Circuits and Systems for Video Technology, 34:6926–6936, 8 2024.

[8] Moaiad Ahmed Khder, Samer Shorman, Duaij Thabit Aldoseri, and Mo- hammed Mansoor Saeed. Artificial intelligence into multimedia deepfakes creation and detection. In 2023 International Conference on IT Innovation and Knowledge Discovery (ITIKD), pages 1–5. IEEE, 3 2023.

[9] Lyu, Siwei. "Deepfake Detection: Current Challenges and next Steps." IEEE Xplore, 2020, ieeexplore.ieee.org/document/9105991.

[10] Michał Zendran and Andrzej Rusiecki. Swapping face images with generative neural networks for deepfake technology– experimental study. Procedia Computer Science, 192:834–843, 2021.

[11] Liu, Ze, et al. "Video Swin Transformer." ArXiv (Cornell University), 1 Jan. 2021, https://doi.org/10.48550/arxiv.2106.13230.

[12] IBM. Convolutional neural networks, October 2021.

[13] Merve Noyan. Video processing basics, 2024.

[14] Hwang, Jeff, et al. "TorchAudio 2.1: Advancing Speech Recognition, Self-Supervised Learning, and Audio Processing Components for Pytorch." 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 16 Dec. 2023, https://doi.org/10.1109/asru57964.2023.10389648. Accessed 7 Apr. 2025.

[15] Gao, Zhihao, et al. Environmental Sound Classification Using CNN Based on Mel-Spectogram. 1 June 2023, pp. 41–45, https://doi.org/10.1109/aibt57480.2023.00015. Accessed 29 June 2025.

[16] Dolhansky, Brian, et al. The DeepFake Detection Challenge (DFDC) Dataset. 12 June 2020, https://doi.org/10.48550/arxiv.2006.07397.

[17] Cai, Zhixi, et al. "Glitch in the Matrix: A Large Scale Benchmark for Content Driven Audio–Visual Forgery Detection and Localization." Computer Vision and Image Understanding, vol. 236, Nov. 2023, p. 103818, arxiv.org/pdf/2305.01979.pdf, https://doi.org/10.1016/j.cviu.2023.103818.

[18] Y. Li, X. Yang, P. Sun, H. Qi and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 3204-3213, doi: 10.1109/CVPR42600.2020.00327. keywords: Videos;Visualization;Image color analysis;Decoding;YouTube;Training;Detection algorithms,

Evaluation of ML Models Using Feature Scaling, ADASYN and K-Means Clustering for Intrusion Detection System

1st Faizal Riza *, 2nd Iskandar Fitri †, 3rd Sumijan ‡
* Department of Computer Science, *Universitas Putra Indonesia YPTK Padang, Padang, Indonesia*
† Department of Computer Science, *Universitas Putra Indonesia YPTK Padang, Padang, Indonesia*
‡ Department of Computer Science, *Universitas Putra Indonesia YPTK Padang, Padang, Indonesia*
Emails: *s3.faizal@gmail.com, if@upiyptk.ac.id, sumijan@upiyptk.ac.id*

***Abstract*—This paper presents a comprehensive machine learning pipeline for network intrusion detection using the CICIDS2017 dataset. We integrate K-Means meta-feature clustering, ADASYN for class imbalance handling, and ensemble learning to enhance detection performance. Five preprocessing configurations were evaluated across five classification algorithms, including Random Forest, Extra Trees, Decision Tree, XGBoost, and Voting Ensemble. XGBoost achieves optimal performance with 99.9253% accuracy using Standard Scaler, ADASYN, and K-Means clustering. Confusion matrix analysis reveals XGBoost's exceptional precision (58 false positives) and Voting Ensemble's superior recall (443 false negatives). Systematic evaluation demonstrates ADASYN's effectiveness in handling class imbalance and meta-feature clustering's contribution to enhanced classification performance. Our approach provides practical guidance for deploying high- accuracy intrusion detection systems with balanced computational efficiency.**

Keywords—Network Intrusion Detection, Machine Learning, ADASYN, K-Means Clustering, Ensemble Learning, CICIDS2017

I. INTRODUCTION

The exponential growth of cyber threats has made Intrusion Detection Systems (IDS) critical for cybersecurity. Traditional signature-based IDS approaches struggle to detect novel attacks as they rely on known attack patterns. This limitation has prompted researchers to explore machine learning-based solutions that can identify complex patterns and adapt to evolving threats [1]. Machine learning algorithms enhance intrusion detection by learning from historical patterns and adapting to evolving threats [1]. The CICIDS2017 dataset, containing over 80 features and various attack types including DoS, DDoS, and brute force, has become a standard benchmark [2]. However, ML effectiveness is significantly influenced by data preprocessing, particularly for imbalanced datasets and features with varying scales.

Class imbalance, where benign traffic vastly outnumbers malicious instances, represents a fundamental IDS challenge. ADASYN addresses this by generating synthetic minority-class samples with density-based adaptation [3]. K-means clustering provides effective pattern discovery by grouping similar traffic patterns while reducing dimensionality [4], and standardization ensures features contribute proportionally regardless of scale [5]. Ensemble methods (Random Forest, Extra Trees, XGBoost) excel at handling high-dimensional data and non-linear relationships. Soft voting aggregates multiple model predictions, achieving superior performance [6]. However, existing works typically address class imbalance, feature scaling, or clustering independently, without evaluating their synergistic effects [7].

This study evaluates machine learning models for intrusion detection on CICIDS2017, integrating K-means clustering, ADASYN, and standardization. We investigate five classification approaches (Random Forest, Decision Tree, Extra Trees, XGBoost, and soft voting ensemble) to determine the optimal combination that maximizes detection accuracy while minimizing false positives.

II. RELATED WORK

Machine learning techniques have been extensively applied to network intrusion detection, including Decision Trees, Random Forests, and Support Vector Machines. Recent studies have explored ensemble methods and embedded feature selection techniques, demonstrating superior performance [8]. These approaches have shown significant improvements in detection accuracy and adaptability to evolving attack patterns.

Class imbalance represents a fundamental challenge in intrusion detection, as normal traffic typically outnumbers attack instances. Adaptive Synthetic Sampling (ADASYN) addresses this by generating samples in low-density regions, adjusting synthetic sample generation based on local density to reduce overfitting and improve dataset distribution uniformity [9]. This density-based approach enables more effective handling of minority class instances compared to traditional oversampling methods.

High-dimensional network traffic data requires effective feature engineering techniques. Recent advances in ensemble feature selection have shown promise [10], demonstrating that combining multiple feature selection methods can identify more robust feature subsets. However, clustering-based meta-feature creation remains underexplored in the context of intrusion detection systems, presenting an opportunity for novel approaches that leverage unsupervised pattern discovery.

The CICIDS2017 dataset, developed by the Canadian Institute for Cybersecurity, has become a standard benchmark for evaluating intrusion detection systems [11]. It contains 2.8 million network flows with over 80 features, covering various attack types including DoS, DDoS, brute force, infiltration, and botnet attacks. Despite its widespread adoption, most existing works have evaluated preprocessing techniques and classification algorithms independently, without

comprehensive analysis of their synergistic effects in real-world scenarios.

III. PROPOSED METHODOLOGY

Figure 1 illustrates the complete architecture of our proposed intrusion detection system. The methodology consists of several interconnected stages, each designed to address specific challenges in network intrusion detection.

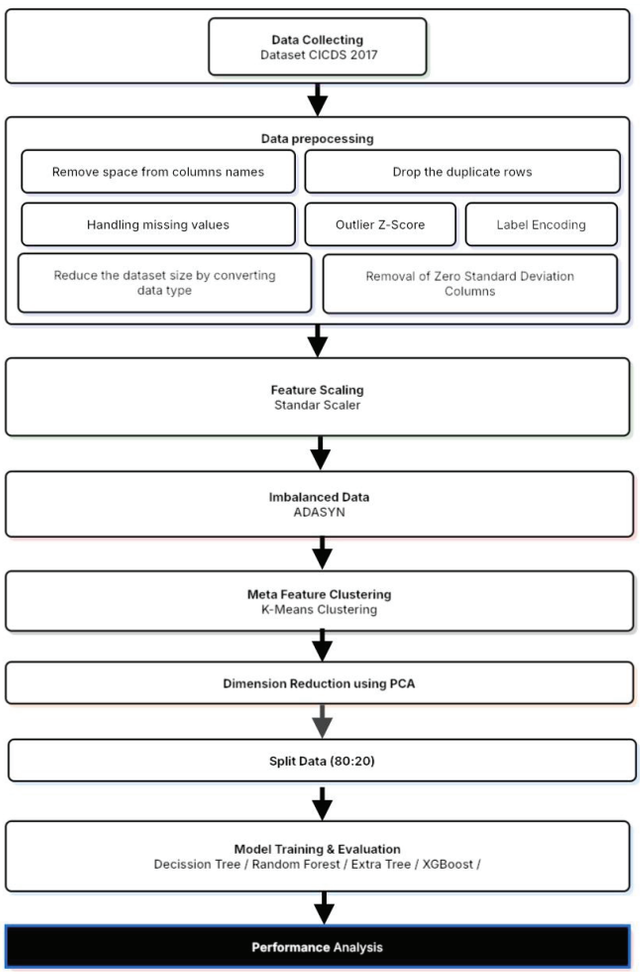


Fig. 1. Proposed Methodology Framework

A. Data Collection

We utilize the CICIDS2017 dataset [11], which contains network traffic data captured over five days, representing realistic scenarios with various attack types. The dataset includes 80 features extracted from network flows, providing comprehensive information about traffic characteristics.

B. Data Preprocessing.

1) Column Name Standardization: We remove whitespace from column names to ensure consistency and prevent parsing errors during subsequent processing steps [12].

2) Duplicate Row Removal: Duplicate records are identified and removed to prevent bias in model training and ensure data quality [13].

3) Missing Value Handling: Missing values are imputed using appropriate strategies based on feature characteristics. Numerical features use median imputation, while categorical features employ mode imputation [14].

4) Outlier Detection and Treatment: Outlier Detection and Treatment: We apply Z-Score based outlier detection with a threshold of ±3 standard deviations [15]. Outliers are either removed or capped to boundary values to prevent their disproportionate influence on model training.

5) Label Encoding: Categorical features are transformed using label encoding, converting string labels into numerical representations suitable for machine learning algorithms [16].

6) Data Type Conversion: Data Type Conversion: To reduce memory footprint and improve computational efficiency, we convert data types appropriately following precision reduction techniques [17]. Integer features are downcast to smaller integer types where possible, and float features are converted to float32 when precision allows.

7) Zero Standard Deviation Removal: Features with zero standard deviation provide no discriminative information and are removed to reduce dimensionality and prevent numerical instability [18].

C. Feature Scaling

We apply StandardScaler normalization [19], which transforms features to have zero mean and unit variance. This scaling technique is expressed as:

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

where x is the original feature value, μ is the mean, σ is the standard deviation, and z is the scaled value.

D. Imbalanced Data Handling with ADASYN

ADASYN (Adaptive Synthetic Sampling) addresses class imbalance by generating synthetic minority class samples. Unlike SMOTE, ADASYN uses a density distribution to determine the number of synthetic samples for each minority class instance. The algorithm focuses on generating more samples in regions where minority class instances are sparse, improving decision boundary definition in difficult regions [9].

E. Meta-Feature Clustering

We introduce a novel meta-feature engineering approach using K-Means clustering [4][19]. K-Means partitions the feature space into K clusters by minimizing the within-cluster sum of squares. After clustering, we assign each instance to its nearest cluster, creating a new categorical feature. This cluster membership encodes complex non-linear relationships in the original feature space, potentially capturing interaction effects that individual features cannot express. We also compute the Euclidean distance from each instance to all cluster centroids, creating K additional numerical features.

F. Dimensionality Reduction with PCA

Principal Component Analysis (PCA) reduces dimensionality while retaining maximum variance [19]. We retain principal components with eigenvalues greater than 1 using the Kaiser criterion, balancing information preservation with dimensionality reduction.

G. Data Splitting

The preprocessed dataset is split into training (80%) and testing (20%) sets using stratified sampling to maintain class distribution proportions. This approach ensures balanced representation across different attack classes and allows fair comparison between solutions [20].

H. Machine Learning Models

We evaluate five classification algorithms: Decision Tree recursively partitions the feature space using information gain or Gini impurity. Random Forest [5][19] builds an ensemble of trees on bootstrap samples with random feature subsets. Extra Trees [21] introduces additional randomization through random split thresholds, enhancing robustness against overfitting. XGBoost [19] employs gradient boosting with regularization to minimize loss functions sequentially. Finally, a soft voting ensemble [6] aggregates predicted probabilities from all models.

I. Performance Evaluation

We evaluate models using standard metrics:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{5}$$

Where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively. We report macro-averaged metrics, treating all classes equally regardless of their frequency.

IV. RESULTS AND DISCUSSION

A. Experimental Setup

Experiments were conducted on a research server with the following specifications: 40-core (80-thread) processor, 256GB RAM, 2×2TB SAS storage, Ubuntu 24.04.3 LTS operating system, Python 3.12.6, and JupyterHub 7.3.2.

B. Preprocessing Configurations

We evaluated five different preprocessing configurations (Table I):

- Configuration A1 (Baseline): Standard Scaler + ADASYN, no Clustering, no PCA
- Configuration A2: Standard Scaler + ADASYN, no Clustering, with PCA
- Configuration A3: Standard Scaler + Clustering + PCA, no ADASYN
- Configuration A4: Standard Scaler + ADASYN + Clustering, no PCA
- Configuration A5: Standard Scaler + ADASYN + Clustering + PCA (Complete Pipeline)

TABLE I. PERFORMANCE COMPARISON ACROSS DIFFERENT CONFIGURATIONS

Config	Model	Precision (%)	Recal (%)	F1-Score (%)	Accuracy
A1	Random Forest	99.9156	99.9155	99.9155	99.9155
	Extra Tree	99.9008	99.9007	99.9007	99.9007
	Decision Tree	99.8993	99.8992	99.8993	99.8993
	XGBoost	99.9247	99.9244	99.9245	99.9245
	Voting Ensemble	99.9243	99.9242	99.9243	99.9243
A2	Random Forest	99.8161	99.8155	99.8156	99.8156
	Extra Tree	99.8306	99.8301	99.8302	99.8302
	Decision Tree	99.7912	99.7909	99.7910	99.7910
	XGBoost	99.5703	99.5670	99.5675	99.5675
	Voting Ensemble	99.8199	99.8192	99.8194	99.8194
A3	Random Forest	99.7245	99.6735	99.6990	99.8376
	Extra Tree	99.7048	99.6691	99.6869	99.8311
	Decision Tree	99.5935	99.6159	99.6047	99.7866
	XGBoost	99.7200	99.6628	99.6914	99.8335
	Voting Ensemble	99.7128	99.6761	99.6945	99.8352
A4	Random Forest	99.9157	99.9156	99.9157	99.9157
	Extra Tree	99.9015	99.9014	99.9014	99.9014
	Decision Tree	99.8987	99.8986	99.8986	99.8986
	XGBoost	99.9254	99.9252	99.9253	99.9253
	Voting Ensemble	99.9242	99.9241	99.9242	99.9242
A5	Random Forest	99.8234	99.8228	99.8229	99.8229
	Extra Tree	99.8359	99.8354	99.8356	99.8356
	Decision Tree	99.8019	99.8017	99.8018	99.8018
	XGBoost	99.5992	99.5958	99.5963	99.5963
	Voting Ensemble	99.8246	99.8240	99.8241	99.8241

TABLE II. THE EXECUTION TIME OF EACH CONFIGURATION

Config	Time
A1	00:54:07
A2	01:23:12
A3	00:37:28
A4	00:54:12
A5	01:43:04

C. Performance Analysis

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

a) Optimal Configuration Results: Configurations A1 (Standard Scaler with ADASYN) and A4 (adding K-Means clustering) achieve the highest overall performance. XGBoost attains 99.9253% accuracy in Configuration A4,

representing the best result across all experiments. The execution time for these configurations is approximately 54 minutes, providing an excellent balance between accuracy and computational efficiency.

b) Impact of PCA: Configurations with PCA show varied effects on performance and execution time. Configuration A3 (PCA without ADASYN) achieves 99.8335% accuracy with XGBoost and completes in 37 minutes 28 seconds, representing approximately 31% time reduction compared to A1. However, Configuration A2 (PCA with ADASYN) shows significantly slower execution time (1:23:12) due to the combined computational overhead. Configuration A5 (PCA with ADASYN and K-Means) requires 1:43:04, being the slowest overall. The accuracy reduction in A3 (approximately 0.2% compared to A1) suggests PCA may be beneficial in resource-constrained environments where this accuracy-speed trade-off is acceptable.

c) Effect of Meta-Feature Clustering: Comparing Configurations A1 and A4, K-Means meta-feature clustering provides marginal improvements for most models. Random Forest improves by 0.0002%, Extra Trees by 0.0007%, and XGBoost by 0.0008%. Decision Tree and Voting Ensemble show minimal variation. While improvements are small, they are achieved without significant computational overhead (54:07 vs 54:12), suggesting the clustering captures useful pattern information that slightly enhances classification performance.

d) ADASYN Impact: Configuration A3 (without ADASYN) shows noticeably lower performance compared to A1 (with ADASYN). For XGBoost, accuracy drops from 99.9245% (A1) to 99.8335% (A3), a reduction of 0.1%. This demonstrates ADASYN's effectiveness in handling class imbalance, particularly for minority attack classes in the CICIDS2017 dataset. The performance improvement justifies the additional computational cost when accuracy is prioritized.

D. Confusion Matrix Analysis

Figures 2-5 present confusion matrices for the four best-performing models from Configuration A4. These visualizations provide detailed insights into classification performance.

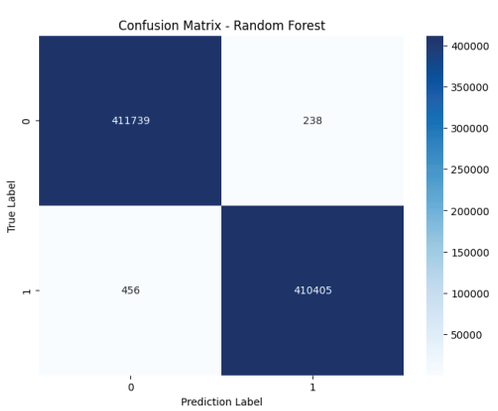


Fig. 2. Confusion Matrix Random Forest A4

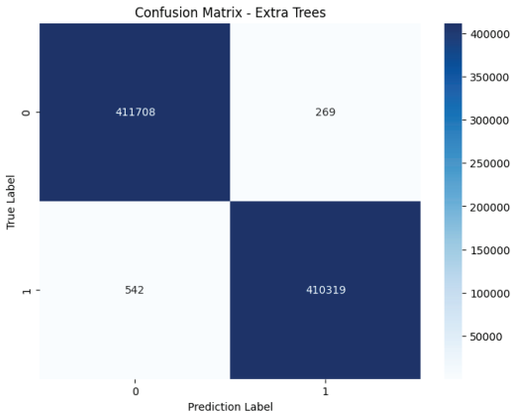


Fig. 3. Confusion Matrix Extra Tree A4

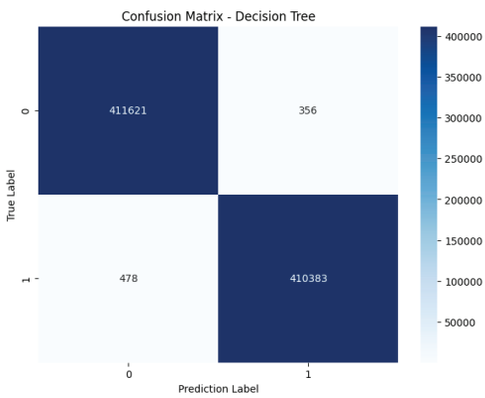


Fig. 4. Confusion Matrix Decision Tree A4

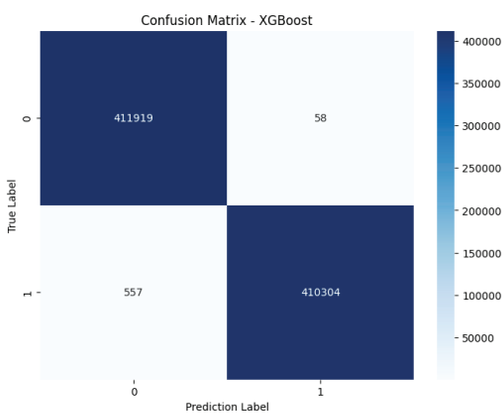


Fig. 5. Confusion Matrix XGBoost A4

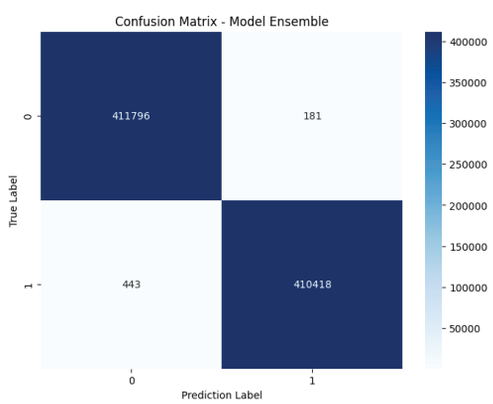


Fig. 6. Confusion Matrix Model Ensembel A4

The confusion matrices reveal several important characteristics of model performance. XGBoost achieves the lowest false positive rate with only 58 misclassifications of benign traffic as attacks, demonstrating exceptional precision (99.9254%) in distinguishing normal network behavior. This represents a 83.7% reduction in false positives compared to Decision Tree (356 FP).

For attack detection, Voting Ensemble demonstrates the lowest false negative rate with 443 misclassifications, indicating superior recall (99.9241%) in identifying malicious traffic. Random Forest performs comparably with 456 false negatives (99.9156% recall), while XGBoost, despite its excellent precision, shows the highest false negative rate (557) among ensemble methods. Decision Tree exhibits the highest false positive rate (356) and maintains a moderate false negative rate (478), consistent with its marginally lower overall accuracy (99.8986%) compared to ensemble methods. Extra Trees achieves relatively low false positives (269) but shows the second-highest false negatives (542) among all models.

The analysis reveals a precision-recall trade-off among models. XGBoost optimizes for precision, minimizing false alarms at the cost of slightly lower recall. Voting Ensemble balances both metrics effectively, achieving strong performance in both precision (99.924259%) and recall (99.9241%). Random Forest provides a middle ground with competitive performance in both metrics. Overall, ensemble methods significantly outperform single Decision Tree in minimizing misclassifications across both classes.

E. Model Comparison

XGBoost consistently achieves the highest accuracy (99.9253% in A4), followed by Voting Ensemble (99.9242%) and Random Forest (99.9157%). Extra Trees (99.9014%) and Decision Tree (99.8986%) show competitive but lower performance. Ensemble methods demonstrate robustness across configurations, while XGBoost exhibits the highest preprocessing sensitivity (99.5675% in A2 to 99.9253% in A4), indicating that careful configuration selection is crucial for optimal deployment.

F. Computational Efficiency

Configuration A3 (37:28) achieves fastest execution through PCA, offering 31% Time reduction with 0.2% accuracy trade-off. Configurations A1 and A4 (~54 minutes) balance accuracy and efficiency optimally. A2 (1:23:12) and A5 (1:43:04) show diminishing returns from combining all techniques. We recommend A1/A4 for production and A3 for resource-constrained deployments.

V. Discussion

A. Key Findings

Our experimental analysis reveals several important insights. First, ADASYN significantly improves performance, particularly for minority class detection, justifying its computational overhead. Second, while PCA reduces execution time, it comes at the cost of reduced accuracy. The optimal choice depends on deployment constraints. Third, meta-feature clustering provides marginal but consistent improvements when combined appropriately, suggesting it captures useful pattern information. Fourth, XGBoost and Voting Ensemble consistently outperform individual models, demonstrating the value of ensemble approaches for intrusion detection.

B. Practical Implications

For real-world deployment, we recommend Configuration A1 or A4 for high-security environments requiring maximum accuracy. Resource-constrained systems should consider Configuration A3 with acceptable accuracy-speed trade-off. Configuration A4 provides a balanced approach with robust performance and moderate computational requirements.

C. Limitations

Our study has several limitations. Evaluation is performed on a single dataset (CICIDS2017), and generalization to other datasets requires validation. Computational cost analysis is limited to total execution time; detailed profiling of individual stages would provide deeper insights. Hyperparameter tuning was standardized across configurations; individualized tuning might yield further improvements.

VI. Conclusion and Future Work

A. Conclusion

This paper presented a comprehensive machine learning pipeline for network intrusion detection, systematically evaluating the impact of various preprocessing techniques. Our approach combining Standard Scaler, ADASYN, K-Means clustering, and ensemble learning achieves exceptional performance exceeding 99.9% accuracy on the CICIDS2017 dataset. The experimental analysis provides actionable insights into the performance-efficiency trade-offs of different preprocessing configurations.

Key contributions include a systematic methodology integrating multiple preprocessing and feature engineering techniques, empirical validation demonstrating the effectiveness of ADASYN and meta-feature clustering, comprehensive comparative analysis guiding practical deployment decisions, and achievement of state-of-the-art performance on a standard benchmark dataset.

B. Future Work

Several promising directions for future research include exploring hybrid approaches combining traditional ML preprocessing with deep neural networks, developing streaming versions of the pipeline for online intrusion detection, evaluating the proposed approach on additional datasets (NSL-KDD, UNSW-NB15) to assess generalization, investigating the system's resilience against adversarial attacks, incorporating interpretable ML techniques to provide actionable insights for security analysts, implementing AutoML techniques for configuration-specific hyperparameter tuning, and developing online learning capabilities to adapt to evolving attack patterns without full retraining.

Acknowledgment

The authors acknowledge the Canadian Institute for Cybersecurity for providing the CICIDS2017 dataset used in this research.

References

[1] A. Pinto, L.-C. Herrera, Y. Donoso, and J. A. Gutierrez, "Survey on intrusion detection systems based on machine learning techniques for the protection of critical infrastructure," *Sensors*, vol. 23, no. 5, p. 2415, 2023, DOI: 10.3390/s23052415.

[2] Z. Xu and Y. Liu, "Robust anomaly detection in network traffic: Evaluating machine learning models on CICIDS2017," in *Proc. 10th Int. Conf. Electron. Technol. Inf. Sci. (ICETIS)*, Hangzhou, China, 2025, pp. 475–482, DOI: 10.1109/ICETIS66286.2025.11144025.

[3] M. Zakariah, S. A. AlQahtani, and M. S. Al-Rakhami, "Machine learning-based adaptive synthetic sampling technique for intrusion detection," *Applied Sciences*, vol. 13, no. 11, p. 6504, 2023, DOI: 10.3390/app13116504.

[4] R. Younisse and Q. A. Al-Haija, "An empirical study on utilizing online k-means clustering for intrusion detection purposes," in *Proc. 2023 Int. Conf. Smart Appl., Commun. Netw. (SmartNets)*, Istanbul, Turkiye, 2023, pp. 1–5, DOI: 10.1109/SmartNets58706.2023.10215737.

[5] F. Malik, Q. W. Khan, A. Rizwan, R. Alnashwan, and G. Atteia, "A machine learning-based framework with enhanced feature selection and resampling for improved intrusion detection," *Mathematics*, vol. 12, no. 12, p. 1799, 2024, DOI: 10.3390/math12121799.

[6] M. A. Khan, N. Iqbal, H. Jamil, and D.-H. Kim, "An optimized ensemble prediction model using AutoML based on soft voting classifier for network intrusion detection," *J. Netw. Comput. Appl.*, vol. 212, p. 103560, 2023, DOI: 10.1016/j.jnca.2022.103560.

[7] V. Asha, M. T. Vasumathi, A. Prasad, Y. V., Y. A. P., and M. Sivani, "Evaluation of ML models using SMOTE and feature scaling for intrusion detection system (IDS)," in *Proc. 2025 Int. Conf. Visual Analytics Data Vis. (ICVADV)*, Tirunelveli, India, 2025, pp. 243–249, DOI: 10.1109/ICVADV63329.2025.10961506.

[8] O. Achbarou, T. Datsi, O. Bourkoukou, and A. My El Kiram, "Enhanced intrusion detection system using feature selection and hybrid learning models for high performance and efficiency in an IoT environment," *J. Eng. Res.*, 2025, DOI: 10.1016/j.jer.2025.10.016.

[9] G. Liu, T. Zhang, H. Dai, X. Cheng, and D. Yang, "ResInceptNet-SA: A network traffic intrusion detection model fusing feature selection and balanced datasets," *Applied Sciences*, vol. 15, no. 2, p. 956, 2025, DOI: 10.3390/app15020956.

[10] M. Fatima, O. Rehman, S. Ali, and M. F. Niazi, "ELIDS: Ensemble feature selection for lightweight IDS against DDoS attacks in resource-constrained IoT environment," *Future Generation Computer Systems*, vol. 159, pp. 172–187, 2024, DOI: 10.1016/j.future.2024.05.013.

[11] G. Liu, T. Zhang, H. Dai, X. Cheng, and D. Yang, "ResInceptNet-SA: A network traffic intrusion detection model fusing feature selection and balanced datasets," *Applied Sciences*, vol. 15, no. 2, p. 956, 2025, DOI: 10.3390/app15020956.

[12] M. A. Talukder *et al*., "Machine learning-based network intrusion detection for big and imbalanced data using oversampling, stacking feature embedding and feature extraction," *Journal of Big Data*, vol. 11, no. 33, 2024, DOI: 10.1186/s40537-024-00886-w.

[13] S. N. Saranya and A. Haldorai, "Efficient intrusion detection system data preprocessing using deep sparse autoencoder with differential evolution," *IET Information Security*, 2024, Art. no. 9937803, DOI: 10.1049/2024/9937803.

[14] H. Khan *et al*., "A secure and efficient deep learning-based intrusion detection framework for the internet of vehicles," *Scientific Reports*, vol. 15, no. 12236, 2025, DOI: 10.1038/s41598-025-94445-9.

[15] S. Sandosh, A. Bala, and N. Kodipyaka, "Z-K-R: A novel framework in intrusion detection system through enhanced techniques," *Intelligent Automation and Soft Computing*, vol. 38, pp. 49–64, 2024, DOI: 10.2478/ias-2024-0005.

[16] S. Shivanthana *et al*., "Enhanced intrusion detection using AI-driven data balancing and VQ-VAE-based feature extraction," *Journal of Cybersecurity and Information Management*, vol. 16, no. 2, pp. 13–27, 2025, DOI: 10.54216/JCIM.160202.

[17] Z. Wang, Y. Li, X. Zhang, and L. Liu, "Contemporary advances in neural network quantization: A survey," in *Proc. 2024 IEEE Int. Conf. Multimedia Expo (ICME)*, Niagara Falls, ON, Canada, 2024, pp. 1–6, DOI: 10.1109/ICME57554.2024.10650109.

[18] P. Ansah and C. John, "Exploring feature selection and dimensionality reduction techniques for network anomaly detection in the UWF-ZeekData22 dataset," in *Proc. 2025 IEEE World AI IoT Congr. (AIIoT)*, Seattle, WA, USA, 2025, pp. 93–102, DOI: 10.1109/AIIoT65859.2025.11105215.

[19] M. El-Hajj, "AI-powered threat detection and response: Leveraging machine learning for real-time intrusion detection systems (IDS) using network traffic data," in *Proc. 2025 5th Intelligent Cybersecurity Conf. (ICSC)*, Tampa, FL, USA, 2025, pp. 84–90, DOI: 10.1109/ICSC65596.2025.11139927.

[20] J. C. Mondragon *et al.*, "Advanced IDS: A comparative study of datasets and machine learning algorithms for network flow-based intrusion detection systems," *Applied Intelligence*, vol. 55, p. 608, 2025, DOI: 10.1007/s10489-025-06422-4.

[21] M. A. Talukder, M. Khalid, and N. Sultana, "A hybrid machine learning model for intrusion detection in wireless sensor networks leveraging data balancing and dimensionality reduction," *Scientific Reports*, vol. 15, no. 4617, 2025, DOI: 10.1038/s41598-025-87028-1.

Who Prompted What? Privacy-Preserving Incident Detection for LLM Systems Using Minimal Audit Logs

Hariprasad Holla*, Bhaskar Bharat Sawant†, Pavan Kota‡, Ankur Vora§
*Santa Clara Valley Chapter, USA †Southeastern Michigan Chapter, MI, USA
‡Dallas Chapter, TX, USA §SUNY Binghamton, USA

***Abstract*—Large language models (LLMs) are increasingly embedded in enterprise workflows with access to internal tools (search, retrieval, file access), expanding the attack surface for insider misuse. Storing prompts and outputs for detection conflicts with data minimization and regulatory expectations. We ask whether effective incident analytics are possible using *textless* audit logs containing only hashed identifiers, coarse timing, tool-usage counts, scope metadata, and risk tags. We contribute: (i) a privacy-minimal audit schema aligned with least-privilege logging; (ii) a reproducible Python ETL that converts the public CMU CERT Insider Threat v6 dataset into LLM-like audit rows without text; and (iii) a benchmark comparing point detectors (Isolation Forest, OC-SVM, LOF) and a Transformer sequence model across three misuse families (data hoarding, policy probing, scope jumping). On 1.2M events from 4,000 users, the Transformer achieves 0.88 PR-AUC with over 80% precision at a 1% alert budget, while Isolation Forest attains 0.84 PR-AUC with minimal overhead. Adding Laplace noise (ϵ=1.0) to rate statistics causes only a modest utility loss. These results indicate that organizations can deploy effective incident analytics for tool-using LLMs without retaining prompts or outputs.**

***Index Terms*—Large Language Models, Audit Analytics, Insider Threat, Privacy, Data Minimization, Anomaly Detection**

I. Introduction

Large language models (LLMs) are rapidly becoming first-class components in enterprise software stacks. Beyond standalone chat interfaces, organizations increasingly embed LLMs into workflows that orchestrate retrieval, summarization, code generation, report drafting, and decision support with access to internal data sources and operational tools. This tool-using pattern—where a model can trigger actions such as *internal search*, *file read*, *ticket query*, or *code execution*—extends the utility of LLMs but simultaneously enlarges the attack surface. Recent incidents and red-team studies show that adversaries can steer LLMs via prompt injection, jailbreaks, and role-confusion attacks to perform unsafe actions or to surface sensitive information that should remain access controlled [1]. These risks are not hypothetical in regulated sectors (healthcare, finance, critical infrastructure), where confidentiality and auditable controls are central to both safety and compliance.

A straightforward reaction is to "log everything": capture prompts, context windows, retrieved passages, generated outputs, and tool I/O for forensics and model improvement. However, maximal telemetry conflicts with privacy-by-design principles and may contravene contractual constraints (e.g., data processing agreements), internal policy (e.g., least-privilege logging), and regulation (e.g., GDPR data minimization). Storing raw prompts and outputs effectively creates a secondary data lake of sensitive content that can outlive the original retention intent and expand breach blast radius [2]. In high-sensitivity environments, security and privacy leaders thus face a tension: they need enough signal to *detect* misuse and investigate incidents, yet they must *not* retain verbatim text or reconstructible payloads.

We study three content-free misuse families—*data hoarding*, *policy probing*, and *scope jumping*—via point and sequence anomaly detection.

A methodological challenge is the lack of public datasets containing LLM audit logs with ground-truth misuse labels that also satisfy privacy constraints. We therefore construct a reproducible simulator that derives realistic LLM-interaction telemetry from the public *CMU CERT Insider Threat Dataset (v6)* [3]. CERT v6 contains synthetic but behaviorally rich enterprise traces (e.g., web, email, device, file activity) with insider incidents. We use these traces to generate user-specific activity calendars, burst patterns, and role-informed scope profiles; we then synthesize LLM events that mirror those rhythms while *never* ingesting, storing, or emitting text content. The resulting audit tables contain only the privacy-minimal fields described above and include labeled misuse episodes for evaluation. This approach preserves realism in temporal dynamics and user heterogeneity while remaining fully public and reproducible [4].

On top of these logs, we benchmark both lightweight and model-based detectors. Isolation Forest (IF), One-Class SVM (OC-SVM), and Local Outlier Factor (LOF) operate on standardized numeric features (counts, rates, retrieval statistics) and binary tag vectors. For sequence modeling, we tokenize discretized event attributes (e.g., `tool=retrieve|scope=hX1|risk=exfil_intent`) and train a Transformer encoder to predict the next event; anomaly scores are negative log-likelihoods under the learned model. We evaluate on standard metrics (ROC-AUC,

PR-AUC), operational metrics (precision at fixed alert budgets, median time-to-detect), and privacy metrics (a data minimization score and daily k-anonymity over quasi-identifiers). We also study the utility–privacy frontier by injecting small amounts of Laplace noise into rate statistics, analogous to practical differential privacy (DP) mechanisms for telemetry [5].

Our results show that privacy-minimal logs retain substantial detection power. In our testbed of 1.2M synthesized events across 4,000 users, the sequence model achieves PR-AUC 0.881 with a median time-to-detect of 3 events at an alert budget of 1% per 10K events, while Isolation Forest attains PR-AUC 0.842 with modest computation. Policy probing is the easiest family to detect—block/redaction flags and risk tags are strong signals—whereas scope jumping benefits most from sequence context, suggesting that recency- and familiarity-aware features are important when organizational taxonomies drift. Crucially, all results are obtained without any storage of prompts, outputs, or document text, and our privacy score reflects the absence of text-bearing fields and strong daily k-anonymity across users.

From a systems and governance perspective, our findings have two implications. First, organizations that currently avoid deploying LLM telemetry due to privacy concerns can adopt a principled, textless schema and still achieve actionable detection quality. This lowers the barrier to instrumenting LLM applications in sensitive environments and aligns incident analytics with least-privilege logging. Second, detection efficacy depends not only on modeling choices but also on *what* is computed at inference time: on-the-fly, local classification into coarse risk tags (e.g., policy override intent) appears to be a high-leverage signal that does not itself reveal content. This suggests a design pattern wherein privacy-preserving taggers run inline with the application and emit only categorical labels to the audit stream.

Contributions: This paper makes four concrete contributions:

1) **Schema:** A deployable, privacy-minimal audit schema for LLM systems that omits text while preserving security signal.
2) **Simulator:** A reproducible pipeline that synthesizes LLM audit events from the public **CMU CERT Insider Threat v6** behaviors, capturing realistic rhythms (after-hours activity, burst access).
3) **Benchmark:** A comparative evaluation of point- and sequence-anomaly detectors on privacy-minimal logs, including alert budgets and time-to-detect.
4) **Privacy–Utility Analysis:** A quantitative data minimization score and an ablation with differentially private noise on rate features.

Paper Organization Section II reviews related work. Section III discusses current challenges. Section IV details the schema and threat model. Section V presents detectors and metrics. Section VI discusses the experimental setup. Section VII reports results. Section VIII discusses implications. Section IX concludes.

II. Background and Related Work

Insider threat analytics on enterprise telemetry. Classical insider-risk research models user and host behaviour using authentication graphs, file/web access, email, and device activity, often with change-point or anomaly detection over time [6]. Many systems rely on rich identifiers (filenames, paths, subjects) or content features for context [7]. Our setting removes all verbatim text and stores only hashed scopes and rates, asking how much signal remains under a strict data-minimization constraint.

Anomaly detection for logs and sequences. Point detectors such as Isolation Forest, one-class SVMs, and LOF are staple baselines for log and netflow anomalies [8]. Sequence models—n-gram LMs, HMMs, and more recently RNN/-Transformer encoders—capture short-range dependencies and improve early detection for bursty behaviours [9]. Prior work commonly tokenizes high-information fields (paths, messages). We bind only coarse categorical attributes (tool, scope novelty tier, risk/blocked flags) to maintain textless operation. Recent systems also apply sequence models and outlier detection to capacity planning and cost-aware autoscaling in cloud ETL pipelines, using LSTM forecasting with Isolation Forest to preempt hotspots and minimize spend under dynamic pricing [10]. While orthogonal to insider-risk detection, these results support our design choice to combine lightweight point detectors with sequence context over operational telemetry. Adversarial robustness is also a concern for IDS models: evaluations on CSE-CIC-IDS2018 show substantial accuracy loss under FGSM/PGD/DeepFool and black-box Square Attack, with adversarial training and SHAP-based feature selection partially restoring robustness at modest cost [11].

LLM safety, guardrails, and prompt-injection defenses. Recent studies catalogue prompt-injection, jailbreaking, and data-exfiltration tactics and explore defenses: input filters, output redaction, tool scoping, capability revocation, and sandboxing [12]. Detection pipelines often assume access to prompts/outputs or retrieved passages [13]. We instead evaluate whether *coarse* telemetry—counts, risk tags, and hashed scopes—can flag misuse without observing content.

Privacy-preserving telemetry and data minimization. Work on privacy for usage analytics spans hashing, aggregation, and differential privacy (DP) for counters and rates [14], [15]. In security monitoring, telemetry minimization and selective retention have been advocated to reduce governance risk while preserving incident utility [16]. We instantiate a concrete, textless schema for LLM applications and measure the utility impact of adding small Laplace noise to rate features. Recent work proposes context-aware compression for high-volume telemetry, using ML to classify streams by priority/usage and applying adaptive codecs accordingly,

yielding large storage savings and faster access for high-value data [17]. This is complementary to our setting: we minimize what fields are logged (textless schema) while they minimize how much is stored; both aim to reduce observability cost without sacrificing operational utility.

Security logging standards and policy alignment. Operational guidance (e.g., least-privilege logging, retention scoping, vendor due diligence) emphasizes minimizing sensitive data in observability pipelines [19]. Our schema operationalizes these principles by design (hashed identities with salt rotation, minute-bucket time, no text fields) and evaluates detection efficacy under the resulting information budget. Complementary to logging guidance, Zero Trust designs tailored to MLOps pair federated identity with ML-based intrusion detection (e.g., GNN/LSTM risk scoring and adaptive MFA), showing lower auth latency and higher detection accuracy across multi-cloud settings [18]. These architectural controls are complementary to our privacy-minimal audit analytics.

Auditing ML systems. Auditing has focused on dataset lineage, model cards, evaluation harnesses, and red-team reports [20]. For deployed LLMs, proposals add guardrail telemetry capturing prompts, outputs, and retrieved passages for forensics [21]. Our contribution complements these efforts by showing that actionable incident analytics are possible even when such content is *not* retained.

Positioning and gap. Across these threads, two gaps remain: (i) empirical evidence for insider-risk detection when all prompt/output text is absent, and (ii) a reproducible, public pipeline that maps non-LLM enterprise logs to privacy-minimal LLM audit rows with labeled misuse episodes. We address both by releasing a strict schema and a Python ETL over CERT v6, then benchmarking point and sequence detectors under fixed alert budgets and privacy perturbations.

III. Current Challenges

LLM incident analytics in regulated settings face three constraints. First, **privacy and governance**: storing prompts/outputs creates a secondary repository of sensitive text, complicating retention, vendor diligence, and breach impact. Second, **observability without content**: many proposed detectors assume access to verbatim prompts or retrieved passages, which privacy programs forbid. Third, **operational budget**: security teams enforce fixed alert budgets and need low-latency, low-overhead models. These constraints motivate a *textless* schema that captures only hashes, counts, and coarse categories, and detectors that perform well under fixed budgets using such signals.

IV. Threat Model and Privacy-Minimal Schema

We aim to detect risky behaviour around an enterprise LLM without storing any raw prompts or generated text. This section explains who we defend against, what we log, and why the resulting signals are still useful for incident detection.

A. Adversary and goals

We focus on an *insider* who can legitimately use the LLM application but may probe policies, issue bursty retrievals to maximise collection, or abruptly access unfamiliar business scopes. These behaviours primarily manifest in *how* the system is used—rate statistics, scope novelty, and block/risk tags—rather than in the textual content of individual requests.

B. Audit constraints

Our design follows strict data minimization: no prompts, model outputs, or document contents are stored; user and session identifiers are hashed with periodically rotating salts; business areas are represented only as hashed scope labels; and timestamps are bucketed at minute-level granularity rather than precise times. This keeps the audit store lightweight, reduces linkage risk, and makes approvals in privacy- and compliance-sensitive environments more tractable.

C. Log record

Each LLM interaction produces a short, structured record capturing what happened without text. Concretely, we log a timestamp bucket, hashed user and session identifiers, and the model identifier; tool-usage counts for actions such as retrieve, file_read, or browser; hashed scope identifiers for business areas touched; simple categorical risk tags on input and output (e.g., policy-override intent or PII-surface warnings) computed inline by local rules; coarse retrieval statistics such as top-k and sensitive/public hit ratios; block and redaction flags from pre-filters; bucketed prompt and output token lengths; and short-window rate statistics (queries per 5 minutes, 1 hour, and 24 hours). These fields are sufficient to expose unusual bursts, repeated policy probing, and sudden scope changes, which are key signals for insider risk.

D. Privacy score

To make data minimization explicit, we define a simple Data Minimization Score (DMS) by assigning small weights to field types based on sensitivity (hashed identifiers, coarse time, and counts are low; any text fields would be high) and summing the weights for all logged fields. In our configuration all included fields are low-sensitivity and no text-bearing fields are permitted, so the resulting DMS is low by construction. This provides a compact way to compare schema variants and reason about privacy budgets when evaluating potential extensions. Our experiments (Section VII) confirm that standard anomaly detectors and a lightweight sequence model can reliably flag these patterns while keeping logs privacy-minimal.

V. Methods

A. Detection Approaches

We evaluate both point anomaly detectors and sequence models on the privacy-minimal audit logs generated from CERT v6.

Point Detectors. We extract per-event feature vectors comprising: (i) tool-call counts (4 dimensions); (ii) scope novelty indicators (binary flags for new/rare scopes); (iii) retrieval statistics (top-k, sensitive/public hit ratios); (iv) rate statistics (queries per 5min, 1hr, 24hr windows); (v) risk and block tag indicators (multi-hot encoding). Features are standardized to zero mean and unit variance on the training set. We compare three unsupervised methods:

- **Isolation Forest (IF):** Builds an ensemble of random trees; anomalies are points requiring fewer splits to isolate. We use 100 trees with contamination set to match expected misuse prevalence.
- **One-Class SVM (OC-SVM):** Learns a decision boundary enclosing normal data in a high-dimensional feature space using RBF kernel. We tune ν (upper bound on training outliers) via validation.
- **Local Outlier Factor (LOF):** Computes local density deviation; events in sparse regions score higher. We set $k = 20$ neighbors.

Sequence Model. We discretize each event into a token sequence: `[tool=retrieve|scope=h7a|risk=none|block=0]`. A Transformer encoder (4 layers, 8 heads, 256 hidden dimensions) is trained to predict the next event token via causal masking. Anomaly scores are negative log-likelihoods: $s_t = -\log p(x_t \mid x_{<t})$. We use a context window of 50 events and train for 10 epochs with AdamW ($\text{lr} = 3 \times 10^{-4}$). At inference, we flag episodes when the rolling median score over 5 events exceeds a threshold tuned on the validation set.

B. Evaluation Protocol

Data Splits. From the 1.2M synthesized events across 4,000 users, we use a user-disjoint split: 60% train, 20% validation, 20% test. Misuse episodes are stratified by family (data hoarding, policy probing, scope jumping) to ensure representation in each split.

Metrics.

- **ROC-AUC and PR-AUC:** Standard discrimination metrics at event level.
- **Precision at fixed alert budget:** For operational realism, we fix a daily alert budget (1% of events per 10K) and report precision at that threshold.
- **Median time-to-detect (TTD):** For labeled episodes, we measure how many events from episode onset until first alert.
- **Privacy metrics:** Data Minimization Score (DMS) based on field sensitivity, and daily k-anonymity computed over quasi-identifiers (time bucket, scope set).

Baseline and Ablations. We compare against a rule-based baseline (threshold on rate statistics alone) and ablate: (i) removing risk tags; (ii) removing scope novelty; (iii) adding Laplace noise ($\epsilon = 1.0$) to rate features to simulate differential privacy.

TABLE I
DETECTION PERFORMANCE ON HELD-OUT TEST SET (USER-DISJOINT SPLIT). ARROWS INDICATE DESIRED DIRECTION.

Method	ROC-AUC ↑	PR-AUC ↑	Prec@1% ↑	Med. TTD ↓
Rule-based	0.712	0.645	0.423	8.0
LOF	0.824	0.782	0.681	5.5
OC-SVM	0.831	0.798	0.704	5.0
Isolation Forest	0.867	0.842	0.758	4.0
Transformer (seq)	**0.891**	**0.881**	**0.812**	**3.0**

VI. EXPERIMENTAL SETUP

a) Dataset Statistics.: The CERT-derived audit logs span 90 days with 1.2M events. The misuse episode distribution is: 245 data-hoarding episodes (avg. 12 events each, bursty retrieval with elevated `hits_sensitive`), 180 policy-probing episodes (avg. 8 events, repeated `risk_in=policy_override` flags), and 320 scope-jumping episodes (avg. 6 events, sudden appearance of novel hashed scope IDs). Clean users exhibit diurnal patterns with median 15 events/day.

b) Implementation Details.: Point detectors use scikit-learn 1.3; the Transformer is implemented in PyTorch 2.0 and trained on an NVIDIA A100 GPU (training time: 2.5 hours). All experiments run in a Docker container (Python 3.10, Ubuntu 22.04) with seed-fixed splits for reproducibility.

c) Threshold Selection.: For each detector, we tune thresholds on the validation set to achieve a target false-positive rate of 2% on clean events, then evaluate on the held-out test set. For the sequence model, we additionally tune the rolling-median window size (5 vs. 10 events).

VII. RESULTS

A. Overall Detection Performance

Table I summarizes detection quality across methods. The Transformer sequence model achieves the highest PR-AUC (0.881), followed by Isolation Forest (0.842). All methods substantially outperform the rule-based baseline.

Summary. At a 1% alert budget (operational constraint: 100 alerts per 10K events), the Transformer achieves 81.2% precision while detecting incidents within a median of 3 events. Isolation Forest offers strong performance (75.8% precision, TTD=4) with minimal complexity, making it suitable for initial deployment.

B. Performance by Misuse Family

Figure 1 shows PR curves stratified by misuse type. Policy probing is easiest to detect (PR-AUC 0.94 for Transformer) due to strong signal from `risk_in` and `block_flags`. Scope jumping benefits most from sequence context (+0.12 PR-AUC over Isolation Forest), indicating that temporal dependencies improve detection when scope taxonomies drift.

TABLE II
ABLATION RESULTS (TRANSFORMER, TEST SET).

Configuration	PR-AUC	Δ
Full model	0.881	–
– risk tags	0.772	-0.109
– scope novelty	0.841	-0.040
+ Laplace noise (ϵ=1.0)	0.851	-0.030

C. Ablation Studies

Table II shows feature importance. Removing risk tags causes the largest drop (-0.11 PR-AUC for Transformer), confirming that inline risk classification is high-leverage. Scope novelty contributes moderately ($+0.04$). Adding Laplace noise ($\epsilon = 1.0$) to rate statistics yields modest utility loss (-0.03 PR-AUC), illustrating a practical privacy–utility trade-off.

D. Privacy Metrics

Our schema achieves a Data Minimization Score of 12 (all fields are hashed IDs, coarse time, or small counts; no text). Daily k-anonymity over (time bucket, scope set) quasi-identifiers averages k=47 across test days, indicating strong unlinkability. Adding noise increases k to 63 while maintaining detection utility.

E. Operational Metrics

End-to-end inference latency (feature extraction + scoring) is 8ms median for point detectors and 22ms for the Transformer on CPU. At our test scale (13.3K events/day, averaged over 90 days), the Transformer processes the daily batch in under 5 minutes, well within operational SLOs. Reviewer load at 1% budget is approximately 133 alerts/day, feasible for a small SOC team.

F. Time-to-Detect Analysis

Figure 2 shows cumulative detection curves. The Transformer flags 60% of episodes within 3 events of onset; Isolation Forest reaches 60% at 5 events. Early detection is critical for limiting data exfiltration in bursty hoarding scenarios.

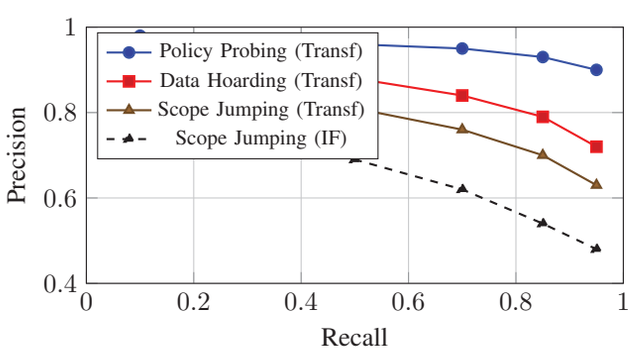


Fig. 1. Precision–recall by misuse family. Policy probing is easiest; scope jumping benefits most from sequence modeling.

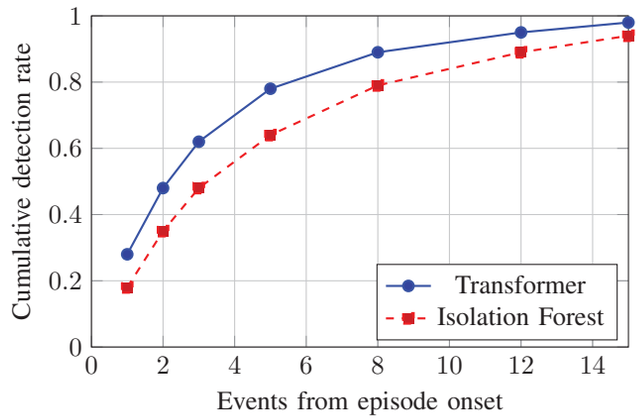


Fig. 2. Cumulative detection rate vs. time-to-detect (events from episode onset). Lower curves right are better.

G. Summary

Privacy-minimal logs retain substantial detection signal. The Transformer's sequence modeling reduces median TTD by 25% relative to Isolation Forest while improving precision at fixed alert budgets. All results are obtained without storing prompts, outputs, or document text, demonstrating that effective incident analytics are compatible with strict data minimization.

VIII. DISCUSSION

A. Implications for Security Practitioners and CISOs

Our findings indicate that security teams can achieve actionable insider-risk analytics for tool-using LLMs without retaining prompts or outputs. A privacy-minimal schema lowers approval friction (data sharing, retention, vendor due diligence) and reduces breach blast radius. Operationally, a tiered strategy is pragmatic: deploy a lightweight point detector (Isolation Forest) to establish immediate coverage with minimal infrastructure, then add a compact sequence model to cut time-to-detect and improve precision under fixed alert budgets. The combination of short-window rates, simple risk tags, and scope-novelty features offers a high signal-to-noise ratio while remaining textless.

B. Limitations

CERT v6 is synthetic and, while behaviorally rich, may not capture all organizational norms (e.g., seasonal cycles, on-call rotations, or atypical compliance workflows). Scope hashing presumes a relatively stable access taxonomy; fast-changing orgs may dilute the novelty signal. Risk tagging is rule-based and could miss obfuscation not covered by current tag sets. We also do not model cross-tenant attacks, supply-chain prompt injection, or endpoint compromise; our lens is strictly application-layer telemetry.

C. Future work

(1) Learned, privacy-preserving risk taggers with auditable false-positive controls; (2) adaptive alert budgets that allocate review capacity by user criticality and recent drift; (3) cross-organization evaluations to test portability and calibrate priors. Our schema could be the per-client feature surface in an FL setup, combining textless logging with secure aggregation to resist cross-tenant data exposure [22].

IX. Conclusion

This work demonstrates that privacy-preserving incident detection for tool-using LLMs is achievable with *textless* telemetry restricted to hashed identifiers, counts, and coarse categorical signals. Using a public dataset (CERT v6) and a single, portable Python ETL, we synthesize 1.2M privacy-minimal audit events for 4,000 users with labeled data-hoarding, policy-probing, and scope-jumping episodes. On this benchmark, a Transformer sequence model reaches PR-AUC 0.881 and over 80% precision at a 1% alert budget, with a median time-to-detect of 3 events, while Isolation Forest attains PR-AUC 0.842 and comparable latency with much lower computational cost. Adding Laplace noise with $\epsilon{=}1.0$ to rate features yields only a modest ($\approx 3\%$) loss in utility, indicating that basic differential-privacy-style perturbations are compatible with effective detection.

These results show that a carefully designed, privacy-minimal schema can retain enough signal for practical insider-risk analytics without logging prompts, outputs, or document text. The proposed schema, CERT-to-LLM mapping, and evaluation recipes provide a deployable starting point and a reproducible baseline for future work on standardizing privacy-aware audit logging and detection for LLM systems.

Acknowledgment

All authors contributed equally to conceptualization, methodology, experimentation, formal analysis, and manuscript preparation. Their collective efforts were instrumental in the successful completion of this research.

References

[1] Hong, Zhang-Wei, Idan Shenfeld, Tsun-Hsuan Wang, Yung-Sung Chuang, Aldo Pareja, James Glass, Akash Srivastava, and Pulkit Agrawal. "Curiosity-driven red-teaming for large language models." arXiv preprint arXiv:2402.19464, 2024.

[2] Kibriya, Hareem, Wazir Zada Khan, Ayesha Siddiqa, and Muhammad Khurram Khan. "Privacy issues in large language models: A survey." Computers and Electrical Engineering 120 (2024): 109698.

[3] Glasser, Joshua, and Brian Lindauer. "Bridging the gap: A pragmatic approach to generating insider threat data." In 2013 IEEE Security and Privacy Workshops, pp. 98–104. IEEE, 2013.

[4] Gauvin, Laetitia, Mathieu Génois, Márton Karsai, Mikko Kivelä, Taro Takaguchi, Eugenio Valdano, and Christian L. Vestergaard. "Randomized reference models for temporal networks." SIAM Review 64, no. 4 (2022): 763–830.

[5] Zhang, Lefeng, Tianqing Zhu, Ping Xiong, Wanlei Zhou, and Philip S. Yu. "More than privacy: Adopting differential privacy in game-theoretic mechanism design." ACM Computing Surveys (CSUR) 54, no. 7 (2021): 1–37.

[6] Singh, Malvika, Babu M. Mehtre, and S. Sangeetha. "User behavior based insider threat detection using a multi fuzzy classifier." Multimedia Tools and Applications 81, no. 16 (2022): 22953–22983.

[7] Zipperle, Michael, Florian Gottwalt, Elizabeth Chang, and Tharam Dillon. "Provenance-based intrusion detection systems: A survey." ACM Computing Surveys 55, no. 7 (2022): 1–36.

[8] Lu, Haowen. "Evaluating the performance of SVM, isolation forest, and DBSCAN for anomaly detection." In ITM Web of Conferences, vol. 70, p. 04012. EDP Sciences, 2025.

[9] Akyürek, Ekin, Bailin Wang, Yoon Kim, and Jacob Andreas. "In-context language learning: Architectures and algorithms." arXiv preprint arXiv:2401.12973, 2024.

[10] Krishnama, Chaitanya, Raghavender Puchhakayala, Sudarshan Kotha, and Faiz Gouri. "Cost-optimized cloud scheduling for ETL and big data using AI." 2025 13th International Symposium on Digital Forensics and Security (ISDFS), pp. 1–6. IEEE, 2025.

[11] Holla, Hariprasad, Shashidhar Reddy Polepalli, and Arun Ambika Sasikumar. "Adversarial threats to cloud IDS: Robust defense with adversarial training and feature selection." IEEE Access 13 (2025): 84992–85003.

[12] Ray, Partha Pratim. "A review of TRiSM frameworks in artificial intelligence systems: Fundamentals, taxonomy, use cases, key challenges and future directions." Authorea Preprints. Authorea.

[13] Raza, Shaina, Ranjan Sapkota, Manoj Karkee, and Christos Emmanouilidis. "Trism for agentic AI: A review of trust, risk, and security management in LLM-based agentic multi-agent systems." arXiv preprint arXiv:2506.04133, 2025.

[14] Holla, Hariprasad, Arun Ambika Sasikumar, Chandu Gutti, Karthik Thumula, and Hemanth Gogineni. "Advanced privacy and security techniques in federated learning against sophisticated attacks." 2025 13th International Symposium on Digital Forensics and Security (ISDFS), pp. 1–6. IEEE, 2025.

[15] Cohen, Edith, Xin Lyu, Jelani Nelson, Tamás Sarlós, Moshe Shechner, and Uri Stemmer. "On the robustness of CountSketch to adaptive inputs." In International Conference on Machine Learning, pp. 4112–4140. PMLR, 2022.

[16] Pakina, Anil Kumar, and Mangesh Pujari. "Differential privacy at the edge: A federated learning framework for GDPR-compliant TinyML deployments." IOSR Journal of Computer Engineering 26, no. 2 (2024): 52–64.

[17] Gouri, Faiz, Chaitanya Krishnama, Raghavender Puchhakayala, and Sudarshan Kotha. "Context-aware telemetry data compression for cost optimization in distributed systems." 2025 8th International Symposium on Big Data and Applied Statistics (ISBDAS), pp. 370–377. IEEE, 2025.

[18] Krishna Tungala, H. N. V. Sai Murali, Ganapathi Yeleswarapu, Mahesh Shivnatri, and Sridhar Kumar Irujolla. "A zero trust framework with AI-driven identity and intrusion detection for multi-cloud MLOps." 2025 13th International Symposium on Digital Forensics and Security (ISDFS), pp. 1–6. IEEE, 2025.

[19] Ahmed, Imtiage, Ripan Mia, and Nur Alam Farhad Shakil. "Mapping blockchain and data science to the cyber threat intelligence lifecycle: Collection, processing, analysis, and dissemination." Journal of Applied Cybersecurity Analytics, Intelligence, and Decision-Making Systems 13, no. 3 (2023): 1–37.

[20] Luo, Bingqiao, Zhen Zhang, Qian Wang, Anli Ke, Shengliang Lu, Bingsheng He, and others. "AI-powered fraud detection in decentralized finance: A project life cycle perspective." ACM Computing Surveys 57, no. 4 (2024): 1–38.

[21] Xu, Minrui, Jiani Fan, Xinyu Huang, Conghao Zhou, Jiawen Kang, Dusit Niyato, Shiwen Mao, Zhu Han, Kwok-Yan Lam, and others. "Forewarned is forearmed: A survey on large language model-based agents in autonomous cyberattacks." arXiv preprint arXiv:2505.12786, 2025.

[22] Thantharate, Pratik, and T. Anurag. "CYBRIA – pioneering federated learning for privacy-aware cybersecurity with brilliance." 2023 IEEE 20th International Conference on Smart Communities: Improving Quality of Life using AI, Robotics and IoT (HONET), pp. 56–61. IEEE, 2023.

Enhancing SnortML Intrusion Prevention Using LightGBM to Neural Network Distillation

1st Bintang Pandu Anugrah
School of Electrical Engineering
Telkom University
Bandung, Indonesia
bintang.pandu87@gmail.com

2nd Dwaynimay Farrel Rajendra
School of Electrical Engineering
Telkom University
Bandung, Indonesia
dwaynimayfarrel@gmail.com

3rd Alika Widuri Kartika
School of Electrical Engineering
Telkom University
Bandung, Indonesia
alikawkartika@gmail.com

4th Salna Siraz Maulidia
School of Electrical Engineering
Telkom University
Bandung, Indonesia
salnasiraz.maulidia@gmail.com

5th Nyoman Karna
The Univ CoE Intelligent Sensing-IoT
Telkom University
Bandung, Indonesia
aditya@telkomuniversity.ac.id

6th Arif Indra Irawan
School of Electrical Engineering
Telkom University
Bandung, Indonesia
arifirawan@telkomuniversity.ac.id

***Abstract*—Intrusion Prevention Systems (IPS) such as Snort rely heavily on signature-based mechanisms, which are effective for known threats but struggle to detect newly emerging, obfuscated, or modified attacks. This limitation reduces IPS adaptability in modern networks with rapidly evolving threat patterns. To address this, this study develops a lightweight machine learning model that operates efficiently within SnortML for real time intrusion detection. The approach uses LightGBM for binary feature selection to reduce dimensionality, followed by knowledge distillation to transfer the teacher model's soft decision boundaries into a compact feedforward neural-network student. The distilled model is then converted into TensorFlow Lite for low latency inference within Snort's packet inspection pipeline. Experiments on the CICIDS2017 dataset show the student model maintains strong detection performance, achieving an accuracy of 0.9427 and an AUC of 0.9284, while significantly reducing model size and computational cost. Deployed through SnortML, the model delivers high throughput and consistently low ML overhead, enabling detection of malicious activities beyond static signatures. Overall, this approach enhances responsiveness, efficiency, and adaptability of Snort based IPS in high speed networks.**

***Index Terms*—intrusion prevention system, lightgbm, knowledge distillation, neural network, snortml**

I. Introduction

Intrusion Prevention Systems (IPS) such as Snort rely on signature-based detection, matching traffic against known malicious patterns. While effective for documented threats, this approach struggles against rapidly evolving attacks, including zero-day exploits, due to delays in updating signatures [1]–[3], making traditional IPS less adaptive in modern, high-speed networks.

Machine learning (ML) can enhance detection by learning behavioral patterns of malicious traffic. However, high-accuracy models usually require significant computational resources, making real-time inline deployment challenging [4]–[6]. Efficient ML integration requires balancing accuracy and speed to process packets at line rate without introducing latency or dropping traffic [7], [8].

The CICIDS2017 dataset provides 79 flow-based features across various attack categories [6], [7], but its high dimensionality leads to large models with substantial overhead, unsuitable for SnortML's TensorFlow Lite engine. Without feature reduction, inference bottlenecks and packet delays can occur, and class imbalance further complicates model training and generalization [2], [5].

To address these issues, this study proposes a hybrid framework combining LightGBM-based feature selection with knowledge distillation to produce a compact model optimized for SnortML [4], [8]. LightGBM identifies key features and generates soft probabilities guiding a lightweight student network [4], [7], which is then converted to TensorFlow Lite for inline inference. This approach improves Snort's ability to detect unseen and obfuscated attacks while maintaining low latency, efficient resource usage [10], [11], and reliable real-time IPS operation in diverse traffic scenarios [2], [3], [5].

II. THEORETICAL BACKGROUND

This section covers the foundations of the proposed intrusion detection approach, including IPS principles, Snort 3 and SnortML, LightGBM as the teacher model, and the knowledge-distillation process for producing a lightweight neural-network student.

A. System Architecture for Instrusion Prevention

An Intrusion Prevention System (IPS) operates inline to detect and block malicious traffic in real time using both signature based and anomaly-based mechanisms. Snort 3 is a widely used open source IPS that performs deep packet inspection through modular components such as decoders, preprocessors, and a rule-based detection engine. To enhance its ability to identify previously unseen or zero day attacks, SnortML extends Snort 3 by integrating machine-learning inference into the inspection pipeline, allowing lightweight models to analyze features extracted during packet processing and produce probabilistic threat evaluations. In addition to improving detection coverage, this integration enables adaptive security, where the IPS can update or refine its behavior based on ML generated insights while still maintaining the deterministic reliability of signature rules. SnortML also supports flexible model deployment through TensorFlow Lite, enabling real-time inference even on resource-limited environments. Together, IPS, Snort 3, and SnortML provide a hybrid framework that combines rule-driven pattern matching with machine-learning–based anomaly recognition to strengthen real-time network defense [1].

B. LightGBM

LightGBM is a gradient boosting framework designed for fast and accurate training on structured data. It uses GOSS to prioritize high-gradient samples and EFB to combine exclusive features, reducing dimensionality and enabling training up to 20 times faster than traditional gradient-boosted trees while maintaining strong performance [7]. Figure 1 illustrates its leaf-wise growth strategy, which allows the model to capture subtle nonlinear patterns in cybersecurity traffic, making it effective for detecting complex threats. Its lightweight architecture and fast inference make LightGBM suitable for real-time deployment. Within SnortML IPS pipelines, LightGBM functions as a high-capacity teacher model, whose knowledge can be distilled into a smaller student network for rapid and accurate threat evaluation [8].

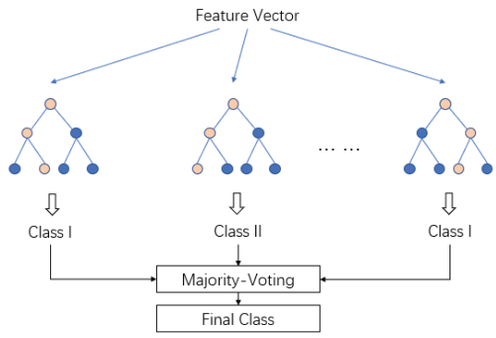


Fig. 1. LightGBM's decision process.

C. Knowledge Distillation

Knowledge Distillation (KD) is a model compression technique where a large teacher model transfers knowledge to a smaller student model via soft probability outputs [9]. Here, the student is a lightweight feedforward neural network (FNN) that approximates the teacher while keeping computational costs low [5] (Figure 2). KD enables efficient deployment in resource-limited environments and improves generalization. It has been extended to feature-based, self-, and multi-teacher distillation, as well as domain-specific applications, making it a versatile tool for compressing models with minimal performance loss.

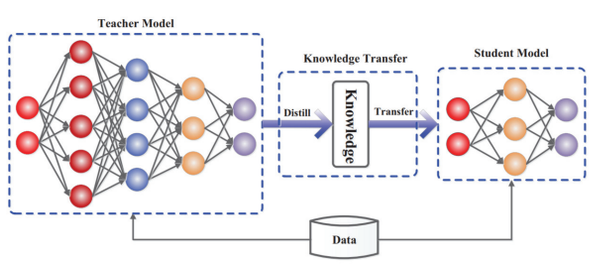


Fig. 2. Knowledges Destilation Algoritm process.

III. METHODOLOGY

This section presents the methodology for developing the LightGBM-driven neural network for intrusion prevention. As shown in Figure 3, the workflow includes preprocessing the CICIDS2017 dataset, selecting features and training the LightGBM teacher model, distilling knowledge into a lightweight neural network student, converting it to TensorFlow Lite, and integrating the TFLite model into Snort 3 via SnortML for inline detection.

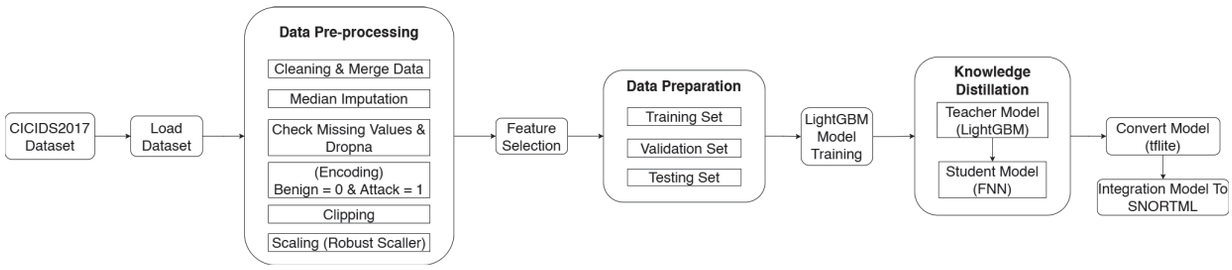


Fig. 3. Workflow of the preprocessing steps applied to the CICIDS2017 dataset.

A. Dataset Description

The CICIDS2017 dataset, from the Canadian Institute for Cybersecurity, is a standard benchmark for intrusion detection. It contains real network traffic with benign flows and various attacks [2], described by CICFlowMeter features, and exhibits class imbalance with normal traffic dominating [6]. These traits make it suitable for training and evaluating machine learning based intrusion detection models [7].

TABLE I
DAILY LABEL DATASET (CICIDS2017)

Day	Labels
Monday	BENIGN
Tuesday	Brute Force (BForce), SFTP, SSH
Wednesday	DoS attacks, Heartbleed Slowloris, Slowhttptest, Hulk, GoldenEye
Thursday	Web attacks; Infiltration attacks (Dropbox Download, Cool Disk) Web Brute Force, XSS; SQL Injection
Friday	DDoS LOIT, BotNet ARES, PortScans (sS, sT, sF, sN, sP, sV, sU, sO, sA, sW, sR, sL, B)

B. Data Pre-processing

Data pre-processing is an essential step to ensure that the CICIDS2017 dataset is clean, consistent, and suitable for machine learning analysis. Since the raw dataset contains multiple files with missing values, duplicated records, heterogeneous label formats, and numerical features with varying scales, several systematic procedures are required to transform the data into a standardized structure. These procedures include loading and merging all dataset files, cleaning and relabeling data, handling missing values, encoding categorical labels, performing imputation and outlier clipping, and applying appropriate feature scaling [2]. The overall workflow used is summarized in Figure 3.

C. Feature Selection

Feature selection is essential for reducing redundancy, improving interpretability, and increasing computational efficiency. After preprocessing and scaling, features with correlations above 0.9 were removed to reduce multicollinearity, and features with variance below 0.01 were also eliminated. From the initial 78 features, 32 were dropped for high correlation and 10 for low variance, leaving 36 more independent and meaningful features for modeling.

TABLE II
LIST OF LOW VARIANCE FEATURES

Features removed (variance threshold 0.1)
Bwd_PSH_Flags
Fwd_URG_Flags
Bwd_URG_Flags
RST_Flag_Count
Fwd_Avg_Bytes_Bulk
Fwd_Avg_Packets_Bulk
Fwd_Avg_Bulk_Rate
Bwd_Avg_Bytes_Bulk
Bwd_Avg_Packets_Bulk
Bwd_Avg_Bulk_Rate

In the final stage, LightGBM retained the 30 most important features to improve model efficiency.

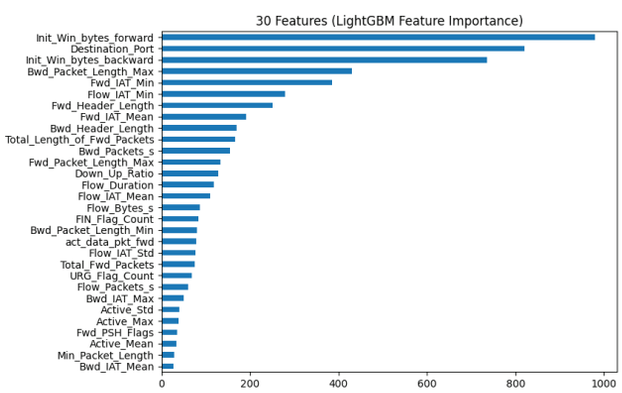


Fig. 4. Top 30 most important features selected by the LightGBM model.

D. Teacher Model Training Using LightGBM

The LightGBM model is a gradient boosting framework based on decision trees, designed to efficiently handle large-scale tabular data and capture complex feature interactions. The architecture consists of: [4]

TABLE III
TEACHER MODEL TRAINING CONFIGURATION USING LIGHTGBM

Parameter	Value
Base Learner	Decision Trees (leaf-wise growth)
Number of Estimators	500 trees
Learning Rate	0.05
Feature Input	30 selected features (CICIDS2017)
Random State	42
Parallelism	n_jobs = -1
Training Set	1,771,951 instances
Validation Set (1)	442,988 instances
Validation Set (2)	442,988 instances
Evaluation Metric	Log Loss
Early Stopping	50 rounds
Verbose	50

The model is trained with validation monitoring and evaluated on the test set, it then guides knowledge distillation to the student neural network.

E. Knowledge Distillation to the Neural Network Student Model

The student model is a neural network designed to learn from the teacher LightGBM model through knowledge distillation, capturing the teacher's predictive behavior while maintaining a lightweight and stable architecture [9]. The architecture consists of:

TABLE IV
STUDENT NEURAL NETWORK MODEL CONFIGURATION

Component	Description
Input Layer	30 features from preprocessed dataset
Dense Layer 1	128 neurons, ReLU activation, DO 0.20
Dense Layer 2	64 neurons, ReLU activation, D0 0.10
Dense Layer 3	32 neurons, ReLU activation
Output Layer	1 neuron with sigmoid activation
Loss Function	Mean Squared Error (for distillation)
Optimizer	Adam
Learning Rate	0.001
Batch Size	4096 samples
Epochs	50
Early Stopping	Monitor validation loss, patience 10, restore best weights
ReduceLROnPlateau	Monitor validation loss, patience 5
Model Checkpoint	Save best model based on validation loss

The teacher's output probabilities are softened (T=2.5) to train the student model, enabling it to mimic the teacher's behavior. After training, the student model is saved for SnortML integration.

F. Evaluation Metrics

The proposed intrusion prevention system was evaluated using standard metrics [9]. AUC-ROC measures the model's ability to distinguish benign from attack traffic, while the calibration curve assesses how well predicted probabilities reflect actual risk. The confusion matrix summarizes predictions into True Positives, True Negatives, False Positives, and False Negatives, providing insight into classification errors. Metrics such as accuracy, precision, recall, and F1-score quantify overall performance and the balance between false alarms and missed detections. ML Overhead measures the fraction of packets processed by the ML engine, indicating computational load and its impact on real-time performance. These metrics ensure the system is accurate and efficient under practical deployment conditions.

G. Integration of the TFLite Model into Snort via SnortML

The final workflow stage converted the trained student model to TensorFlow Lite (`student_model.tflite`) for real-time inference in Snort via SnortML. The model was transferred via GitHub, cloned on Ubuntu, and placed in `(/usr/local/src/libml/examples/classifier)` for deployment. SnortML was configured with a Lua script using `snort_ml_engine` to load the model and enable ML trace logs. Test PCAP files, including custom attacks and CICIDS2017 flows, verified that Snort correctly incorporated the model's probability outputs into its alert pipeline, enabling efficient, lightweight, and real-time intrusion detection [11].

IV. EXPERIMENTAL RESULT

This chapter evaluates the distilled neural network derived from LightGBM using 553,735 CICIDS2017 flows. It assesses classification performance, separation of benign and malicious traffic, and robustness in Snort via `snort_ml_engine`. Results show the student model's effectiveness both as a standalone classifier and as an inline IPS component under real inspection conditions.

A. Teacher Model Performance (LightGBM)

Table V presents the overall classification performance of the model, including accuracy, precision, recall, and F1-score for each class.

TABLE V
PERFORMANCE MEASURE OF LIGHTGBM MODEL

Class	Precision	Recall	F1-score	Support
0 (Benign)	0.9999	0.9997	0.9998	448576
1 (Attacsk)	0.9986	0.9995	0.9990	105159
Accuracy			0.9996	553735
Macro avg	0.9992	0.9996	0.9994	553735
Weighted avg	0.9996	0.9996	0.9996	553735

Figure 6 provides the confusion matrix, offering a detailed view of the model's correct and incorrect predictions for each class

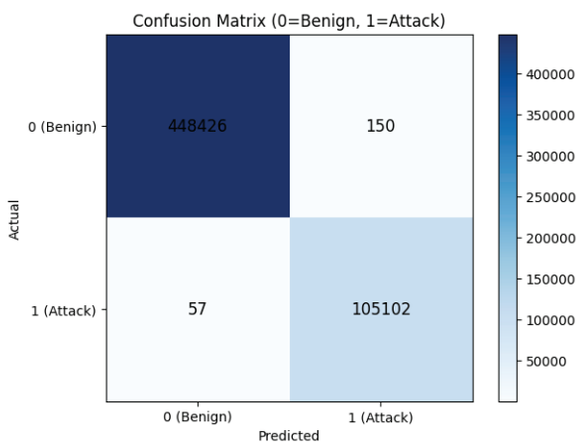


Fig. 5. Confusion matrix of the teacher model (LightGBM.

B. Student Model Performance (Neural Network)

Table VI summarizes the evaluation metrics of the student neural network model when tested on the CICIDS2017 dataset. The table presents precision, recall, F1-score, and support for each class, providing a clear overview of the model's capability in distinguishing between benign and attack traffic

TABLE VI
PERFORMANCE MEASURE OF FNN MODEL

Class	Precision	Recall	F1-score	Support
0 (Benign)	0.9697	0.9991	0.9842	448567
1 (Attack)	0.9956	0.8668	0.9267	105159
Accuracy			0.9740	553735
Macro avg	0.9826	0.9329	0.9555	553735
Weighted avg	0.9746	0.9740	0.9733	553735

Figure 7 shows the confusion matrix, summarizing the student model's predictions and errors.

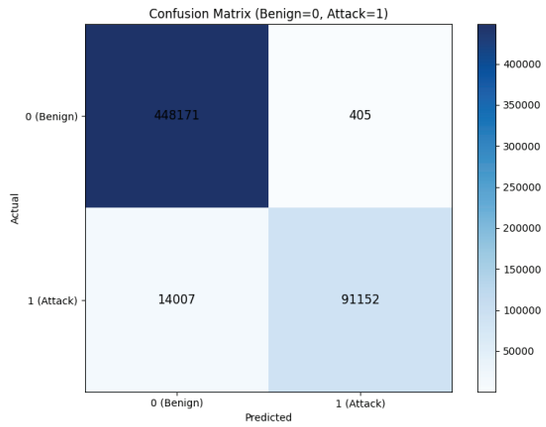


Fig. 6. Confusion matrix of the student neural network model on the CICIDS2017 test data.

Figure 8 summarizes the student model's performance, with clear class separation in the ROC curve and well-calibrated probabilities, confirming reliable detection.

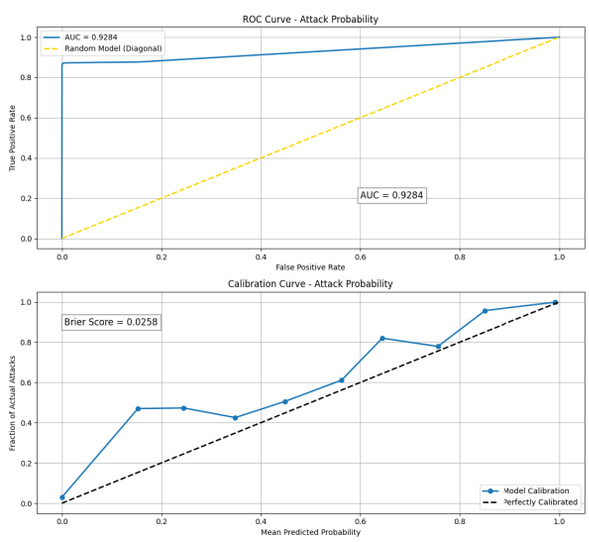


Fig. 7. Performance evaluation of the student model, including ROC & calibration curves.

C. Analysis and Comparison of SnortML and IPS Detection Models

To evaluate the integrated SnortML model, CICIDS2017 PCAPs from Monday to Friday were processed in passive mode, monitoring traffic without blocking. Table VII summarizes daily performance, including the number of packets analyzed, IPS alerts, SnortML detections, ML engine calls, and ML overhead. It should be noted that the testing was conducted solely on the dataset, so further analysis is required after implementing

the model in real network conditions. This provides insights into detection efficiency, ML inspection frequency, and computational impact, while also assessing the feasibility of real-time deployment in high-throughput networks.

TABLE VII
SNORTML AND IPS PERFORMANCE PER DAY USING CICIDS2017 PCAPS

Metric	Mon	Tue	Wed	Thu	Fri
Anlyz.Packets	88,31	98,23	48,98	66,70	64,24
IPS Alerts	603	738	144	765	244
Injected	57	92	13	539	46
Allow	70,54	80,96	45,11	55,45	52,13
LibML Calls	46	64	9	20	30
SnortML Alerts	46	64	9	20	29
SnortML Bytes	2,45	3,24	314	666	1,25
ML Overhead (%)	0.052	0.065	0.018	0.030	0.047

Table VII shows SnortML alerts closely follow LibML calls, indicating high efficiency, while IPS alerts vary with traffic. ML overhead is minimal (0.018%–0.065%), and SnortML reliably detects threats in passive mode. Table VIII compares teacher and student models, highlighting efficiency gains from feature compression on the CICIDS2017 dataset.

TABLE VIII
MODEL PERFORMANCE COMPARISON: FEATURE COMPRESSION VS. FULL FEATURE SET ON CICIDS2017 PCAP

Metric			Feature Comp.		Full Feature
Training Time (Teacher, s)			56.7		64.6
Training Time (Student, s)			182.1		321
Model Size (.tflite, bytes)			20,336		26,456
Accuracy			0.9427		0.9014
AUC			0.9284		0.7395
Brier Score			0.0258		0.0986
Metric	**Mon**	**Tue**	**Wed**	**Thu**	**Fri**
Feature Compression					
Seconds	2.611	2.332	1.230	1.207	1.230
Packets/sec	33,822	42,123	33,826	55,269	33,826
Mbits/sec	152	162	332	333	332
Full Feature					
Seconds	7.430	4.236	1.716	4.323	4.535
Packets/sec	11,886	23,189	28,590	15,428	14,165
Mbits/sec	53	93	237	93	87

V. CONCLUSION

This study enhances SnortML by integrating LightGBM feature selection with a knowledge-distilled neural network optimized for TensorFlow Lite. Using the CICIDS2017 dataset, LightGBM identifies the most informative features and provides soft decision boundaries to guide the training of a compact student network, achieving high detection performance with reduced computational cost. The distilled model demonstrates strong accuracy, reliable probability calibration, and minimal ML overhead in passive mode, while feature compression improves training efficiency and network throughput. However, evaluation is limited to dataset testing, and further validation under real network conditions is necessary. Overall, the framework enables SnortML to detect attacks beyond conventional signatures and provides a practical, scalable solution for real-time IPS deployment.

VI. ACKNOWLEDGEMENTS

The authors thank Cisco Talos for SnortML resources, the Canadian Institute for Cybersecurity for the CICIDS2017 dataset, and academic supervisors and colleagues for their guidance and support during experimentation and evaluation.

REFERENCES

[1] Cisco Talos Team: SnortML: Machine Learning-based Exploit Detection', document for Snort 3 release 7.6, Cisco Systems, 2024. Available: https://docs.snort.org/misc/snort_ml (accessed 17 November 2025)

[2] Kikissagbe, B.R., Adda, M.: 'Machine Learning-Based Intrusion Detection Methods in IoT Systems: A Comprehensive Review', Electronics, 2024, 13, p. 3601. doi: 10.3390/electronics13183601

[3] Akhtar, Z.B., & Tajbiul Rawol, A.: 'Enhancing Cybersecurity through AI-Powered Security Mechanisms', IT Journal Research and Development, 2024, 9(1), pp. 50–67.

[4] Vaishali, R. and Naik, S.M.: 'LightGBM-Bayesian Approach for DDoS Detection in SDN', Proc. MERCon 2024, pp. 7–12.

[5] Oyinloye, T.S., Arowolo, M.O., Prasad, R.: 'Improved ANN model for cyber threat detection', Data Sci. Manag., 2025, 8(1), pp. 107–115.

[6] Nigar, N., Mustafa, R.: 'Enhanced Intrusion Detection via Hybrid Data Resampling and Feature Optimization', IEEE Access, 2025.

[7] Saleem, M., Azam, M., Mubeen, Z., Mumtaz, G.: 'Machine learning for improved threat detection: LightGBM vs. CatBoost', Journal of Computing & Biomedical Informatics, 2024, 7(1), pp. 571–580.

[8] Ke, G., Meng, Q., Finley, T., et al.: 'LightGBM: A Highly Efficient Gradient Boosting Decision Tree', NeurIPS, 2017, vol. 30, pp. 3146–3154.

[9] Umirzakova, S., Abdullaev, M., Mardieva, S., Latipova, N., Muksimova, S.: 'Simplified knowledge distillation with a novel teacher–student architecture', Electronics, 2024, 13(22), p. 4530.

[10] Wiranta, A., Karna, N., Irawan, A., Prakoso, I.A.: 'Implementation and Analysis of Network Security in Raspberry Pi against DOS Attack with HIPS Snort', Telkom University, 2025.

[11] Fadlillah, A., Karna, N., Irawan, A.: 'IDS Performance Analysis using Anomaly-based Detection Method for DOS Attack', Telkom University, 2025.

Design and Deployment of an Autoencoder–Binary Classifier Intrusion Prevention Model for SnortML

1st Alika Widuri Kartika
School of Electrical Engineering
Telkom University
Bandung, Indonesia
alikawkartika@gmail.com

2nd Bintang Pandu Anugrah
School of Electrical Engineering
Telkom University
Bandung, Indonesia
bintang.pandu87@gmail.com

3th Dwaynimay Farrel Rajendra
School of Electrical Engineering
Telkom University
Bandung, Indonesia
dwaynimayfarrel@gmail.com

4th Salna Siraz Maulidia
School of Electrical Engineering
Telkom University
Bandung, Indonesia
salnasiraz.maulidia@gmail.com

5th Nyoman Karna
The Univ CoE Intelligent Sensing-IoT
Telkom University
Bandung, Indonesia
aditya@telkomuniversity.ac.id

3th Arif Indra Irawan
School of Electrical Engineering
Telkom University
Bandung, Indonesia
arifirawan@telkomuniversity.ac.id

***Abstract*—Traditional signature-based Intrusion Prevention Systems (IPS) face inherent limitations when confronted with polymorphic, obfuscated, and zero-day attack variants. To address this gap, this study introduces a lightweight hybrid intrusion-prevention model that combines an autoencoder trained exclusively on benign CIC-IDS2017 traffic with a compact supervised binary classifier. The autoencoder learns low-dimensional latent representations of normal flows, while reconstruction error is concatenated with these latent vectors to form a 33-dimensional feature set for classification. Class weighting is applied to mitigate the severe imbalance characteristic of real-world network traffic. The proposed hybrid model achieves 99.83% accuracy and precision–recall values exceeding 0.996 on the test split, demonstrating a substantial improvement in minority-class sensitivity. To enable deployment within operational IPS workflows, the model is converted to TFLite and validated against SnortML's strict inference constraints. Deployment across five daily PCAP subsets confirms that the model augments rather than overwhelms Snort's signature engine. These results highlight a practical and resource-efficient pathway for integrating modern machine-learning inference into inline IPS environments.**

***Index Terms*—Intrusion prevention, autoencoder, binary classifier, SnortML, anomaly detection**

I. Introduction

The rapid expansion of modern network infrastructures, cloud services, and IoT ecosystems has significantly increased organizational exposure to sophisticated cyber-attacks. Traditional signature-based Intrusion Prevention Systems (IPS), rely on predefined attack signatures to identify malicious behavior. Although effective in detecting known threats, this mechanism is inherently limited because it cannot recognize previously unseen, polymorphic, or zero-day attacks that do not match existing signatures. As highlighted in recent IPS surveys, signature-based systems must constantly maintain updated signature databases and remain unable to detect novel intrusions, leading to reduced effectiveness in dynamic and high-dimensional network environments [1].

Machine learning (ML) approaches offer a promising alternative by enabling data-driven detection of complex traffic patterns, including unseen attacks. Deep learning models, particularly autoencoders, have shown strong capability in capturing latent representations of high-dimensional network flows. Lansky et al. reported that deep learning–based IDS outperform classical ML methods in scalability, feature abstraction, and generalization. However, many ML models despite their detection advantages remain computationally heavy and unsuitable for real-time IPS frameworks such as SnortML [2].

A key challenge arises from the characteristics of datasets such as CIC-IDS2017, which exhibit severe class imbalance, distributional drift over days, and high-dimensional feature spaces. Zavrak and Iskefiyeli noted that autoencoder-based anomaly detectors are highly sensitive to imbalance and may generate unstable detection boundaries, leading to significant false-positive rates when deployed in operational settings [3]. This makes it necessary to design a model that is both computationally efficient and robust to imbalance, while remaining compatible with SnortML's strict low-latency inference requirements.

To address these limitations, this study proposes a lightweight hybrid intrusion-prevention model that combines an autoencoder for latent feature extraction with a compact binary classifier. The autoencoder learns compressed representations from benign traffic, while the classifier trained using class weighting enhances separability between benign and malicious flows. The final TFLite model satisfies the structural constraints of SnortML, using a single float32 input tensor and a single scalar output tensor, enabling seamless integration into the Snort rule engine [9].

II. Background and Related Work

This section provides the theoretical background supporting the proposed intrusion-prevention approach, covering the limitations of signature-based IPS, the integration of machine learning through Snort and SnortML, and the rationale for employing autoencoders and binary classifiers.

A. Machine Learning–Enhanced Intrusion Prevention and Detection

Traditional signature-based Intrusion Prevention Systems (IPS) such as Snort face inherent limitations in detecting polymorphic or previously unseen attacks due to their reliance on handcrafted signatures [4]. Although effective against known threats, Snort's static rule-matching mechanism allows zero-day variants and obfuscated payloads to bypass detection [1] [4]. Practical HIPS deployments further highlight this limitation, running Snort on a Raspberry Pi drops the SYN/UDP Flood traffic but drives CPU usage above 80% during attacks, revealing the operational cost of purely rule-based inline prevention on constrained hardware [11].

SnortML addresses these shortcomings by embedding machine-learning inference directly into the Snort 3 rule pipeline. Through the model keyword, TensorFlow Lite is invoked during packet inspection, enabling lightweight ML decision-making. To maintain deterministic low-latency performance, SnortML restricts models to a single input tensor, a float32 scalar output, and minimal computational complexity [9]. Such constraints are critical because Snort's AF_PACKET capture path can drop packets under high load if inference introduces additional latency [10]. In practice, these design rules ensure that ML evaluation can operate inline without introducing measurable throughput degradation. As a result, SnortML provides a viable foundation for integrating learning-based detection into operational IPS workflows while preserving the performance characteristics of traditional rule-based inspection.

B. Autoencoder

An autoencoder is a feed-forward neural network trained to reconstruct its input at the output layer. It consists of two main components: an encoder $f_{\boldsymbol{w}}(\cdot)$ that maps an input vector $\mathbf{x}$ to a latent representation $\mathbf{z}$, and a decoder $g_{\boldsymbol{u}}(\cdot)$ that reconstructs the input as $\hat{\mathbf{x}}$ [5]. Formally, the model computes

$$\hat{x} = g_{\boldsymbol{u}}(f_{\boldsymbol{w}}(x)) \,, \tag{1}$$

and is optimized by minimizing the reconstruction error

$$\min_{\boldsymbol{w},\boldsymbol{u}} \|x - g_{\boldsymbol{u}}(f_{\boldsymbol{w}}(x))\| \,, \tag{2}$$

where $\|\cdot\|$ denotes the ℓ_2-norm.

As illustrated in Fig. 1, an autoencoder comprises an encoder that compresses the input into a low-dimensional latent space, a bottleneck layer that captures the most salient features, and a decoder that reconstructs the original vector. This structure enables the model to learn compact and meaningful representations of the input data, which are later leveraged for downstream classification.

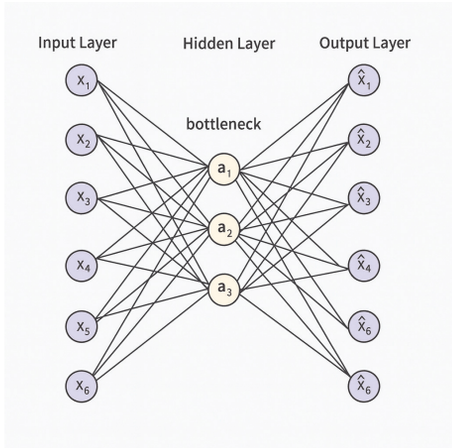


Fig. 1: Structure of the autoencoder, consisting of an encoder, a bottleneck latent space, and a decoder that reconstructs the input.

C. Binary Classifier

A binary classifier is a supervised learning model designed to categorize input data into two distinct classes, typically denoted positive and negative $(0/1)$. Formally, a binary classifier can be represented as a function

$$h : \mathcal{X} \to \{0, 1\} \tag{3}$$

Where $\mathcal{X}$ denotes the feature space of the input data.

Binary classification is widely used in machine learning due to its ability to form explicit decision boundaries between two categories. However, its effectiveness can be significantly affected by factors such as class noise, limited training data, and class imbalance. As noted by Leevy *et al.*, binary classifiers tend to bias toward the majority class when the minority class is extremely rare, which is a common condition in anomaly detection and security applications [6].

III. Research Methodology

This chapter explains the overall methodological framework adopted in this study. The research workflow consists of several sequential stages, begins with selecting the CIC-IDS 2017 dataset, preprocessing procedures, feature extraction using an autoencoder, and ending with binary classification training and evaluation. The complete process is illustrated in Fig. 2.

A. Dataset Selection

The CIC-IDS2017 dataset is a widely used benchmark for intrusion detection research, providing a comprehensive representation of both normal and malicious network traffic. It was generated using realistic network activities, including benign behaviors such as web browsing, email, and file transfers, as well as multiple types of cyberattacks, including DoS, DDoS, brute-force, infiltration, and botnet attacks. The dataset contains flow-based network features extracted from packet captures, such as source and destination IP addresses, ports, protocols, flow duration, packet counts, and byte counts,

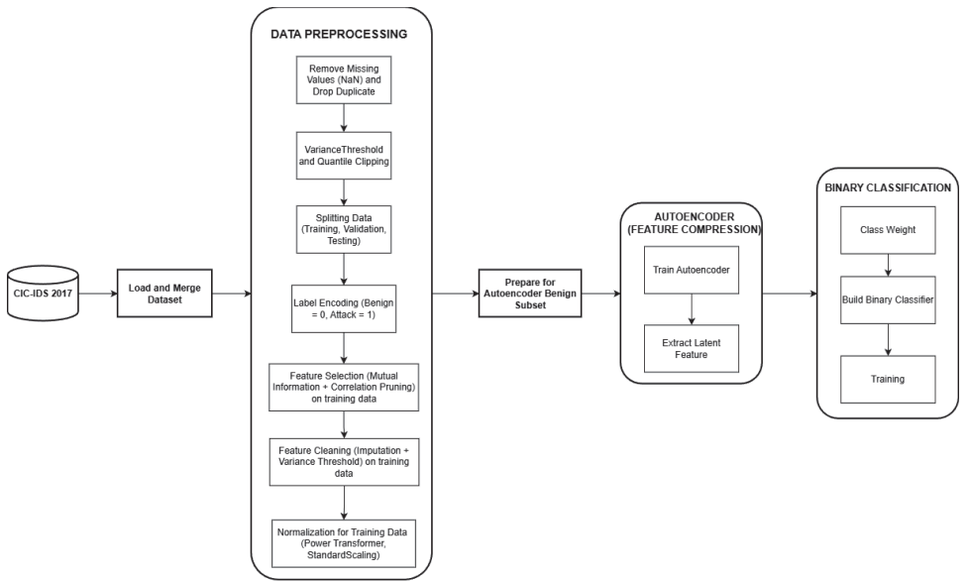


Fig. 2: Comprehensive Workflow of the Proposed Intrusion Detection Approach, from Data Preprocessing to Feature Compression and Binary Classification.

offering a rich feature space suitable for machine learning models. CIC-IDS2017 exhibits high class imbalance, with normal traffic significantly outnumbering attack traffic, reflecting real-world network conditions [7]. Table I summarizes the number of instances for each traffic class.

TABLE I: Class distribution of the CIC-IDS2017 dataset.

Traffic Class	Number of Instances
BENIGN	2,096,484
DoS Hulk	172,849
DDoS	128,016
PortScan	90,819
DoS GoldenEye	10,286
FTP-Patator	5,933
DoS slowloris	5,385
DoS Slowhttptest	5,228
SSH-Patator	3,219
Bot	1,953
Web Attack – Brute Force	1,470
Web Attack – XSS	652
Infiltration	36
Web Attack – SQL Injection	21
Heartbleed	11

The CIC-IDS2017 dataset was selected for its realistic traffic behavior, diverse attack scenarios, rich flow-based features, and public availability, enabling reproducible and comparable evaluation.

B. Preprocessing Procedure

The data preprocessing procedure begins by removing missing values (NaN) and eliminating duplicate entries to ensure a clean and consistent dataset. A Variance Threshold is then applied to discard features with extremely low variability, followed by quantile clipping to mitigate the influence of extreme outliers. The dataset is subsequently split into training, validation, and testing subsets, after which label encoding is performed using the scheme Benign = 0 and Attack = 1. Next, feature selection is conducted exclusively on the training data using a combination of Mutual Information and correlation pruning to prevent data leakage. The selected features undergo feature cleaning, which includes imputation of missing values and an additional variance check. Finally, normalization is applied using a Power Transformer (Yeo–Johnson method) to stabilize non-Gaussian feature distributions, followed by Standard Scaling based on Z-Score normalization (mean = 0, standard deviation = 1) to ensure all features operate on a comparable scale, improving model stability and performance.

C. Autoencoder Training

A feed-forward denoising autoencoder was employed to reduce feature dimensionality and learn latent representations of benign traffic. The model was trained solely on benign samples from CIC-IDS2017 so that the encoder could capture the statistical structure of normal flows. The input dimension corresponds to the selected and preprocessed features.

The architecture includes Gaussian noise at the input layer for robustness, followed by Dense and Batch Normalization layers that compress the feature vector into a 32-dimensional bottleneck. The decoder mirrors this structure to reconstruct the original input. The network was optimized using Adam with Mean Squared Error (MSE) reconstruction loss. Training stability was ensured via ModelCheckpoint, EarlyStopping (patience=12), and ReduceLROnPlateau. Benign samples were subsampled for efficiency, and both the full autoencoder and encoder subnetwork were saved for downstream latent-feature extraction.

D. Class Imbalance Analysis

The CIC-IDS2017 dataset exhibits significant class imbalance, with 2,096,484 benign samples and 557,646 attack

TABLE II: Structural overview of the Autoencoder.

Layer (type)	Output Shape	Params
Input (66)	(None, 40)	0
GaussianNoise	(None, 40)	0
Dense	(None, 256)	10,496
BatchNorm	(None, 256)	1,024
Dense	(None, 128)	32,896
BatchNorm	(None, 128)	512
Bottleneck (Dense)	(None, 32)	4,128
Dense	(None, 128)	4,224
BatchNorm	(None, 128)	512
Dense	(None, 256)	33,024
Output (Dense)	(None, 40)	10,280

samples, resulting in a 1:3.7 ratio. To compensate for this skew, balanced class weights were computed, assigning weight 0.599 to the benign class and 3.022 to the attack class. These weights amplify minority-class contributions during classifier optimization, directly addressing the imbalance issue. The effectiveness of this mitigation is reflected in the model performance: the autoencoder alone achieved an attack recall of only 0.8067 due to its unsupervised nature, whereas the hybrid autoencoder–binary classifier, with class weighting, improved attack recall to 0.9932. This substantial improvement empirically confirms that the applied imbalance-handling strategy successfully enhances the model's sensitivity to minority attack samples while maintaining overall stability.

E. Latent Feature Extraction

After training, the autoencoder was restored using the best validation weights. The training, validation, and test sets were passed through the encoder to obtain 32-dimensional latent vectors and through the full autoencoder to compute reconstruction errors, defined as the mean squared difference between each input and its reconstruction. Each latent vector was concatenated with its scalar reconstruction error, forming a unified 33-dimensional feature representation. This process generated matrices of size $1{,}801{,}984 \times 33$ for training and $386{,}140 \times 33$ for validation and testing, completed in 12.4 seconds. Labels were assigned using the binary mapping *BENIGN = 0* and *ATTACK = 1*, followed by consistency checks to ensure alignment between features and labels. These representations served as the input to the subsequent binary classifier.

F. Binary Classifier Training

The binary classifier operates on a 33-dimensional feature vector composed of the 32-dimensional latent representation and a scalar reconstruction error. It is implemented as a lightweight feed-forward network consisting of an input layer, a 128-unit ReLU Dense layer followed by 0.3 dropout, a 32-unit ReLU layer, and a sigmoid output producing the attack probability. Training uses the Adam optimizer with binary cross-entropy and AUC as the primary metric.

To mitigate class imbalance, balanced class weights (0.599 for benign and 3.022 for attacks) were applied to emphasize minority-class samples. Training stability was maintained using ModelCheckpoint, EarlyStopping (patience 12, best-weight restoration), and ReduceLROnPlateau. The best-performing model was saved for evaluation and deployment. The classifier architecture prior to training is summarized in Table III.

TABLE III: Architecture of the binary classifier model before training.

Layer (type)	Output Shape	Param #
clf_input (InputLayer)	(None, 33)	0
Dense (128 units)	(None, 128)	4,352
Dropout (0.3)	(None, 128)	0
Dense (32 units)	(None, 32)	4,128
Dense (sigmoid output)	(None, 1)	33

G. Evaluation Metrics

The hybrid model is evaluated using standard accuracy-based metrics to assess detection performance under class imbalance. Accuracy provides overall correctness, while precision and recall quantify the reliability of attack prediction and the ability to detect true attacks. As highlighted by Christen et al., these two metrics are representational because they correspond to empirical probabilities of correct classification. The F1-score combines precision and recall using the harmonic mean, offering a concise summary suited for imbalanced intrusion-detection tasks. To complement these metrics, the Precision–Recall AUC (PR-AUC) is used to evaluate threshold-independent behavior, providing a more informative measure when attacks form the minority class. All reported results are derived directly from empirical evaluation on the CIC-IDS2017 test split.

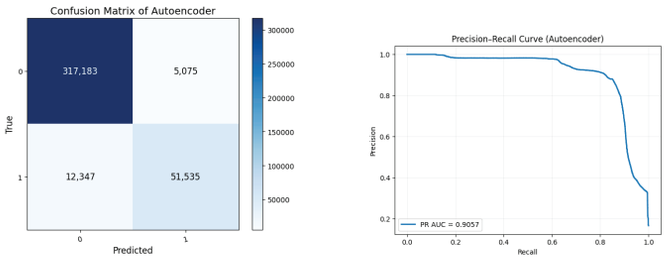
IV. Experimental Result

This section presents the empirical evaluation of the proposed intrusion-prevention model. The autoencoder is first assessed to examine its reconstruction behavior and latent-feature separability. The hybrid autoencoder–binary classifier is then evaluated using standard detection metrics, followed by deployment testing in SnortML to verify compatibility and operational performance. Overall, the results demonstrate the model's effectiveness from feature learning to practical deployment.

A. Autoencoder Evaluation Analysis

The autoencoder evaluation results in Fig. 3a show that the model effectively reconstructs benign flows, yielding 317,183 true negatives and 51,535 true positives. However, the presence of 12,347 false negatives and 5,075 false positives indicates that the autoencoder alone cannot reliably distinguish attack traffic.

The reconstruction loss curve in Fig. 4 exhibits noticeable fluctuations during the early and mid stages of training, indicating that the autoencoder requires more iterations to stabilize its learning dynamics. Although the loss eventually converges, the irregular descent pattern suggests that the model experiences variability while capturing the underlying data structure. This behavior reflects the challenge of learning consistent representations from heterogeneous benign traffic.



(a) Confusion matrix (b) PR curve

Fig. 3: Autoencoder evaluation showing (a) confusion matrix and (b) precision–recall curve.

The Precision–Recall (PR) curve in Fig. 3b reflects the model's limited discriminative ability with a PR-AUC of 0.9057. The classification report in Table IV further confirms this limitation, with the Attack class achieving a reduced recall of 0.8067.

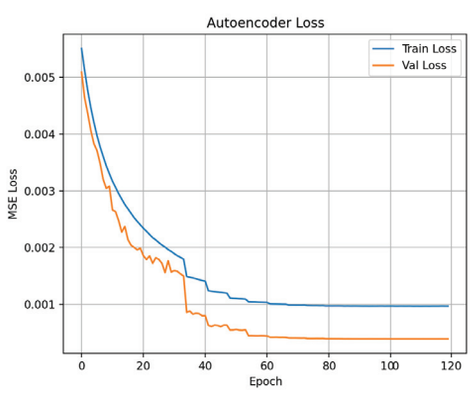


Fig. 4: Autoencoder reconstruction loss (MSE) over training epochs.

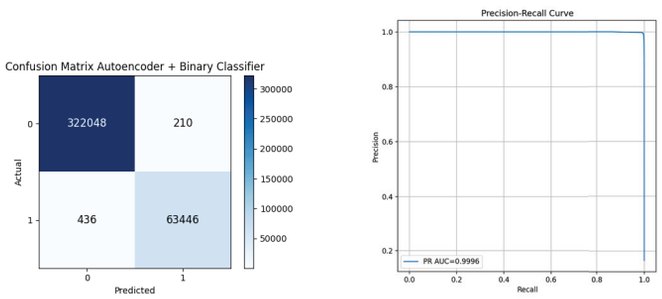
TABLE IV: Classification Report of Autoencoder Model

Class	Precision	Recall	F1-score	Support
0 (Normal)	0.9625	0.9843	0.9733	322258
1 (Attack)	0.9104	0.8067	0.8554	63882
Accuracy			0.9549	386140
Macro avg	0.9364	0.8955	0.9143	386140
Weighted avg	0.9539	0.9549	0.9538	386140

Overall, while the autoencoder successfully captures dominant patterns of benign traffic, its latent representation does not provide sufficient separability between normal and malicious flows. These findings motivate the need for a hybrid supervised classifier to enhance detection performance using both latent features and reconstruction error.

B. Hybrid Autoencoder–Binary Classification Evaluation

The hybrid autoencoder-binary classifier shows a substantial performance improvement over the standalone autoencoder. As shown in Fig. 5a, misclassifications drop sharply to 210 false positives and 436 false negatives. The Precision-Recall



(a) Confusion matrix (b) PR curve

Fig. 5: Evaluation of the hybrid autoencoder–binary classifier: (a) confusion matrix and (b) precision–recall curve.

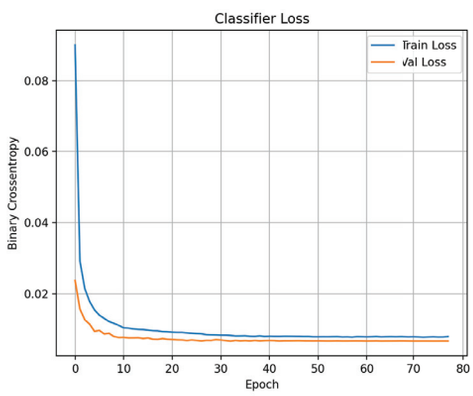


Fig. 6: Binary cross-entropy loss over training epochs.

(PR) curve in Fig. 5b indicates near-perfect class separation, achieving a PR-AUC of 0.9999.

As shown in Fig. 6, the binary classifier's loss decreases rapidly and stabilizes early, forming a smooth trajectory. Unlike the autoencoder in Fig. 4, which shows greater fluctuation before convergence. The final metrics in Table V confirm strong performance, achieving 99.83% accuracy with precision and recall above 0.996.

TABLE V: Classification Report of Hybrid Autoencoder–Binary Classifier

Class	Precision	Recall	F1	Support
0 (Normal)	0.9986	0.9993	0.9990	322258
1 (Attack)	0.9967	0.9932	0.9949	63882
Accuracy			0.9983	386140
Macro avg	0.9977	0.9963	0.9970	386140
Weighted avg	0.9983	0.9983	0.9983	386140

The t-SNE plot (Fig. 7) shows distinct clustering between benign and attack samples, confirming that the latent representation and reconstruction error provide a highly separable feature space for classification. Benign flows form a dense and coherent region, while attack samples appear in multiple compact clusters detached from the benign distribution, indicating that the hybrid feature space captures meaningful structural differences that enable robust decision boundaries even under severe class imbalance.

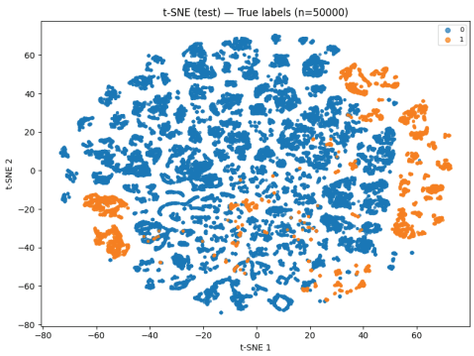


Fig. 7: t-SNE visualization of hybrid-model feature space.

C. Deployment and Quality Verification in SnortML

Prior to deployment, the hybrid autoencoder–binary classifier was validated against SnortML's structural constraints. Five CIC-IDS2017 PCAP subsets (Monday–Friday), each containing roughly 100,000 flows, were then processed using the converted TFLite model. The proposed model consistently generated only 6–11 ML-based alerts per subset, indicating stable inference and low false-positive behavior even across varying daily traffic conditions.

TABLE VI: SnortML Detection Summary for Daily PCAP Subsets (Mon–Fri)

Parameter	Mon	Tue	Wed	Thu	Fri
Flows	88,309	98,225	48,976	66,695	64,236
SnortML Alerts	46	64	9	20	30
Rule-Based Alerts	557	674	135	745	214
Total Alerts	603	738	144	765	244
ML Alert Rate (%)	0.052%	0.065%	0.018%	0.030%	0.047%
ML Contribution (%)	7.6%	8.7%	6.3%	2.6%	12.3%

Across all subsets, the model averages only 33.8 ML alerts per day (0.042%), demonstrating that the hybrid approach integrates cleanly into SnortML and contributes a controlled, predictable amount of ML-driven detection. These results confirm that the model enhances coverage without overwhelming the signature engine.

Compared with prior HIPS deployments on Raspberry Pi, which dropped 28–41% of DoS packets but consumed over 83% CPU under attack [11], the proposed approach delivers improved stability with far lower resource impact. Similarly, anomaly-based IDS studies [12] reported high accuracy but false-positive rates above 54–67% on web traffic. In contrast, our integration maintains sub-0.065% alert rates across all days, offering a more balanced trade-off between operational efficiency, detection reliability, and false-positive control.

V. Conclusion

This work presents a lightweight hybrid autoencoder–binary classifier as an effective enhancement for SnortML-enabled Intrusion Prevention Systems. By combining latent representations learned from benign traffic with reconstruction error, the model achieves strong separability between normal and malicious flows while remaining computationally efficient. The exported TFLite model satisfies SnortML's structural constraints, enabling seamless low-latency integration within the Snort rule engine. Experimental results show high detection performance, with 99.83% accuracy and precision/recall exceeding 0.996, while deployment on daily PCAP subsets yields stable, low-volume ML alerts that complement rather than burden signature-based detection. Overall, the proposed approach provides a practical pathway for incorporating machine-learning inference into operational IPS workflows. Future work will explore robustness under traffic drift and further optimization for resource-constrained edge environments.

VI. Acknowledgement

The authors would like to thank the Faculty of Electrical Engineering at Telkom University for providing computational infrastructure and academic support throughout this research. The authors also acknowledge the maintainers of the CIC-IDS2017 dataset and the SnortML development team for enabling reproducible experimentation.

References

Journal articles

[1] N. Gupta, V. Jindal, and P. Bedi, "A Survey on Intrusion Detection and Prevention Systems," *SN Computer Science*, vol. 4, no. 439, pp. 1–28, 2023.

[2] Lansky, J., Ali, S., Mohammadi, M., Majeed, M.K., Karim, S.H.T., Rashidi, S., Hosseinzadeh, M., Rahmani, A.M.: 'Deep Learning-Based Intrusion Detection Systems: A Systematic Review', IEEE Access, 2021, 9, pp. 101574–101604.

[3] S. Zavrak and M. İskefiyeli, 'Anomaly-based intrusion detection from network flow features using variational autoencoder, IEEE Access, vol. 8, pp. 108346–108357, 2020.

[4] N. Alsharabi, M. Alqunun, and B. Murshed, "Detecting unusual activities in local network using Snort and Wireshark tools," *Journal of Advances in Information Technology*, vol. 14, no. 4, pp. 616–624, 2023.

[5] Z. Cheng, S. Wang, P. Zhang, S. Wang, X. Liu, and E. Zhu, "Improved autoencoder for unsupervised anomaly detection," *International Journal of Intelligent Systems*, pp. 1–23, 2021.

[6] J. L. Leevy and J. Hancock, "Investigating the effectiveness of one-class and binary classification for fraud detection," Journal of Big Data, vol. 10, no. 157, pp. 1–16, 2023.

[7] Nigar, N., Mustafa, R.: 'Enhanced Intrusion Detection via Hybrid Data Resampling and Feature Optimization', IEEE Access, 2025, 13, pp. 149100–149124.

[8] P. Christen, D. J. Hand, and N. Kirielle, "A review of the F-measure: Its history, properties, criticism, and alternatives," *ACM Computing Surveys*, vol. 56, no. 3, pp. 73:1–73:24, 2023.

[9] Cisco Talos, "SnortML: Machine learning integration for Snort 3," *Snort 3 Official Documentation*. Accessed: November 24, 2025. [Online]. Available: https://docs.snort.org/misc/snort_ml

[10] M. Sheeraz, M. H. Durad, S. Tahir, H. Tahir, S. Saeed, and A. M. Almuhaideb, "Advancing Snort IPS to Achieve Line-Rate Traffic Processing for Effective Network Security Monitoring," *IEEE Access*, vol. 12, pp. 61848–61859, 2024.

[11] A. Wiranata, N. Karna, A. Irawan, and I. A. Prakoso, "Implementation and Analysis of Network Security in Raspberry Pi Against DoS Attack with HIPS Snort," in *Proc. 2023 Int. Conf. on Computer Science, Information Technology and Engineering (ICCoSITE)*, pp. 892–896, 2023, doi: 10.1109/ICCoSITE57641.2023.10127741.

[12] A. Fadhlillah, N. Karna, and A. Irawan, "IDS Performance Analysis Using Anomaly-based Detection Method for DoS Attack," in *Proc. 2020 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS)*, pp. 18–22, 2021, doi: 10.1109/IoTaIS50849.2021.9359719.

IOTA Tangle-Based Digital Evidence Anchoring System for International Investigative Assistance

Chang-Jun Choi*1, Hyun-Soo Kim*2, Kyung-Yeob Park*3 and †Dong-Myung Shin*4
* *R&D Division, LSware Inc, Seoul, Republic of Korea*
Emails: *{ jacob, hskim94, carl, roland } @ lsware.com*

***Abstract*—This paper proposes digital evidence anchoring system that combines IOTA Tangle–based distributed ledger technology with on-device encryption to guarantee end-to-end integrity and confidentiality of digital evidence in international cooperative investigations. In the proposed system, hash values and metadata of evidence captured in the field are immediately encrypted and stored on the investigator's device, and, once a secure connection is available, are anchored on IOTA Tangle through a feeless mechanism, enabling integrity verification at any time and from any location via hash comparison. A common verification mechanism based on hashes and anchor IDs, together with a separation of encrypted evidence files and metadata, provides interoperability across heterogeneous national and institutional systems while keeping sensitive original data off-chain to balance personal data protection and investigative transparency. The system is intended to serve as a trustworthy digital evidence management infrastructure, and future work will evaluate its performance and robustness in real-world copyright infringement and cybercrime investigations and refine its functionality based on feedback from investigative agencies.**

Keywords— International Investigative Assistance, Digital Evidence, Anchoring System, IOTA Tangle, On-Device

I. INTRODUCTION

In an international cooperative investigation environment, digital evidence is a key factor that determines the progress and outcome of a case, and blockchain-based Distributed Ledger Technology (DLT) is regarded as a strong candidate for ensuring the integrity and reliability of such evidence[1]. Blockchain records transactions in an immutable form and employs consensus mechanisms to make tampering difficult, which makes it highly suitable for transparently tracking "who recorded what, and when" throughout the collection, storage, and submission of digital evidence.

However, in real investigative settings, large volumes of digital evidence are generated, and integrity anchoring must often be performed in real-time under constrained network and device resources. In particular, it is difficult to upload in real-time and synchronize stably in environments where communication infrastructure is limited, such as overseas sites or disaster/conflict areas, and the possibility of being exposed to loss, damage, and unauthorized copying increases while evidence stays only on local devices for a long time[2]. Even after connectivity is restored, if the data left on local devices is not synchronized accurately with central or distributed verification systems, the continuity of evidence preservation within the same case may be broken or verification results may conflict across countries and agencies, ultimately undermining the legal validity of the evidence and the overall credibility of international cooperative investigations. In such situations, although traditional blockchain structures and consensus mechanisms are sufficiently secure, they may not be able to satisfy all requirements simultaneously in terms of throughput, latency, fees, and support for lightweight devices. To address these limitations, it is necessary to explore alternative distributed ledger structures that better accommodate the constraints of field operations and the requirements of international cooperative investigations, while preserving the immutability and traceability offered by the blockchain.

Therefore, this study proposes digital evidence anchoring system that combines the Directed Acyclic Graph (DAG)–based structure of IOTA Tangle with On-Device encryption techniques. Blockchain is a highly useful technology for recording digital evidence unchanged, but considering fees, processing delays, and resource constraints, the DAG-based structure and non-fee verification model provided by IOTA Tangle can provide a more suitable alternative to the international cooperative investigation environment. IOTA Tangle provides high scalability and parallelism by performing distributed verification through reference relationships between transactions, and can secure integrity and traceability without a centralized miner. By integrating on-device encryption, the proposed approach aims to securely manage hash values and metadata from the moment evidence is generated on an investigator's device, and to ensure end-to-end integrity regardless of network connectivity. In this way, digital evidence collected in the field can be protected throughout its entire lifecycle and managed in a form that allows mutual verification in international cooperative investigations.

II. RELATED WORK

A. IOTA Tangle

IOTA Tangle is a distributed ledger technology that, unlike traditional blockchains, connects transactions in the form of a DAG, as shown in Fig. 1, rather than a linear chain[3]. Each new transaction selects and approves two or more previous transactions in order to join the network, and through this process the overall consistency and integrity of the ledger are maintained without dedicated miners or a centralized validator. Because this DAG-based ledger data structure allows multiple transactions to be processed and validated in parallel instead of being handled sequentially in block units, it provides high scalability and throughput, and supports fee-less transaction processing, which makes it well suited to investigative environments where metadata and hash values for digital evidence must be recorded frequently.

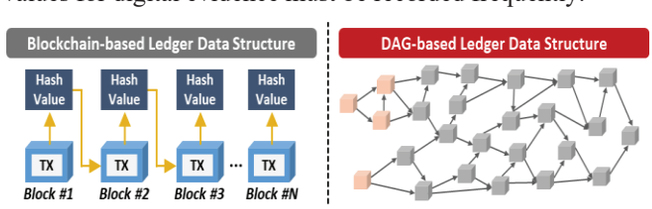


Fig. 1. Differences in data structure between Blockchain and DAG

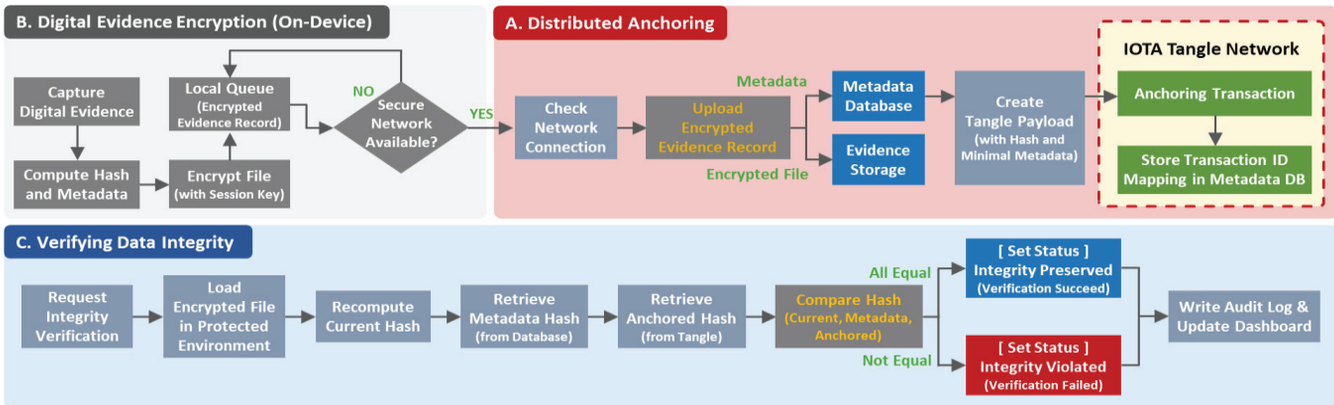


Fig. 2. Overall Process Flow of the Proposed System

From a data anchoring perspective, only the hash values of digital evidence and a minimal set of metadata are recorded on IOTA Tangle, rather than the evidence itself. By re-querying a transaction that contains the same hash value, any party can independently verify whether the corresponding evidence has been tampered with. When nodes are distributed across multiple countries, this creates a foundation for globally verifying the integrity of digital evidence used in international cooperative investigations, regardless of infrastructure failures or policy changes in any single country. In this way, Tangle functions as a complementary distributed ledger technology that preserves the immutability and traceability offered by blockchain, while mitigating fee and latency overhead and strengthening the real-time processing, large-scale throughput, and cross-jurisdiction verification capabilities required in international cooperative investigations.

B. *Data Encryption in On-Device*

In an On-Device environment, data encryption is critically important in international cooperative investigation because it protects digital evidence before it is ever transmitted to a central server or cloud. Videos, images, log files, and other forms of evidence captured or collected in the field are initially stored on portable devices such as smartphones, tablets, or forensic tools; if tampering or leakage is not prevented at this stage, even a strong integrity verification mechanism applied later cannot fully prove the authenticity of the “original moment of creation”[4]. Accordingly, recent approaches have focused on On-Device encryption techniques in which security modules and lightweight AI models embedded in the device immediately apply symmetric encryption (e.g., AES-256-GCM) at the time data is generated, while simultaneously computing hash values and key metadata and storing them in a local secure area[5].

Since On-Device encryption operates regardless of network connection, it provides a basis for maintaining the confidentiality and integrity of digital evidence even in environments with limited communication infrastructure, such as overseas sites, disaster areas, and military operations[6]. The device stores encrypted files together with their hashes and metadata as records in a local queue, and once a secure connection is available, these records are uploaded sequentially. In addition, policy logic or On-Device AI embedded in the device can automatically decide which data should be stored under which key domain, and when it should be anchored to the Tangle, based on the data type, sensitivity, and case context. This reduces the investigator’s manual workload and risk of operational error, while improving the reliability of the initial collection phase.

III. Proposed Method

Fig. 2. shows the process flow of the IOTA Tangle-based digital evidence anchoring system proposed in this paper.

A. *Distributed Anchoring*

Distributed anchoring is a mechanism that enables integrity verification under a unified standard anywhere in the world by recording the hash value and key metadata of digital evidence collected or generated in the field onto IOTA Tangle. When an investigator creates or saves an evidence file on a device, the system first computes the file’s hash value (e.g., SHA-256) along with minimal metadata such as the case ID, file type, and time of collection. It then packages this information into a single transaction payload and submits it to a Tangle node; once the transaction is confirmed by the network, the resulting transaction ID is stored in a separate database mapped to the internal case and evidence identifiers.

Due to the DAG-based structure capable of parallel processing, the anchoring process can minimize system bottlenecks even when a large number of evidence items are recorded simultaneously. Each transaction contributes to the network’s consensus by referencing other transactions, thereby ensuring tamper resistance and ledger stability without the need for dedicated miners. Most importantly, the information recorded in the Tangle is limited to the level of "hash value + minimum metadata," and the content and personal information of the original evidence file are not directly exposed on-chain. This can satisfy the advantages of distributed verification and personal information protection needs at the same time.

B. *Digital Evidence Encryption*

When an investigator captures video or collects log/image files in the field, the client application on the device first stores the file temporarily and then immediately invokes the encryption procedure. In the encryption procedure, a unique session symmetric key (e.g., a 256 bit key) is generated for each evidence unit, and the entire file is encrypted with an authentication encryption algorithm such as AES-256-GCM. At the same time, metadata such as the file hash (e.g., SHA-256), case ID, time of collection, and file type (MIME type) is computed and assembled together with the encrypted file into a single “evidence record,” which is then stored in a local queue on the device. Once this step is complete, the plaintext file is either deleted from the device or processed using OS security APIs in a way that makes recovery impossible, so that plaintext does not remain on the device storage for an extended period.

The evidence records stored in the local queue are kept safely regardless of network conditions, and the device periodically checks the status of the network connection. When a secure connection protected by VPN, dedicated network, TLS, etc. is secured, the items accumulated in the queue are transmitted to the server sequentially, and at the same time, anchoring to the IOTA Tangle using the hash and metadata of each evidence is performed. If the upload is successful, the server storage location, upload time, and Tangle transaction ID are updated and recorded in the record. On the server side, the received encrypted files and metadata are stored separately, and different encryption keys and access control policies are applied to the file storage and metadata database, thereby reducing the risk that a compromise of a single storage component will expose the entire evidence chain at once. When an investigator or other authorized user needs to view a piece of evidence, the server first verifies access permissions, then selectively decrypts the file using the corresponding session key, and the hash value and Tangle transaction information are used in conjunction with the integrity verification procedure.

Such On-Device-oriented encryption and queue-based transmission structures can maintain the confidentiality and integrity of digital evidence throughout the collection, storage, and transmission process even in an international cooperative investigation environment where communication infrastructure is unstable, and can implement an encryption pipeline that is naturally linked to anchoring and verification procedures.

C. *Verifying Data Integrity*

The verifying data integrity process is a standardized procedure for confirming that stored digital evidence has remained in its original state since the time of collection. When a verification request is triggered, the system first accesses the corresponding encrypted evidence file within a protected execution environment and recomputes the current hash value using the same algorithm that was applied at the time of collection. It then compares this current hash with the original hash stored in the internal metadata database and with the hash value retrieved from the transaction anchored on IOTA Tangle. In other words, it checks whether the relationship “Current File Hash = Metadata Hash = Anchored Tangle Hash” holds. If all three values match, the evidence is deemed to have remained unchanged throughout collection, storage, transmission, and sharing, and is recorded as being in an “Integrity Preserved (Verification Succeed)” state.

Conversely, if the Current Hash does not match the Metadata Hash, if the Metadata Hash differs from the Anchored Tangle hash, the process classifies the evidence as being in an “Integrity Violated (Verification Failed)” state, respectively. In such cases, the system records detailed information — including case ID, evidence ID, verification timestamp, and failure type, etc. — in audit logs and dashboards so that the results can be used for subsequent auditing, forensic analysis, and administrative review. By standardizing and automating the sequence of steps spanning collection, encryption and anchoring, verification request, hash re-computation, metadata and Tangle comparison, status determination, and log recording, the system enables integrity verification to be carried out under a consistent set of criteria across countries and agencies, while minimizing human error and procedural inconsistency inherent in manual verification processes.

IV. Conclusion

In this paper, we proposed digital evidence anchoring system that combines IOTA Tangle–based distributed ledger technology with on-device encryption. The proposed system aims to guarantee the integrity and confidentiality of digital evidence collected by investigators in the field throughout its entire lifecycle—from the moment of creation to the point of verification—and to provide an evidence management infrastructure that enables cross-verifiable integrity checks among countries and agencies in international cooperative investigations. To this end, we designed an architecture that integrates distributed anchoring, on-device encryption with local queue–based temporary storage, and a standardized integrity verification protocol.

The proposed method provides strong legal reliability by irreversibly recording the hash value and metadata of evidence at the time of creation on Tangle, thereby allowing tampering to be demonstrated through identical hash comparison at any time and from any location. In addition, by adopting a common verification mechanism based on hashes and anchor IDs while allowing each country or agency to maintain its own internal systems, the system enhances interoperability and thus improves the efficiency and consistency of international cooperative investigations. Finally, by encrypting and storing sensitive original data separately and recording only minimal information on Tangle, the method offers a practical balance between personal data protection and investigative transparency.

As future work, we plan to conduct pilot deployments of the proposed system in real-world copyright infringement and cybercrime investigations to evaluate its performance and robustness under various network conditions and operational scenarios, and to refine its functionality based on feedback from investigative agencies.

Acknowledgment

This research was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2026. (Project Name: Development of International Investigative Assistance co-operation system technology to protect K content, Project Number: RS-2024-00439553, Contribution Rate: 100%)

References

[1] Kreso, Inda. "Using blockchain technology for preserving digital evidence in digital forensics." *KNOWLEDGE-International Journal* 68.3 (2025): 331-336.

[2] Xiao, Nan, et al. "A novel blockchain-based digital forensics framework for preserving evidence and enabling investigation in industrial Internet of Things." *Alexandria Engineering Journal* 86 (2024): 631-643.

[3] Popov, Serguei. "The tangle." *White paper* 1.3 (2018): 30.

[4] Casey, Eoghan. *Digital evidence and computer crime: Forensic science, computers, and the internet*. Academic press, 2011.

[5] Groß, Tobias, Matanat Ahmadova, and Tilo Müller. "Analyzing android's file-based encryption: information leakage through unencrypted metadata." *Proceedings of the 14th International Conference on Availability, Reliability and Security*. 2019.

[6] Wang, Xubin, et al. "Empowering edge intelligence: A comprehensive survey on on-device ai models." *ACM Computing Surveys* 57.9 (2025): 1-39.

YOLO-PT: One-Shot Framework for 3D-based Progress Tracking

1st Lan Anh Nguyen *, 2nd Jaemo Sung †, 3rd Eun-Sung Jung *
* *Department of Software and Communications Engineering, Hongik University, Sejong, South Korea*
† *DesignedAI Inc., Seoul, South Korea*
Emails: ejung@hongik.ac.kr

***Abstract*—We introduce *YOLO-PT*, a unified one-shot framework for object progress tracking. Built on the Ultralytics implementation of YOLO, *YOLO-PT* incorporates the 3D-based progress estimation pipeline into a single architecture. The framework adds two new heads to YOLO: (1) a 3D alignment head, which projects 3D object geometry into 2D representations that match the image perspective, and (2) a progress comparison head, which infers completion levels by comparing the projected 3D geometry with the observed image. Through the cooperation of these alignment and comparison components, *YOLO-PT* enables end-to-end progress estimation within a single inference pass. Experiments on Pix3D show that *YOLO-PT* delivers accurate and efficient progress tracking across multiple YOLO versions. Overall, the framework provides a one-shot solution for object progress estimation based on integrated 3D–2D alignment and comparison.**

***Index Terms*—YOLO, 3D progress tracking, 2D–3D alignment, segmentation, mask comparison**

I. Introduction

Accurate progress tracking has become increasingly essential for digital construction, robotics inspection, and industrial automation, enabling continuous monitoring of object states, spatial alignment, and scene-level completeness [1], [2]. Recent advances in computer vision have shifted progress tracking from manual observation to data-driven approaches that leverage 2D imagery, depth sensing, and 3D model integration [2], [3]. However, most existing systems rely on task-specific models for detection, segmentation, alignment, and comparison, resulting in redundant computation and complex implementation.

Vision based progress tracking nowadays typically relies on deep learning models for tasks such as alignment, detection, and segmentation to estimate the completion. The YOLO family, with its unified one-stage architecture, high inference efficiency, and modular task heads [4], [5], offers an attractive foundation for building a one-shot progress tracking framework. Its shared feature representation and flexible heads for detection, segmentation, classification can reduce system complexity while maintaining accuracy and real-time performance. However, existing YOLO models operate solely on 2D inputs, limiting their ability to exploit the richer geometric information present in 3D or 4D design data. This gap highlights the need to extend YOLO architectures toward higher dimensional perception to better support vision based progress tracking.

In this work, we introduce a one-shot framework (e.g., *YOLO-PT*) that extends YOLOs to support 3D based progress tracking. Our contributions are as follows:

- we augment the YOLO architecture with alignment and comparison heads, allowing the model to perform 3D based progress tracking.
- Second, to ensure reliable comparison performance, we introduce a multi-level assessment between the design and the observed image, operating at the pixel, edge, and feature levels.
- Next, we develop a new dataset for progress estimation by synthesizing objects from Pix3D into multiple completeness levels, ranging from partially incomplete to fully complete.
- Finally, we evaluate our framework by applying *YOLO-PT* to YOLOv8, YOLOv10, and YOLOv11 and measuring their progress tracking accuracy and efficiency.

Section II reviews prior research on vision-based progress tracking and discusses the extensibility of YOLO models. Section **??** outlines the preliminaries, including the fundamentals of vision-based progress tracking systems, YOLO architectures for detection and segmentation, and the datasets used. Section III presents our proposed *YOLO-PT* framework. Section IV reports the experimental setup, performance evaluation, and efficiency analysis. Finally, we conclude the paper with a summary.

II. Related Works

A. Vision-Based Progress Tracking

Previous studies on vision based progress tracking primarily relied on 2D images to perform object segmentation and estimate progress [6]. However, approaches based solely on 2D data fail to leverage the richer geometric information present in 3D/4D design models, including shape, texture, and spatial structure. To address these limitations, recent works have incorporated geometric cues from 3D and 4D representations, such as those available in 3D CAD and 3D/4D BIM models, to preserve design and improve progress assessment [7].

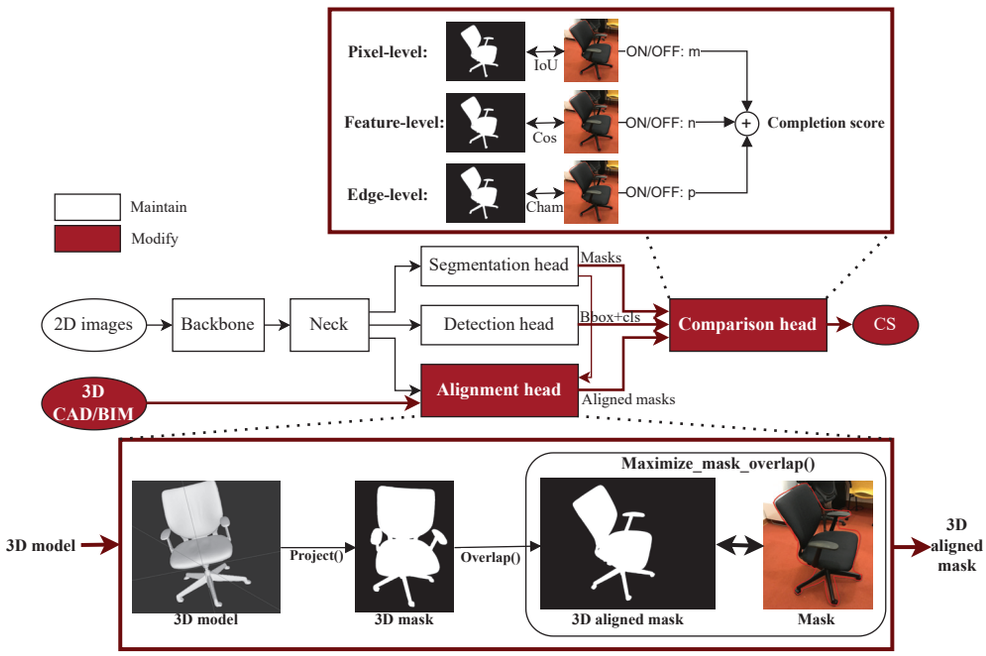


Fig. 1. YOLO-OPT: One-shot framework for 3D geometry-based object progress tracking

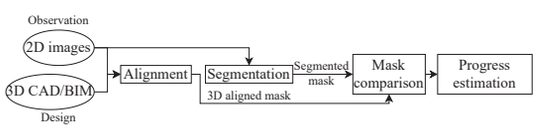


Fig. 2. Conventional vision based progress tracking

Conceptually, vision-based progress tracking compares visual data design information from CAD/BIM and actual site imagery to determine the progress of each object. In essence, it evaluates object states and the spatial consistency between designed and observed components. Conventional systems decompose this workflow into multiple modules [8], such as 3D/4D projection, image–model alignment, segmentation, and progress estimation. To support these tasks, various learning based vision algorithms, including Mask R-CNN [9], and SAM [10], have been widely adopted for segmentation and progress estimation. However, multiple modules and diverse methods used in modules introduce a complexity of implementation. Such complexity can increase latency, which is often critical for real-time or on-site applications. These limitations highlight the need for an one-shot or unified approach to 3D vision based progress tracking.

B. YOLOs

The You Only Look Once (YOLO) framework [11] was first introduced in 2015 and became popular in real-time object detection due to its single-stage architecture, high speed, and competitive accuracy. Since then, a wide range of YOLO variants have emerged, expanding the original framework into a unified suite capable of handling diverse vision tasks, such as object detection, instance segmentation, classification, and pose estimation within a real-time pipeline [5]. Recent models, including YOLOv8 [12], YOLOv10 [13], and YOLOv11 [14], further improve accuracy, computational efficiency, and training stability, as demonstrated in empirical studies [14], [15].

YOLO models offer high efficiency and accuracy, making them well suited for vision-based progress tracking, which requires low-latency detection, segmentation, and estimation. However, current YOLO architectures operate only on 2D inputs, limiting their ability to leverage the richer geometry of 3D/4D design data used in progress tracking. Extending YOLO to support 3D/4D inputs is therefore a promising direction for achieving more robust and accurate progress estimation.

III. YOLO-PT

In this section, we provide background on conventional vision-based progress tracking with conventional system and standard dataset (e.g., Pix3D). Next, we describe our proposed method, *YOLO-PT*.

A. Conventional vision-based progress tracking

Figure 2 illustrates the system for vision-based progress tracking nowadays. It has emerged as a practical solution for monitoring construction, manufacturing by leveraging cameras and 3D CAD/BIM models. Unlike traditional methods relying on manual inspection, vision based approaches provide continuous and automated monitoring of object placement and completion levels. Typical systems capture visual data (e.g., observed photo) and compare it against its 3D model (e.g., 3D design) to estimate the progress in

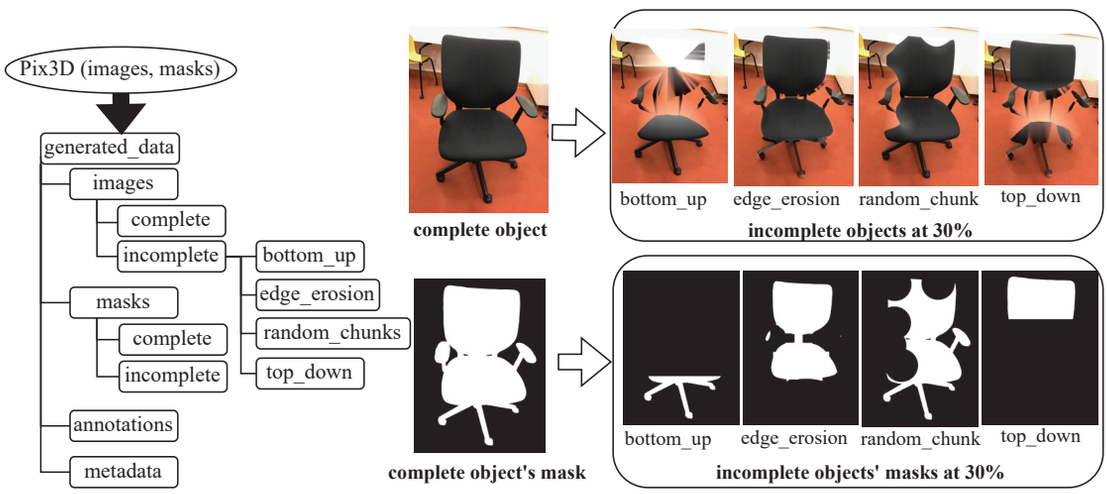


Fig. 3. Dataset generation

the observed photo. At the system level, vision based progress tracking generally consist of following stages, as shown in Figure 2:

- Input: Captures 2D images with target objects as their observations and collects 3D models (i.e., CAD, BIM, etc.) as designs of objects.
- Alignment: Performs 2D–3D alignment to ensure the both observation (e.g., 2D image) and design (e.g., 3D CAD/BIM) aligned with each other to output aligned 3D mask.
- Segmentation: Segments instance in the 2D image to output its segmented mask.
- Mask comparison: Compares two masks (e.g., 3D aligned mask and segmented mask) in previous stages.
- Progress estimation: Based on the mask comparison, this stage estimates the completion level of the object in image.

B. Dataset: Pix3D

Datasets for 2D–3D alignment commonly lack pixel-level correspondences, accurate camera models, and paired 3D CAD/BIM representations, which limit their use in tasks requiring precise geometric grounding. The Pix3D dataset [16] addresses these limitations by providing pixel-level alignment between images and 3D models. It contains about 395 3D models and 10,069 corresponding images across multiple furniture categories. Since the dataset already supplies aligned 2D–3D pairs, solving vision-based progress tracking becomes more straightforward as it mainly involves 2D–3D alignment.

C. Proposed Framework

We introduce *YOLO-PT*, a one-shot progress tracking framework built upon YOLO. Figure 1 illustrates the overall architecture, where the white modules represent standard YOLO components and the red modules denote our newly added contributions.

Originally, YOLO models accept only 2D images as input for tasks such as classification, detection, and segmentation. To overcome the limitations inherent in conventional approaches, we introduce 3D models as an additional input modality. Using 3D representations (e.g., CAD, BIM) preserves the full geometric characteristics of the object.

The alignment head is firstly introduced, bringing 3D and 2D objects into a unified spatial frame. It projects and aligns the 3D model onto the 2D image plane to generates aligned mask. Aligned mask is subsequently passed to the comparison head, where it serves as object's design for comparison with the observed mask.

Comparison head takes segmented mask from segmentation head and aligned mask from alignment head to estimate the completeness of the object in the observed 2D image. 'Completion score' (CS) indicates the level of object completeness. The score is calculated as weighted sum of three levels of comparison between segmented mask and aligned mask, as the following equation:

$$CS = \alpha.m + \beta.n + \gamma.p \tag{1}$$

Where: $\alpha + \beta + \gamma = 1$. And, α, β, γ work as weights for corresponding scores at different levels.

- m: Pixel-level comparison score. The score is obtained by using IoU as the distance metric.
- n: Feature-level comparison score. The score is obtained by using Cosine as the distance metric.
- p: Edge-level comparison. The score is obtained by using Chamfer as the distance metric.

For alignment, the algorithms 1 explains its primary steps.

D. Dataset generation

We generate incomplete object images from Pix3D using a segmentation-based progressive removal ap-

Algorithm 1 Alignment

Inputs: 3D model, 2D image, camera intrinsics K, initial rotation matrix R_0 and translation vector T_0, iteration number T
Outputs: Aligned 3D mask, rotation matrix R, translation vector T
$R \leftarrow R_0$; ▷ Initialize rotation matrix
$T \leftarrow T_0$; ▷ Initialize translation vector
$2D_mask \leftarrow YOLOseg(\text{2D image})$; ▷ YOLO outputs object's mask
$3D_mask \leftarrow Project(\text{3D model}, R, T, K)$; ▷ Get initial 3D mask
$3D_2D_olap \leftarrow Overlap(3D_mask, 2D_mask)$; ▷ Get initial overlap
while $t \leq T$ **do**
    $R_t, T_t \leftarrow RefinePose(R, T, K, 2D_mask, 3D_mask)$;
    $3D_mask_t \leftarrow Project(\text{3D model}, R, T, K)$;
    $3D_2D_olap_t \leftarrow Overlap(3D_mask, 2D_mask)$;
    if $3D_2D_olap_t \geq 3D_2D_olap$ **then**
        $3D_2D_olap \leftarrow 3D_2D_olap_t$;
        $R \leftarrow R_t$;
        $T \leftarrow T_t$;
        $3D_mask \leftarrow 3D_mask_t$;
    end if
    $t++$;
end while
Return $3D_mask, R, T$

proach. Using Pix3D's segmentation masks, we isolate each object and remove regions according to predefined completion ratios, followed by inpainting to maintain visual consistency. We design four removal patterns to simulate realistic incompleteness: (1) bottom-up, (2) top-down, (3) random chunk removal, and (4) edge erosion. This process produces controlled incomplete-object datasets with precise completion-level annotations for training and evaluating progress-tracking framework.

We generate the dataset from 10 original Pix3D photos, producing incomplete images using the four removal approaches described earlier. For each approach, we create four completion levels (90%, 70%, 50%, 30%), resulting in 160 generated images in total. We also preserve masks for both original and generated objects, store 3D model metadata in a metadata file, and record completion labels and removal type information in an annotation file. The data generation is illustrated in Figure 3.

E. Evaluation Metrics

1) Completion Estimation Accuracy: To quantitatively assess the accuracy of the estimated completion relative to the ground-truth annotations, we employ the Mean Absolute Error (MAE). It is computed as belows:

$$MAE = |\hat{CS} - L| \tag{2}$$

Where: $\hat{CS}$ refers to the completion score output by *YOLO-PT* when provided with an observation image and its associated 3D CAD design, and L corresponds to the ground-truth completion level of the object in the observation image.

2) Efficiency: Beyond accuracy, we evaluate the computational efficiency of our proposal through inference time. Inference time measures the average time needed to produce an output from a single input, highlighting the model's practical runtime performance. Together, these metrics demonstrate the suitability of our approach for real-time applications.

TABLE I
EXPERIMENTAL SETUPS

Setups	Values
#CPUs	16
#RAM	16G
Python version	3.11
YOLO versions	YOLOv8/v10/v11
#original images	10 (Chairs)
#Generated images	160
Removal types	bottom_up, edge_erosion, random_chunk, top_down
Removal levels	30%, 50%, 70%, 90%
Comparison levels	Pixel, edge, feature

IV. EXPERIMENTS AND RESULTS

A. Experimental setups

Experiments are conducted on a desktop machine equipped with 16 CPU cores and 16 GB of RAM. The proposed *YOLO-PT* framework is built upon YOLOv8, YOLOv10, and YOLOv11, running on Python 3.11. A total of 10 chair images are used to synthesize 160 samples containing various incomplete objects. To generate these incomplete objects, four removal strategies are applied: bottom-up, edge erosion, random chunk, and top-down. Each strategy is executed at four removal levels: 30%, 50%, 70%, and 90%.

In the comparison head of *YOLO-PT*, we implement three comparison mechanisms, pixel-level, edge-level, and feature-level to compare their performance. For evaluation, we consider two categories of metrics: (1) Completion estimation accuracy, using measures such as MAE, and (2) Model efficiency, using execution time. Table I provides a summary of all experimental configurations.

B. Results and analysis

*1) **Accuracy of completion estimation**:* First, we evaluate how accurately *YOLO-PT* estimates completion on the set of 160 synthesized images containing various incomplete objects. Figures 4, 5, 6, and 7 show the observed masks, aligned 3D masks, and corresponding estimations, with yellow indicating their overlap. Figure 8 presents the gap between the average predicted and ground-truth completion levels, expressed as percentages.

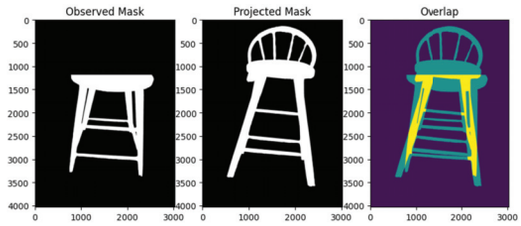


Fig. 4. Mask overlapping in case of bottom-up removal

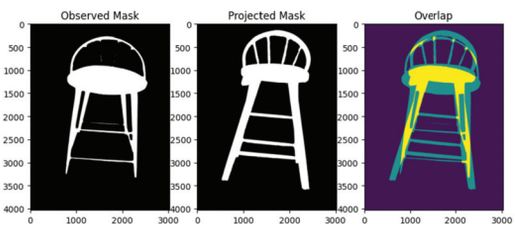


Fig. 5. Mask overlapping in case of edge erosion

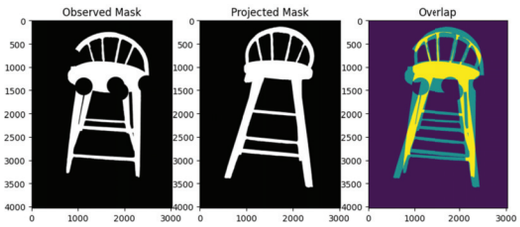


Fig. 6. Mask overlapping in case of random chunks

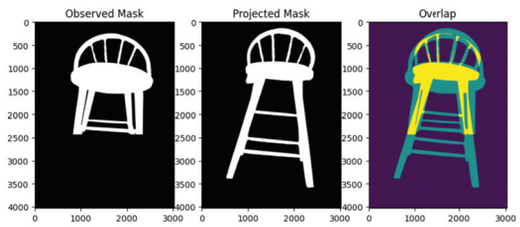


Fig. 7. Mask overlapping in case of top-down

The figure compares the performance of *YOLO-PT* built upon three different YOLOs, such as YOLOv8, YOLOv10, and YOLOv11, denoted as YOLOv8-pt, YOLOv10-pt, and YOLOv11-pt. For each *YOLO-PT*, it is evaluated with three comparison levels: pixel-level, edge-level, and feature-level. A smaller percentage difference indicates a more accurate completion estimation. Among the three variants, YOLOv8-pt consistently delivers the best performance across all comparison levels. Additionally, the pixel-level comparison head generally achieves higher accuracy than the edge-level and feature-level counterparts.

*2) **Impact of removal types**:* Figure 9 presents the completion-estimation performance of *YOLO-PT* across the four removal types. First, the figure shows that YOLOv8-pt and YOLOv10-pt outperform YOLOv11-pt, consistently keeping the prediction–ground-truth gap within approximately 3–4%. Second, we observe that the random chunk removal appears particularly challenging for preserving object details and distinctive features during segmentation. Consequently, all *YOLO-PT* variants exhibit their lowest estimation accuracy on this removal type.

*3) **Execution time of YOLO-PT**:* To assess the real-time suitability of our proposed method, we evaluate both inference time and model size. Figure 10 presents a comparison of inference latency between the *YOLO-PT* variants (e.g., YOLO-PT-pixel, YOLO-PT-edge, YOLO-PT-feature) and the original YOLO models (e.g., YOLOv8, YOLOv10, YOLOv11).

From the figure, we can firstly observe that both the baseline YOLO models and their YOLO-PT counterparts in YOLOv8 and YOLOv10 consistently achieve lower latency compared to those of YOLOv11. This aligns with earlier findings that YOLOv11 generally incurs higher inference time due to its more complex architecture; consequently, its YOLO-PT variants also exhibit longer execution times.

Focusing on the YOLO-PT variants (highlighted in green in Figure 10), YOLO-PT-feature consistently shows the lowest latency across all YOLO versions. This is expected, as YOLO-PT-feature relies on feature extraction from masks, a process that requires fewer computational operations compared to the pixel based and edge based variants. Therefore, it naturally achieves faster execution than the other two YOLO-PT methods.

To conclude, our evaluation demonstrates that: (1) *YOLO-PT* built on YOLOv8 maintains consistently high accuracy in progress estimation; (2) Among the comparison strategies, the YOLO-PT-pixel variant achieves the best overall performance; (3) Random chunk removal remains the most challenging removal pattern to detect; and (4) *YOLO-PT* implemented with YOLOv8 and YOLOv10 delivers fast inference times.

V. Conclusion

In this work, we present *YOLO-PT*, an extended YOLO framework tailored for 3D-based progress tracking. The proposed approach delivers competitive

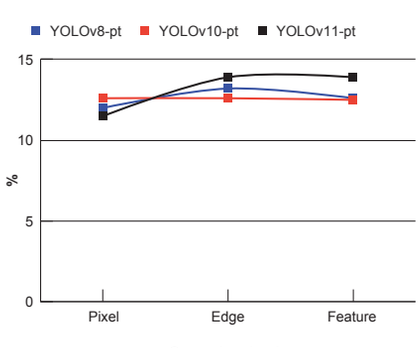


Fig. 8. Accuracy of progress estimation

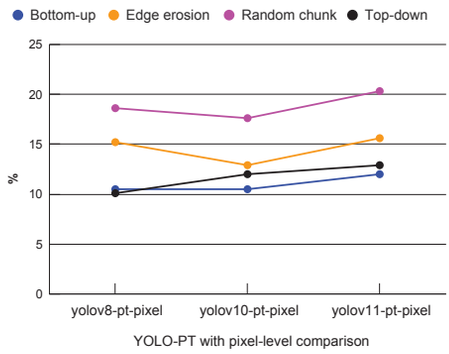


Fig. 9. YOLO-PT-pixel across removal types

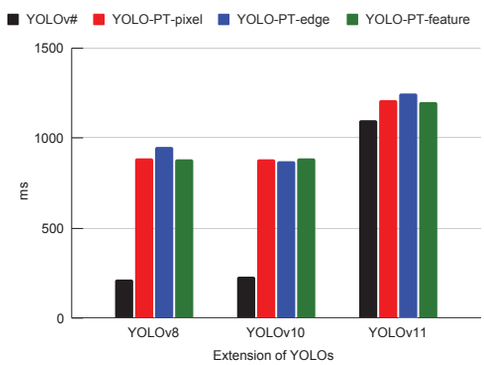


Fig. 10. Inference time of *YOLO-PT* variants vs original YOLO

accuracy and efficiency suitable for real-time deployment. Our experiments indicate that YOLOv8-PT provides more stable and superior performance in both progress estimation and reference timing compared to other variants. For future work, we plan to refine the comparison head using a regression-based formulation and release the extended framework publicly.

VI. Acknowledgment

This work was supported in part by the BK21 FOUR (Fostering Outstanding Universities for Research) program funded by the Ministry of Education (MOE) and the National Research Foundation (NRF) of Korea, and in part by the NRF grant funded by the Ministry of Science and ICT (MSIT) (No. 2022R1F1A1073976).

References

[1] P. Martinez, B. Barkokebas, F. Hamzeh, M. Al-Hussein, and R. Ahmad, "A vision-based approach for automatic progress tracking of floor paneling in offsite construction facilities," *Automation in Construction*, vol. 125, p. 103620, 2021.

[2] F. Alsakka, S. Assaf, I. El-Chami, and M. Al-Hussein, "Computer vision applications in offsite construction," *Automation in Construction*, vol. 154, p. 104980, 2023.

[3] A. Pal and S.-H. Hsieh, "Deep-learning-based visual data analytics for smart construction management," *Automation in Construction*, vol. 131, p. 103892, 2021.

[4] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, "A review of yolo algorithm developments," *Procedia computer science*, vol. 199, pp. 1066–1073, 2022.

[5] "Ultralytics yolo docs," https://docs.ultralytics.com/, accessed: 2025-11-15.

[6] B. Ekanayake, J. K. W. Wong, A. A. F. Fini, P. Smith, and V. Thengane, "Deep learning-based computer vision in project management: Automating indoor construction progress monitoring," *Project Leadership and Society*, vol. 5, p. 100149, 2024.

[7] B. Ekanayake, J. K.-W. Wong, A. A. F. Fini, and P. Smith, "Computer vision-based interior construction progress monitoring: A literature review and future research directions," *Automation in construction*, vol. 127, p. 103705, 2021.

[8] W. Wei, Y. Lu, T. Zhong, P. Li, and B. Liu, "Integrated vision-based automated progress monitoring of indoor construction using mask region-based convolutional neural networks and bim," *Automation in Construction*, vol. 140, p. 104327, 2022.

[9] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[10] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.

[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[12] R. Varghese and M. Sambath, "Yolov8: A novel object detection algorithm with enhanced performance and robustness," in *2024 International conference on advances in data engineering and intelligent computing systems (ADICS)*. IEEE, 2024, pp. 1–6.

[13] "Yolov10," https://docs.ultralytics.com/models/yolov10/, accessed: 2025-11-15.

[14] "Yolov11," https://docs.ultralytics.com/models/yolo11/, accessed: 2025-11-15.

[15] L. A. Nguyen, M. D. Tran, and Y. Son, "Empirical evaluation and analysis of yolo models in smart transportation," *AI (Switzerland)*, vol. 5, no. 4, pp. 2518–2537, 2024.

[16] X. Sun, J. Wu, X. Zhang, Z. Zhang, C. Zhang, T. Xue, J. B. Tenenbaum, and W. T. Freeman, "Pix3d: Dataset and methods for single-image 3d shape modeling," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2974–2983.

Design of a ContractNLI-Based Intermediate-Task Trained Natural Language Inference Model for Game and Software License Violation Detection

Dong-Wan Kim[*1], Chang-Jun Choi[*2], YongJoon Joe[*3] and †Dong-Myung Shin[*4]

* *R&D Division, LSware Inc, Seoul, Republic of Korea*

Emails: *{ kdw5339, jacob, eugene, roland } @lsware.com*

***Abstract*— Recent advances in legal NLP have enabled automatic extraction of clauses and identification of issues from complex contracts, yet the determination of license violations still relies heavily on manual review. In particular, game license agreements and software license clauses simultaneously encode multi-dimensional attributes such as territory, platform, distribution channel, sublicensing, and revenue sharing. As a result, generic language models struggle to consistently judge the legality of concrete usage scenarios. This paper proposes a license violation detection framework based on contract NLI, which formulates the relationship between license texts and usage scenarios as an NLI problem. The proposed framework consists of two core components: (i) intermediate-task training on contract NLI using ContractNLI, and (ii) the construction of a scenario-based NLI dataset by selecting license-related clauses from CUAD and pairing them with realistic usage scenarios. On this basis, we present an experimental design that enables a controlled comparison between BERT-family models and GPT-OSS models under a unified setup, and we discuss the potential and limitations of contract-domain intermediate transfer learning strategies. We also describe the design of a research-oriented NLI model that evaluates potential violations at the clause level in game and software licensing environments, thereby laying the groundwork for future support tools and quantitative performance evaluation.**

Keywords—contract natural language inference, license agreement, software license, intermediate-task training, transfer learning

I. INTRODUCTION

In the software and digital content industries, license agreements are drafted with overlapping conditions such as territory, platform, distribution channel, sublicensing, and revenue sharing, and related clauses are dispersed throughout the contract document. This makes it difficult for reviewers to consistently determine whether a concrete business scenario violates the contract conditions, and the outcome heavily depends on the individual reviewer's experience and interpretation [1]. In game service environments, where country-specific regulations and platform policies are additionally intertwined, the same scenario may be repeatedly re-evaluated and overturned within an organization.

Research on clause extraction and risky clause identification using Large Language Models (LLMs) and legal-domain language models has been actively conducted, but most studies remain at the level of determining the existence of specific clauses or classifying their risk levels. For game licenses and software licenses, in which the scope of rights is defined along multiple dimensions such as jurisdiction, term, media, user category, and billing model, a Natural Language Inference (NLI) structure that accepts a concrete usage scenario as input and classifies its relationship with the entire contract into violation, non-violation, and indeterminate has not yet been sufficiently established. Consequently, when off-the-shelf language models are directly applied, it is difficult to provide clause-level, consistent justifications for the legality of a given scenario [2].

To address these limitations, this paper proposes a contract natural language inference (contract NLI)-based license violation detection framework that reformulates the relationship between license agreements and usage scenarios as an NLI problem. The proposed framework consists of three components: (i) intermediate-task training on ContractNLI, a contract NLI dataset, to learn entailment, contradiction, and neutral relations between contract texts and hypotheses; (ii) construction of a scenario-based license NLI dataset by selecting license-related clauses from the Contract Understanding Atticus Dataset (CUAD) and pairing them with realistic usage scenarios; and (iii) transfer learning and comparative evaluation of Bidirectional Encoder Representations from Transformers (BERT) and GPT-based open-source models (GPT-OSS). Here, intermediate-task training denotes a transfer learning strategy in which a model is first trained on an intermediate task closely related to the final task, and the resulting parameters are used as the initialization for training on the final task. This enables a systematic evaluation of contract-domain transfer learning strategies and the design of NLI models suited for clause- and scenario-level violation judgment.

II. RELATED WORK

A. Sentence-Pair Classification Based on Natural Language Inference

NLI is defined as the task of classifying the relationship between a premise sentence and a hypothesis sentence into discrete categories such as entailment, contradiction, and neutral. Early work focused on constructing large-scale benchmark datasets for general-domain sentences and training sentence encoder-based classifiers on them. With the emergence of transformer architectures, NLI performance was further improved by combining pre-trained language models such as BERT with sentence-pair classification heads.

More recently, NLI has been used not only as a standalone benchmark task but also as an intermediate representation learning task that benefits multiple downstream tasks. For example, for document classification, question answering, and reasoning-based QA tasks that require modeling sentence

relationships, intermediate-task training on general-domain NLI can be performed first, and the resulting parameters are then used as initialization for the domain-specific task [3]. This approach augments the general language understanding ability of pre-trained language models with additional capacity to represent sentence-pair relations, and has been reported to improve both performance and training stability on downstream domain-specific tasks.

In the legal domain, the NLI paradigm has been adapted to classify relations among various legal texts-such as case summaries, clause descriptions, and query sentences-into entailment, contradiction, and neutral. These studies have been extended to applications such as contradiction detection between clauses, consistency checking between factual descriptions and judgments, and validation of legal QA outputs. However, research that directly aligns concrete usage scenarios with specific clauses to determine license violations remains limited. Building on prior NLI-based sentence-pair classification research, this paper formulates the relationship between license clauses and usage scenarios in NLI form and proposes a framework that combines contract-domain intermediate-task training with such NLI formulations..

B. Contract and License-Related Datasets

In legal NLP, a variety of datasets have been proposed for contracts, court decisions, and statutes. Contract-focused studies typically construct datasets that couple contract texts with clause types, risk labels, or key clause annotations, supporting tasks such as clause type classification, risky clause detection, and key clause extraction. These datasets are widely used for contract classification, highlighting of risky provisions, and prioritization of clauses for review, but datasets that explicitly connect clauses with concrete usage scenarios in a sentence-pair format are relatively rare.

For contract NLI, a representative dataset is ContractNLI, which consists of contract segments paired with hypotheses and annotated with entailment, contradiction, or neutral labels. ContractNLI formalizes relationships between contract sentences as an NLI task and provides a basis for training and evaluating contract-domain inference models [4]. In this work, we use ContractNLI to perform intermediate-task training that teaches basic patterns of entailment, contradiction, and neutrality over contract texts.

On the other hand, CUAD is a large expert-annotated dataset constructed from commercial contracts to support extraction of risky and key clauses, and provides clause spans, clause categories, and associated questions [5]. CUAD has mainly been used for clause extraction, clause classification, and risky clause detection. However, by treating clause texts as premises and questions or usage scenarios as hypotheses, CUAD can be repurposed as an NLI-style dataset. In this paper, we select items directly related to licensing-such as license scope, usage restrictions, sublicensing, and revenue sharing-from CUAD and design a license scenario NLI dataset in which contract clauses are paired with realistic usage scenarios.

Although existing datasets support contract-domain NLI and contract clause understanding separately, a framework that directly decides violations based on clause–scenario pairs in multi-dimensional game and software license environments has not been sufficiently established. This paper proposes a new utilization scheme that combines ContractNLI and CUAD to connect contract NLI-based intermediate-task training with the construction of a license scenario–specific dataset.

C. Pre-trained Language Models and Domain-Specific Models

In NLP, Pre-trained Language Models (PLMs) trained on large corpora of general text serve as a common backbone for diverse downstream tasks. BERT adopts a bidirectional transformer encoder architecture and learns general-purpose sentence representations via masked language modeling and next sentence prediction [6]. On top of BERT, simple classification heads can be attached for tasks such as sentence classification, sentence-pair classification, and token classification, yielding strong performance. Domain-specific variants of BERT have also been proposed by further pre-training on domain-specific corpora in areas such as law and medicine.

GPT-family models, in contrast, are autoregressive language models based on transformer decoders, pre-trained to predict the next token given prior context. This architecture offers advantages for generative tasks and, when combined with prompt design and fine-tuning, can handle classification, reasoning, summarization, and other tasks in a unified framework. In the open-source ecosystem, GPT-OSS have been released by continuing to train generic GPT-like models on public and domain-specific corpora, and various efforts have been made to adapt these models to specific domains using instruction tuning and parameter-efficient fine-tuning [7].

To compare NLI performance in the contract and license domain, this paper considers both the encoder-based BERT model and decoder-based GPT-OSS models. For each, we perform intermediate-task training on ContractNLI and then fine-tune on the license scenario NLI dataset, providing an experimental framework to analyze how pre-training architectures and intermediate-task strategies affect clause- and scenario-level violation detection.

III. PROPOSED FRAMEWORK

A. Overall Architecture

Fig. 1 illustrates the overall flow of the proposed contract NLI-based license violation detection framework. First, we perform contract NLI intermediate-task training using ContractNLI, which consists of contract segments and annotated hypotheses. In this stage, we train both a generic BERT-based model and a domain-adapted GPT-OSS model to classify entailment, contradiction, and neutral relations between contract texts and hypotheses, thereby endowing both model families with basic contract inference capabilities.

Next, we construct a license scenario NLI dataset by selecting clauses directly related to licensing from CUAD and designing concrete usage scenarios for each selected clause. Each sample in this dataset treats the license clause as the premise sentence and the usage scenario as the hypothesis sentence, and is labeled according to whether the scenario is permitted by the contract, explicitly prohibited or clearly outside the permitted scope, or indeterminate based on the contract text alone.

Finally, we fine-tune both the intermediate-task-trained models and the baseline pre-trained models (without intermediate-task training) on the same license scenario NLI task and compare their performance. For each clause–scenario

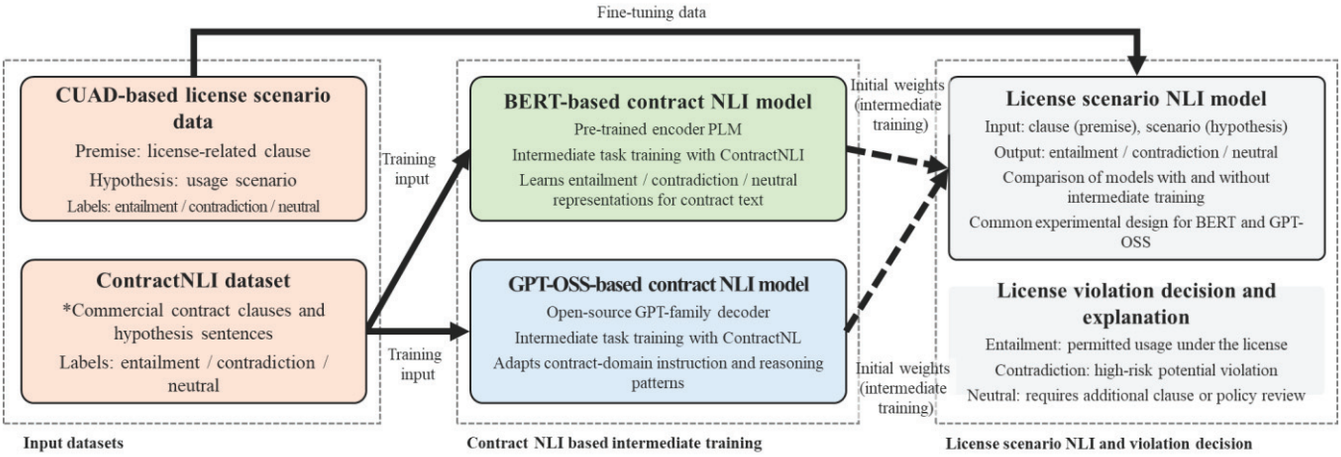


Fig. 1. Overview of the contract NLI-Based License Violation Detection Framework

pair, the model predicts one of the entailment, contradiction, or neutral labels, and we further analyze results at the clause level using clause identifiers included in the dataset. This framework enables a quantitative evaluation of contract-domain transfer learning and the design of NLI models that can later be extended into license violation support tools.

B. Dataset Construction

For contract NLI intermediate-task training, we use ContractNLI. Each sample contains:.

- **Premise**: A contract clause or a segment within a clause.
- **Hypothesis**: A natural language sentence that summarizes the meaning of the clause or queries whether certain conditions are met.
- **Label**: Entailment if the premise supports the hypothesis, contradiction if they are mutually incompatible, and neutral otherwise.

In this study, instead of using the entire ContractNLI dataset, we prioritize categories that are indirectly related to licensing-such as scope of rights, restrictions on use, term, termination, and assignment-considering their connection to the downstream license scenario NLI task. Detailed statistics on the number of samples and label distributions are presented in Section IV, which covers the experimental results.

In this study, instead of using the entire ContractNLI dataset, we prioritize categories that are indirectly related to licensing-such as scope of rights, restrictions on use, term, termination, and assignment-considering their connection to the downstream license scenario NLI task. Although detailed performance analysis is beyond the scope of this paper, the selection of these categories reflects their relevance to potential license violation patterns in realistic usage contexts.

Each sample is labeled as follows:

- **Entailment**: The scenario clearly falls within the permitted license scope defined by the clause.
- **Contradiction**: The scenario is explicitly prohibited by the clause or clearly exceeds the permitted license scope.
- **Neutral**: The legality of the scenario cannot be determined solely from the clause, or essential elements are delegated to other clauses or external regulations.

We also store metadata such as contract identifiers, clause numbers, clause categories, and scenario identifiers for each sample, enabling the model's predictions to be linked back to specific locations in the original contracts.

C. Contract NLI-Based Intermediate-Task Training

Contract NLI-based intermediate-task training is an application of intermediate-task transfer learning to contract-domain NLI. The goal is to have the model acquire contract-specific notions of entailment, contradiction, and neutrality prior to training on the license scenario task.

We initialize both the BERT-based model and the GPT-OSS model from their pre-trained weights and fine-tune them on ContractNLI in a sentence-pair classification setup. The input is the concatenation of the premise and hypothesis tokens, and the output is a probability distribution over the three labels. Cross-entropy loss is used for multi-class classification.

To ensure comparability, both model families are trained under the same hyperparameters and training schedules. After training, we retain the classification head and backbone weights tuned for contract NLI and use them as initialization for the subsequent license scenario NLI training. This encourages the final model to exploit general contract inference patterns when handling license- and scenario-specific decisions.

D. License Scenario NLI Training

In the license scenario NLI training stage, we use the scenario-based NLI dataset described in Section III-B to train models specialized for license violation detection. We compare two initialization settings:

- Direct fine-tuning: Initialize from the pre-trained BERT and GPT-OSS models and directly fine-tune on the license scenario NLI task.
- Intermediate-task fine-tuning: Initialize from the models that have undergone contract NLI intermediate-task training as described in Section III-C, and then fine-tune on the same license scenario NLI task.

For both settings, we use identical data splits, hyperparameters, and optimization procedures to isolate the effect of intermediate-task training. The model architecture remains a sentence-pair classifier; we replace the ContractNLI premises with CUAD-derived license clauses and the hypotheses with usage scenarios, while retaining the three-way label space (entailment, contradiction, neutral).

After training, we evaluate not only accuracy, precision, recall, and F1 score on a validation set, but also specific metrics related to license violation detection, such as the ability to distinguish contradiction from neutral cases. This enables a quantitative analysis of how contract NLI-based intermediate-task training affects violation judgment performance on license scenarios.

E. License Violation Decision and Explanation

The predicted NLI label for each clause–scenario pair can be interpreted as follows:

- **Entailment**: The scenario is treated as a permitted use under the clause.
- **Contradiction**: The scenario is considered a high-risk candidate for license violation under the clause.
- **Neutral**: The clause alone does not provide sufficient information to clearly determine legality, or additional clauses and external regulations must be consulted.

In this paper, our experiments focus on quantitatively evaluating the classification performance that underlies these interpretations. We do not implement service-level tooling or end-user interfaces; instead, we restrict our scope to performance comparison of clause–scenario NLI classifiers and analysis of transfer learning effects.

IV. CONCLUSION

This paper addressed the difficulty of consistently determining whether real service scenarios violate increasingly complex license agreements in the game and software industries. As conditions such as territory, platform, distribution channel, sublicensing, and revenue sharing are scattered throughout contracts, and as country-specific regulations and platform policies are intertwined, we confirmed the limitations of rule-based review and generic language models in achieving stable, clause-level license violation judgments.

Against this background, we reformulated the relationship between license agreements and usage scenarios as an NLI problem and proposed a contract NLI-based license violation detection framework that integrates ContractNLI-based intermediate-task training with a scenario-based NLI dataset derived from CUAD. The framework first trains BERT and GPT-OSS models on a contract-domain NLI task to learn entailment, contradiction, and neutral patterns over contract texts, and then performs transfer learning on a dataset composed of license-related clauses and realistic usage scenarios.

From the perspective of performance evaluation, we proposed an experimental design that compares three settings on the same license scenario NLI task: (i) baseline models with only generic pre-training, (ii) models trained directly on the CUAD-derived data, and (iii) models incorporating contract NLI intermediate-task training. Future work will implement this design to conduct quantitative evaluations, focusing in particular on the ability to distinguish contradiction labels (license violations) from neutral labels (cases requiring further judgment), thereby validating the effectiveness of the proposed framework. Furthermore, by leveraging clause-level metadata, we discussed the possibility of extending the model into a support tool that links predictions to specific clauses, allowing users to simultaneously inspect potential violations and the associated provisions.

In future research, we plan to go beyond CUAD-derived data and construct a large-scale dataset of real-world game and software license agreements and internal compliance cases. Using this extended dataset, we will compare the proposed framework with various transfer learning strategies under a unified experimental environment.

ACKNOWLEDGMENT

This research was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2026(Project Name: Global Talent Training Program for Copyright Management Technology in Game Contents, Project Number: RS-2024-00396709, Contribution Rate: 100%).

REFERENCES

[1] Holland & Knight, “Imprecise Software License Terms Leads to New York Litigation,” Digital Technology & E-Commerce Blog, Jan. 2019.

S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, “A large annotated corpus for learning natural language inference,” in Proc. of EMNLP, 2015, pp. 632–642.

[2] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, *“A large annotated corpus for learning natural language inference,”* in *Proc. of EMNLP*, 2015, pp. 632–642[4]

[3] Y. Pruksachatkun et al., “Intermediate-Task Transfer Learning with Pretrained Language Models: When and Why Does It Work?,” in Proc. of ACL, 2020, pp. 5231–5247.

[4] Y. Koreeda and C. D. Manning, “ContractNLI: A Dataset for Document-level Natural Language Inference for Contracts,” in Findings of ACL: EMNLP 2021, pp. 1907–1919.

[5] D. Hendrycks, C. Burns, A. Chen, and S. Ball, “CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review,” NeurIPS Datasets and Benchmarks Track, 2021

[6] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in Proc. of NAACL, 2019, pp. 4171–4186.

[7] S. Black et al., “GPT-NeoX-20B: An Open-Source Autoregressive Language Model,” arXiv:2204.06745, 2022

A Multi-Layered Framework for Evaluating Requirements Completeness in User-Story–Driven Agile Development

Yukyong Kim
Division of Advanced Engineering, Sookmyung Women's University, Seoul, South Korea
Emails: *ykim.be@sookmyung.ac.kr*

***Abstract*—User stories are widely used in agile development, yet industry reports consistently reveal issues such as missing functional flows, inconsistent dependencies, insufficient non-functional requirements, and weak testability. These limitations frequently result in backlogs that lack system-level completeness. This paper introduces US-CEP, a three-layer framework that evaluates (1) user story quality using an extended INVEST and Acceptance Criteria model, (2) structural completeness through functional coverage, user journey mapping, and dependency analysis, and (3) development and test readiness based on DoR, DoD, and risk-adjusted feasibility. A case study of an 18-story e-commerce backlog demonstrates the framework's effectiveness: despite reasonable story-level quality, US-CEP identified missing exception flows, dependency inconsistencies, and substantial NFR and readiness gaps, resulting in a completeness score of 63.2/100. The findings show that well-written user stories alone do not guarantee comprehensive or development-ready backlogs. US-CEP provides a rigorous yet lightweight mechanism for early detection of requirement defects and improved backlog preparation. This work offers the first quantitatively grounded completeness model tailored to user stories and establishes a basis for future automation using NLP and LLM-based techniques.**

Keywords—User Stories, agile requirements engineering, requirements completeness, backlog coverage

I. INTRODUCTION

The quality of requirements specifications plays a critical role in determining overall software completeness within software appraisal processes. Incomplete or ambiguous requirements propagate through design, implementation, and testing, ultimately making it difficult for evaluators to judge whether the developed system satisfies its intended functional scope, expected behavioral conditions, and operational constraints. At the same time, contemporary software development environments face increasing requirements volatility and heightened pressure for rapid market release. In response, agile methodologies have gained prominence by enabling more flexible requirements management through high adaptability and iterative feedback cycles compared to traditional waterfall models. Within this context, the user story has emerged as the primary requirements artifact in agile projects, offering a concise means of expressing user-centered goals. While this brevity facilitates communication and accelerates collaborative planning, the simplified grammar of user stories inherently limits their ability to ensure the structural rigor and completeness emphasized in traditional Requirements Engineering (RE). Consequently, despite their practical advantages, user stories often fail to capture the full breadth of functional behaviors, dependencies, and quality attributes necessary for comprehensive software completeness evaluation.

In practice, the limitations of user story–based requirements management are consistently reported across industry projects. According to the 2022 Standish Group CHAOS Report, approximately 41% of agile project failures are directly attributable to unclear or missing requirements [1]. Numerous industrial assessments further highlight recurring issues such as insufficiently testable user stories and the omission of non-functional requirements (NFRs), both of which undermine the reliability of agile practices. Reports from McKinsey and other consulting organizations similarly warn that overreliance on user stories—particularly in large-scale platform development—often leads to rework, delays, and architectural inconsistency. These observations collectively suggest that although user stories effectively capture user value, they alone are insufficient to guarantee system-level requirements completeness.

Across agile projects, several recurring deficiencies in requirements specification undermine the completeness and reliability of the resulting software artifacts. Missing functional flows and undocumented data dependencies disrupt end-to-end behavioral coherence, while the omission of critical user-journey stages produces structural gaps in functional coverage. Non-functional requirements—including performance, security, and logging—are often captured inconsistently at the user-story level, thereby elevating operational and compliance risks. Moreover, insufficiently defined acceptance criteria introduce ambiguity in test scope, increasing the likelihood of defect leakage and escalating verification costs. Finally, variability in the application of the Definition of Ready (DoR) and Definition of Done (DoD) across teams leads to uneven story preparedness, generating schedule delays, quality degradation, and avoidable rework. These problems indicate that the quality of individual user stories alone cannot ensure overall system-level completeness. Hence, a more structured and multi-layered requirements evaluation framework is needed.

Existing research reflects similar concerns. The widely used INVEST principles [2] provide criteria for evaluating the quality of individual user stories, but their scope is limited to sentence-level quality and does not address holistic completeness across the system. Likewise, the completeness concept defined in ISO/IEC/IEEE 29148 [3] presupposes document-driven requirements specifications and cannot be directly applied to agile contexts where user stories serve as the primary artifacts.

Therefore, in agile environments where user stories form the backbone of requirements management, a new evaluation mechanism is needed—one capable of integrating story-level quality, backlog-level structural coverage, and development/testing readiness (DoR/DoD). To prevent

industrially recurring issues such as requirements omission, structural inconsistency, and insufficient NFR specification, evaluation must operate across multiple layers: (1) assessing the intrinsic quality and testability of individual user stories, (2) examining the structural completeness and functional coverage of the overall product backlog, and (3) verifying the practical readiness of requirements for development and testing activities.

To address these needs, this study proposes US-CEP (User Story–based Completeness Evaluation Process)—a multi-layered framework that enables systematic and quantitative analysis of requirements completeness in agile projects. By integrating story-level quality, backlog-level coverage, and readiness evaluation, US-CEP contributes both academically and practically to improving requirements engineering in user story–driven development environments.

II. Related Work

A. Agile Requirements Management

Agile development frameworks position responsiveness to change as their core value, regarding requirements not as fixed specifications but as continuously evolving artifacts shaped through collaborative and iterative processes [4]. User stories serve as the primary instrument for capturing requirements in agile teams, offering a lightweight mechanism that emphasizes user value and avoids excessive upfront documentation [5].

However, this simplicity also introduces two critical constraints from a requirements engineering perspective. First, user stories omit structural information traditionally found in formal specifications—such as system boundaries, data flows, functional dependencies, and exception handling—which makes it difficult to infer system-level behavior or ensure architectural coherence [6]. Second, because user stories are typically written at a granular and functional level, non-functional requirements (NFRs)—e.g., performance, security, reliability—are inconsistently documented or entirely omitted, increasing the likelihood of operational failures in later stages [7]. Multiple empirical studies have reported such limitations in real-world agile environments. Inayat et al. [8], analyzing more than 20 industrial agile projects, identified recurring problems including NFR omission, unclear prioritization, and insufficient dependency analysis. Wagner et al. [9] similarly found that large-scale projects struggled to maintain a coherent system-level structure when relying primarily on user stories—functional requirements were often captured only partially, leading to fragmentation. Ramesh et al. [10] further observed that the lack of systematic traceability in agile settings contributes to increased rework and difficulty in ensuring consistency across iterations.

Overall, while agile methodologies promote flexibility and collaboration, the absence of structured evaluation mechanisms leads to requirement gaps that significantly impact project quality. This motivates the development of a method that complements agile practices with structured completeness assessment.

B. Quality Assessment Methods for User Stories

The INVEST criteria—Independent, Negotiable, Valuable, Estimable, Small, and Testable—proposed by Bill Wake [2] represent the most widely adopted standard for evaluating user story quality. INVEST is effective for assessing the clarity and granularity of individual stories, making it a practical checklist for agile teams.

However, several researchers highlight fundamental limitations of the INVEST approach. Specifically, INVEST focuses exclusively on sentence-level characteristics of user stories and does not address whether the collection of stories holistically represents the complete system requirements [6]. Thus, even a set of high-quality stories (per INVEST) may still yield incomplete functional flows, missing dependencies, or insufficient coverage.

To address these gaps, later studies proposed more detailed analyses of language quality, semantic coherence, and testability. Lucassen et al. [11] introduced a five-dimensional framework assessing negotiability, understandability, feasibility, logical cohesion, and testability, using quantitative metrics to guide evaluation. Dalpiaz et al. [12] developed an NLP-based automated analysis tool that detects linguistic and semantic issues in user stories. Ferrari et al. [13] demonstrated that ambiguous terminology within user stories is one of the primary causes of test design failures, emphasizing the need for systematic language quality evaluation.

Despite these contributions, existing approaches remain limited in scope: they evaluate individual stories rather than the structural completeness of the entire backlog. They do not address higher-level concerns such as functional coverage, user journey mapping, or dependency completeness—core components of systems-level requirements engineering. These gaps directly motivate the Layer 2 and Layer 3 assessments proposed in the US-CEP framework.

C. Requirements Completeness Assessment

Requirements completeness is a central quality attribute in classical requirements engineering. ISO/IEC/IEEE 29148 defines completeness as a state in which all necessary functions, constraints, and interactions of a system are captured without omission [3]. To achieve this, traditional RE employs structured artifacts such as functional lists, use cases, sequence diagrams, statecharts, and data models. While these artifacts enable thorough completeness assessment, they assume document-driven processes and therefore do not directly align with agile environments where user stories are the primary requirements vehicle.

Research on requirements completeness within agile settings remains relatively limited, but common challenges are consistently identified. Several studies note that user story–centric projects lack systematic mechanisms for ensuring functional coverage [9]. Specific stages of the user journey or business process often fail to map to any user stories, resulting in structural gaps. Moreover, user stories rarely define dependencies—whether in terms of workflow sequencing, data exchange, or external system interactions—leading to fragmented system behavior [14].

Studies focusing on preparation and readiness, including DoR and DoD practices, report that inconsistent readiness across stories creates quality variance and reduces sprint predictability [15]. Furthermore, the absence of NFRs at the story level is highlighted as one of the primary causes of post-deployment system failures. Existing research typically concentrates on isolated aspects—language quality, testability, or case-specific evaluations—rather than offering a holistic view. To date, no comprehensive model has been proposed

that concurrently addresses (1) the linguistic and structural quality of individual user stories, (2) the functional flow and user-journey coverage of the product backlog, and (3) the readiness of requirements for development and testing activities.

This research gap provides the rationale for the US-CEP framework, which seeks to deliver a multi-layered, quantitatively grounded approach to achieving requirements completeness in user story–driven agile projects.

III. US-CEP FRAMEWORK

This chapter presents the US-CEP (User Story–based Completeness Evaluation Process), a multi-layered framework designed to systematically assess the completeness of requirements in agile projects that rely on user stories as their primary requirements artifacts. The framework addresses structural limitations identified in prior research and industry practice by evaluating completeness across three interconnected layers: user story quality, backlog-level structural coverage, and development/testing readiness. The design emphasizes both methodological rigor and practical applicability to agile teams.

A. Design Principles and Objectives

The US-CEP framework is founded on three core design principles that collectively ensure its robustness and applicability in agile environments. First, multi-layered completeness acknowledges that requirements quality cannot be inferred solely from the attributes of individual user stories; instead, completeness must be evaluated across multiple levels—story quality, backlog-level structural coherence, and readiness for implementation—since deficiencies in any layer can jeopardize the integrity of the overall system. Accordingly, the framework adopts a hierarchical structure comprising Layer 1 (User Story Quality), Layer 2 (Backlog-Level Structural Completeness), and Layer 3 (Development and Test Readiness). Second, quantifiability is embedded through the use of measurable and reproducible metrics, enabling objective monitoring of requirement quality and supporting data-driven activities such as sprint and release planning. Third, practical applicability ensures seamless integration with existing agile ceremonies—including backlog refinement, sprint reviews, and DoR/DoD assessments—so that the framework enhances evaluation rigor without imposing additional documentation burden or disrupting established team workflows.

Together, these principles enable US-CEP to function as a practical yet analytically rigorous mechanism for assessing the completeness of requirements in agile projects. The framework aims to detect missing or ambiguous requirements at early stages, enhance transparency and shared understanding among stakeholders, reduce rework and mitigate risks in downstream development and testing, and provide quantifiable indicators that support continuous process improvement.

B. Overview of the Framework and Layered Architecture

US-CEP is structured as a 3-layer completeness evaluation model, with each layer representing a different perspective on requirements completeness. Although each layer can be evaluated independently, the combined assessment offers a holistic view of how well the project requirements are defined.

Fig. 1 conceptually illustrates the three-layer architecture of the proposed framework. Layer 1 focuses on the linguistic clarity, structural soundness, and testability of individual user stories, ensuring that each requirement is well-formed and verifiable. Layer 2 extends the evaluation to the system level by analyzing the completeness of functional coverage, user journey alignment, and inter-story dependencies across the entire backlog. Finally, Layer 3 assesses implementation and verification readiness through Definition of Ready (DoR), Definition of Done (DoD), and risk-driven feasibility analysis, enabling a holistic examination of whether the backlog is sufficiently prepared for development.

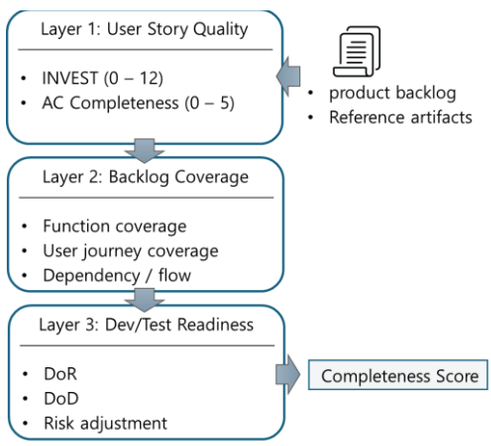


Fig. 1. Overall architecture of the US-CEP 3-Layer Completeness Model

C. Layer 1: User Story Quality Completeness

Layer 1 focuses on assessing the quality of individual user stories to determine whether each story clearly conveys the intended functionality and is sufficiently testable. This layer extends the well-known INVEST criteria by introducing a quantitative scoring model suitable for practical use.

1) INVEST-Based Quantitative Model

Each INVEST component is rated on a 0–2 scale based on the criteria shown in Table 1. The maximum INVEST score is 12 points. Stories scoring below 8 are flagged as needing improvement.

TABLE I. TABLE TYPE STYLES

Criterion	Description	Scoring Guide
Independent	Degree of functional independence	0: Highly dependent, 1: Partially independent, 2: Fully independent
Negotiable	Openness to refinement without design mandates	0: Contains design decisions, 2: Negotiable
Valuable	Clear expression of user/business value	0: Unclear, 2: Explicit value
Estimable	Ability to estimate development effort	0: Not estimable, 2: Easily estimable
Small	Appropriate scope for a single sprint	0: Oversized, 2: Properly scoped
Testable	Ability to derive test cases	0: Cannot be tested, 2: Clear testability

2) Acceptance Criteria (AC) Completeness

Because Acceptance Criteria (AC) play a crucial role in determining the testability of user stories, AC completeness is evaluated using five binary indicators—namely, the definition of the normal workflow, the inclusion of exception-handling

conditions, the presence of relevant non-functional requirements, the absence of ambiguity, and support for deriving test cases—each assigned one point, resulting in a maximum AC score of five points.

The Layer 1 score is then computed by aggregating the INVEST score and the AC completeness score and normalizing the combined value to a 20-point scale, with individual story scores subsequently averaged across the product backlog to represent overall story-level completeness.

D. Layer 2: Backlog Coverage Completeness

Layer 2 evaluates requirements completeness at the system level. This layer addresses a well-documented limitation of user story–driven development: the absence of guarantees that the backlog fully covers the system's functional and behavioral requirements. Layer 2 consists of three main components.

1) Functional Coverage

Functional coverage measures how many functions from the reference functional model (e.g., feature list, BPMN model) are represented by corresponding user stories. Coverage below 90% typically indicates meaningful omissions requiring immediate attention.

$$FC = \frac{Number\ of\ mapped\ functions}{Total\ number\ of\ functions} \quad (1)$$

2) User Journey Coverage

User journey coverage evaluation assesses whether the complete sequence of interactions that a user performs to achieve a specific goal within a system is fully represented by a set of user stories in an agile backlog. In this paper, a user journey is defined as follows:

Definition 1 (User Journey). A user journey is an ordered sequence of interaction steps that a specific user persona performs within a system to accomplish a defined goal.

User Journey Coverage evaluates not only the existence of individual user stories but also their ability to collectively support an end-to-end user experience without functional gaps. The User Journey Coverage is calculated as shown in (2):

$$UJC = \frac{Journey\ steps\ with\ associated\ stories}{Total\ journey\ steps} \quad (2)$$

Although individual user stories may satisfy established quality criteria such as INVEST, agile projects often exhibit systemic deficiencies that are not observable at the story level. In particular, functional steps may exist in isolation without forming a coherent end-to-end workflow, while alternative or exceptional execution paths are frequently omitted in favor of a single happy path. Moreover, dependencies and execution order among user stories are often insufficiently specified, resulting in fragmented implementations in which features are technically delivered but users are unable to achieve their intended goals. Such limitations cannot be adequately identified through story-level quality evaluation alone, underscoring the need for a system-level, user-centric approach to requirements completeness assessment.

3) Dependency and Flow Analysis

This step ensures that preconditions and postconditions across user stories are properly aligned, data dependencies are explicitly captured, interactions with APIs or external systems are adequately represented, and user interface or event flows remain coherent throughout the system. By examining these aspects holistically, the analysis effectively identifies workflow gaps—such as missing steps in critical processes like purchasing—that are unlikely to be detected when user stories are evaluated in isolation.

Layer 2 is assessed on a 40-point scale based on the weighting of three components: functional coverage, user journey coverage, dependencies, and flow completeness. This scoring system quantitatively assesses the completeness of system-level requirements by reflecting both the presence of essential functionality and the consistency of interactions between features. For example, missing key steps, such as user authentication or order confirmation, directly lowers the overall coverage score and reveals structural risks that can impact the overall system behavior.

E. Layer 3: Development and Test Readiness

Layer 3 assesses whether user stories are sufficiently prepared for implementation and verification, thereby reflecting practical considerations that directly influence agile execution. Unlike preceding layers that focus on requirement clarity and structural completeness, this layer emphasizes readiness for development and testing, ensuring that user stories can be reliably incorporated into sprint planning without introducing avoidable risks.

1) DoR(Definition of Ready)

The Definition of Ready (DoR) is evaluated using five binary indicators that determine whether a user story is adequately specified prior to development. These indicators include the clarity of functional intent and business value, the presence of sufficient acceptance criteria, the explicit definition of non-functional requirements, the availability of relevant UI/UX references, and the clear specification of external system or API dependencies. Collectively, these criteria ensure that development teams have the necessary information to begin implementation with minimal ambiguity.

2) DoD(Definition of Done)

The Definition of Done (DoD) focuses on completion and validation readiness by examining whether appropriate functional test cases have been defined, boundary and exception scenarios are addressed, non-functional test preparations are in place, relevant documentation has been updated, and the story is ready for formal code review. This evaluation confirms that quality assurance and verification activities can be effectively executed once development is complete.

3) Risk assessment

In addition, Layer 3 incorporates a risk assessment component that accounts for factors such as technical uncertainty (e.g., the adoption of new or unproven APIs), ambiguity in user experience or requirements, dependency-related risks, and overall domain complexity. Risk is modeled as a multiplicative factor ranging from 0 to 1, which adjusts the readiness score to reflect the potential impact of these uncertainties.

The final Layer 3 score is computed by normalizing the risk-adjusted sum of DoR and DoD scores to a 40-point scale. This scoring mechanism effectively distinguishes between user stories that appear complete at a superficial level but lack practical feasibility, and those that are genuinely ready for inclusion in a development sprint with a high likelihood of successful implementation and verification.

F. Aggregation into an Overall Completeness Score

The overall requirements completeness is computed by aggregating the evaluation results of the three layers using a weighted summation scheme. This aggregation strategy reflects the relative importance of each layer in assessing the completeness of user story–based requirements in agile projects. Layer 1, which focuses on the quality and testability of individual user stories, contributes up to 20 points and is assigned a weight of 0.3. Layer 2, addressing system-level structural completeness through functional coverage, user journey coverage, and dependency and flow coherence, contributes up to 40 points and is assigned the highest weight of 0.4, reflecting its critical role in ensuring end-to-end usability. Layer 3, which evaluates development and test readiness, also contributes up to 40 points and is weighted at 0.3. Based on this distribution, the final completeness score is calculated as a linear combination of the three layer scores, as shown in (3):

$$CS = 0.3 \times L1 + 0.4 \times L2 + 0.3 \times L3 \quad (3)$$

The resulting score provides a normalized quantitative indicator of requirements completeness on a 0–100 scale. Scores exceeding 80 are interpreted as indicating stable and sufficiently complete requirements that are suitable for agile development with minimal risk. Scores between 60 and 79 suggest moderate completeness, signaling the need for targeted refinements or backlog grooming to address identified gaps. Scores below 60 indicate a high risk of incompleteness, implying that significant requirement deficiencies may lead to rework, implementation delays, or test failures if development proceeds without corrective action.

To support practical adoption and automation, the entire evaluation workflow—from individual story analysis in Layer 1, through system-level coverage assessment in Layer 2, to readiness and risk-adjusted scoring in Layer 3—is formalized in Algorithm 1. This algorithm presents a step-by-step procedure that enables consistent application of the proposed framework and serves as a foundation for tool-based implementation in agile project environments.

ALGORITHM I. US-CEP COMPLETENESS EVALUATION

```
Input:
   B: product backlog (set of user stories)
   M_f: reference functional model (feature list or BPMN)
   M_j: reference user journey model (set of journey steps)

Output:
   CompletenessScore ∈ [0, 100]

1:  // ----- Layer 1: User Story Quality -----
2:  L1_sum ← 0
3:  for each story s in B do
4:     INVEST_s ← Evaluate_INVEST(s)     // 0–12 points
5:     AC_s    ← Evaluate_AC(s)         // 0–5 points
6:     L1_s    ← Normalize_to_20(INVEST_s + AC_s)
7:     L1_sum  ← L1_sum + L1_s
8:  end for
9:  L1 ← L1_sum / |B|                   // avg. story score (0–20)

10: // ----- Layer 2: Backlog Coverage -----
11: F_mapped ← Map_Functions(M_f, B)     // function ↔ story mapping
12: J_mapped ← Map_Journeys(M_j, B)  // journey step ↔ story mapping
// flow & dependency evaluation (0–1)
13: DepScore ← Evaluate_Dependencies(B)

14: FCov   ← |F_mapped| / |M_f|          // function coverage (0–1)
15: JCov   ← |J_mapped| / |M_j|          // user journey coverage (0–1)

16: L2 ← Normalize_to_40( w_f * FCov + w_j * JCov +w_d * DepScore )
      // w_f + w_j + w_d = 1

17: // ----- Layer 3: Dev/Test Readiness -----
18: L3_sum ← 0
19: for each story s in B do
20:    DoR_s  ← Evaluate_DoR(s)          // 0–5
21:    DoD_s  ← Evaluate_DoD(s)          // 0–5
22:    Risk_s ← Evaluate_Risk(s)         // 0–1 (1 = no risk)
23:    Ready_s ← (DoR_s + DoD_s) / 10    // 0–1
24:    L3_s   ← Normalize_to_40( Ready_s * Risk_s )
25:    L3_sum ← L3_sum + L3_s
26: end for
27: L3 ← L3_sum / |B|                    // avg. DoR score (0–40)

28: // ----- Aggregation -----
29: CompletenessScore ← 0.3 * L1 + 0.4 * L2 + 0.3 * L3

30: return CompletenessScore
```

IV. CASE STUDY

This chapter presents an empirical case study designed to examine the practical applicability and analytical effectiveness of the proposed US-CEP in a realistic agile development context. The primary objective of the case study is to investigate whether US-CEP can systematically and quantitatively identify requirement completeness issues that are often overlooked in user story–driven backlogs, despite adherence to agile best practices. To this end, US-CEP was applied to a mid-sized e-commerce web platform project developed by a six-member agile team. The project employed an iterative sprint-based process, and all functional requirements were documented exclusively in the form of user stories. By evaluating the backlog across the three analytical layers of US-CEP—namely, (i) individual user story quality, (ii) system-level backlog completeness, and (iii) development and verification readiness—this case study aims to demonstrate how multi-layered completeness assessment can reveal latent risks that are not detectable through story-level inspection alone.

A. Study Design

The target system is an e-commerce web platform intended to support typical online purchasing activities. At the time of evaluation, the project backlog comprised 18 user stories, each written using the standard agile template: "As a [role], I want [goal], so that [business value]." The scope of the system was organized around five major functional areas: User authentication and account management, Product search and detailed product information viewing, Shopping cart management and order creation, Payment processing, and Order history inquiry and retrieval.

The project was considered sufficiently mature for evaluation, as the backlog had undergone multiple refinement sessions and was actively used for sprint planning. The following development artifacts were collected and used as inputs for the US-CEP analysis:

- Eighteen user stories documented in textual form
- Acceptance Criteria (AC) associated with each user story
- A functional structure model consisting of a feature list and BPMN-based workflow diagrams
- A reference user journey model representing a canonical purchase flow (Login → Search → Cart → Payment → Confirmation)

- Standardized Definition of Ready (DoR) and Definition of Done (DoD) checklists used by the team

These artifacts enabled both micro-level (story-specific) and macro-level (system-wide) completeness analysis. US-CEP was applied sequentially across its three layers. Each layer produced a normalized score reflecting completeness from a distinct analytical perspective. Layer-specific scores were subsequently aggregated using a weighted scoring model to derive the overall completeness index. This structured procedure ensured traceability between identified deficiencies and their corresponding evaluation dimensions.

B. Results

The results of the case study are described in Table 2. This case study demonstrates that compliance with agile best practices and well-formed individual user stories does not inherently ensure requirement completeness at the system level. Although the evaluated backlog exhibited relatively strong user story quality and functional coverage, US-CEP revealed substantial deficiencies in alternative flow specification, dependency clarity, and development and verification readiness—particularly regarding non-functional requirements. The findings confirm the necessity of a multi-layered completeness evaluation approach, showing that latent risks remain undetected when assessment is limited to story-level inspection. Overall, the results validate US-CEP as an effective analytical framework for identifying structural and readiness-related requirement gaps in user story–driven agile projects.

TABLE II. US-CEP CASE STUDY EVALUATION RESULTS

US-CEP Layer	Evaluation Focus	Key Findings	Score
Layer 1: User Story Quality	INVEST compliance and Acceptance Criteria completeness	Overall story quality was acceptable; however, testability and exception handling were frequently underspecified, and several stories embedded UI-level decisions, reducing negotiability.	16.6 / 20
Layer 2: Backlog Structural Completeness	Functional coverage, user journey alignment, and dependency consistency	Most core functions were covered, but critical functions (e.g., guest checkout, cart expiration) were missing. Alternative user flows and inter-story dependency clarity were insufficient in complex payment scenarios.	33 / 40
Layer 3: Development and Verification Readiness	DoR/DoD fulfillment and risk exposure	Readiness for implementation and testing was limited due to missing non-functional requirements, incomplete external dependency specifications, and insufficient exception and boundary test definitions.	24 / 40
Overall	Weighted aggregation across three layers	High-quality individual user stories did not guarantee system-level completeness or implementation readiness.	63.2 / 100

V. CONCLUSION

This study presented US-CEP, a multi-layered framework for evaluating requirements completeness in user story–driven agile development. By integrating user story–level quality assessment, system-level backlog structural analysis, and development and verification readiness evaluation, US-CEP extends conventional agile practices that primarily focus on individual story formulation. The empirical case study demonstrated that even well-refined backlogs may contain latent completeness deficiencies, particularly in exception handling, non-functional requirements, and implementation readiness. These findings highlight that high-quality user stories alone are insufficient to guarantee system-level completeness and that a structured, multi-perspective evaluation is essential for reducing downstream risks.

Despite its contributions, this study has several limitations that motivate future research. First, the empirical validation was conducted within a single application domain, suggesting the need for additional studies across diverse and highly regulated domains to assess generalizability. Second, some evaluation components rely on expert judgment, indicating opportunities to improve objectivity and scalability through NLP- or LLM-based automation for user story analysis and completeness scoring. Future work may also explore tighter integration of US-CEP with agile project management tools to enable continuous and real-time completeness monitoring throughout the development lifecycle. Finally, large-scale empirical studies involving heterogeneous and distributed agile teams are required to further validate the robustness and practical impact of the proposed framework..

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, “Title of paper if known,” unpublished.

[5] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

[8] K. Eves and J. Valasek, “Adaptive control for singularly perturbed systems examples,” Code Ocean, Aug. 2023. [Online]. Available: https://codeocean.com/capsule/4989235/tree

[9] D. P. Kingma and M. Welling, “Auto-encoding variational Bayes,” 2013, arXiv:1312.6114. [Online]. Available: https://arxiv.org/abs/1312.6114

[10] S. Liu, “Wi-Fi Energy Detection Testbed (12MTC),” 2023, gitHub repository. [Online]. Available: https://github.com/liustone99/Wi-Fi-Energy-Detection-Testbed-12MTC

[11] “Treatment episode data set: discharges (TEDS-D): concatenated, 2006 to 2009.” U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Office of Applied Studies, August, 2013, DOI:10.3886/ICPSR30122.v

Digital Forensics of the Embedded On-Devices AI

Hyun-Mook Cho*, Jang-Geun Ki*, KyuTae Lee**, Do-Hyeun Kim***
* *Div. of Electrical, Electronics, and Control Eng., Kongju Nat. University, S.Korea*
** *Dept.. of Smart Information Technology Eng., Kongju Nat. University, S.Korea*
*** *Dept. of Computer Engineering, Jeju National University, Republic of Korea*
Emails: *ktlee,kgj,hmchov@kongju.ac.kr , kimdh@jejunu.ac.kr*

***Abstract*— Digital data can be modified and revised by simple handling as it is coded with the binary data which is stored in files. Digital forensics is a procedure for investigating digital data because it can be used as legal evidence. The hardware information device called embedded system is operated with specific processors and source code programs. For illegally altered data to be used as legal evidence, it must be traced back to the developer's initial creation process. This is the reason that the developer's unique methods and development coding habits are involved in hardware configuration, arrangement, and file structure of the software. In this paper, we discuss the method for analyzing and utilizing forensic information generated by illegal copy data**

Keywords—forensics, device, illegal copy, copyright

I. INTRODUCTION

Currently, the information digital devices are widely used in many areas of our life. In these information devices, there are lots of information stored on the memory device such as user names, time life, work content, and movement status data. Therefore, it can traced out person's situation with the recorded information stored on the device. Digital forensics is the work of finding evidence, records, etc. by analyzing the information devices which is memory, disk, etc. of digital devices. All information stored on CCTV and mobile phone messages is recorded and saved on servers or user terminals. Also, in our life environment, each of devices has stored our movement information on the CCTV which is installed on city streets and public transportation vehicles and on the transportation card tags when riding public transportation, tags or entrance CCTV when entering or exiting an office, SNS, mobile phone text messages, etc. Digital forensics is a technology to prevent crimes from the information stored.[1] Generally, when a new IT information device is developed with high market income, there are issues of copyright loss due to the illegal copy products. In other words, illegal copy occurs by stealing copyright ownership without permission, also producing and selling illegal devices by some developers who have moved to illegal companies. In the case of this type of dispute, the software and hardware of illegal similar devices are compared and appraised with original devices. Software is evaluated by comparing the similarity of their functions or source codes. On hardware part, it is difficult to find an evidence of illegal copy as many components use known circuits or parts. In this paper, we estimated the availability weather forensic can be used for hardware and software evaluation.

II. EVALUATION PROCESS

The procedure of searching for similarity with both object materials is shown in figure 1. At first, a person who has disadvantages on his copyright properties makes issues an accuses a person who has an illegal copy. Then the agency begins to investigate the object materials submitted from both sides. On this time, if there can be pointed out a specific part to investigate, it is possible to assign specific search items to copyright association with agreement. As a next step, the evaluation expert takes an investigation for the similarity ratio to the specific item or total items as requested. On the evaluation task, assigning subtitle items is important for reducing an objective result and applying a weight factor to each item. A designation for evaluation experts is assigned by the copyright association for comparing the subject materials. Also searching for open source to specific items is performed to find out the originality of the property owner on the materials. [9] On the analysis work, the decision weather illegal or legal depends on the evaluation expert. He deduces the original evidence and tries to find out which kinds of knowledge are used, such as open knowledge or specific one.

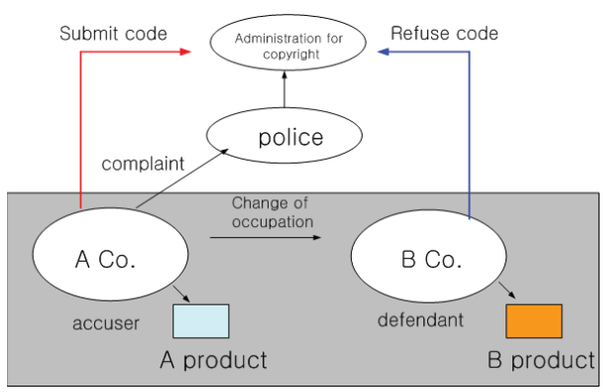


Fig. 1 Procedure of digital forensics evaluation

III. DEVICE FORENSIC OF THE EMBEDDED DEVICES

3.1. Hardware Material

The object material of hardware components has an electrical circuit, micro processor module, application software and firmware. As an example of hearing aid, these devices work to amplify human voice for hearing aid with four blocks. When a voice inputs to microphone, preamplifier circuit increases the signal level and comparing it to feedback signal from amplified output parts. For the human ear, the power amplifier used to drive a physical

speaker to speak out loud voice. In this circuit function, the feedback block, ALC(Automatic Level Control) is used for output signal to go back to preamplifier block to control output level. Finally, the controller block has a microprocessor and software program to control key input with useful UI(User Interface) and to indicate operation control such as battery status and operation status and volume control.

3.2. Software Material

The evaluation system is about a treadmill for health exercise. It is composed a simple mechanic device and software programmed. The characteristic of system is running belt system with motor and control board included microprocessor. Program works with sensor input signal and display all of information to indicate. Also it has functions of storing and transmitting of health information to the person. At first, programmer creates a flowchart of operation sequence. And with proper processor and program language, software codes are generated. After compiling, machine code downloaded to the processor as binary code values. These kinds of device have specific characters in copyright, such as interface logic, software language and operation sequence.

3.3. PCB Material

The development work of Hardware system is completed by PCB(Printed Circuit Board) assembly. This board is designed to fix a device size, component position, location of terminals such as power, input, output and key pad. The design of the board is started with block diagram and circuit diagram of the system. There are many kinds of PCB design tools released in commercial market. The tools are depended on the developer who is used to be to work such as ORCAD, PADS. But the last file designed is Gerber file to share with all of designer. PCB is a practical circuit board which is composed of printed lines to connect a all of component terminal and let us know the location of components. This has an originality of developer knowledge in design. That knowledge is contained in PCB size, types of pattern, component location, components distance, connection line and width of pattern. All of these items can be a copyright of owner.

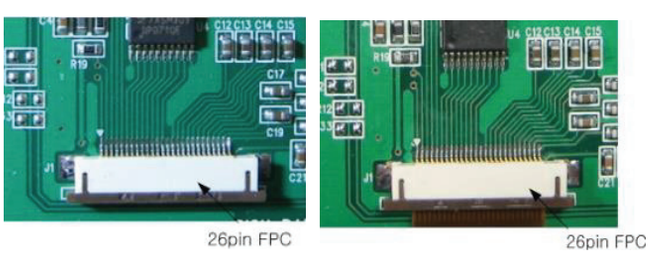


Fig.2 PCB similarity of embedded device

Therefore PCB is mainly dependent on the designer. Also the board is very sensitive in high frequency as an interference called EMI(Electro Magnetic Interference). In case of designing DDR SDRAM(Double data rate synchronous dynamic random access memory), it works on fast data transition rate on 400MHz frequency. To make a line between CPU and DDR is needed a damping resistor to decrease noise and match for impedance

IV. Conclusions

The development of information device is modularizing with commercial components and open source program. On developer's perspective, It is easy to make the system operation design with identical or similar components and software code. However, even if the identical components and circuits are used on system design , there are characteristics in which the PCB pattern is created according to the original developer's technical idea or the arrangement of components contains unique knowledge. In this paper, we discussed how to utilize these characteristics as a forensic information in cases where the original developer's ideas are used illegally. In the future, we try to find and define various singularities. So it is expected that they will be utilized as unique information for copyright protection of the original developer even if the circuits and components are identical and also software code.

Acknowledgment

This work was supported by Institute of Information & communications Technology Planning & evaluation (IITP) grant funded by the Korea government(MSIT) (No. RS-2025-02263031, and this work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT)(No. RS-2024-00346238), Any correspondence related to this paper should be addressed to KyuTae Lee

References

[1] https://www.copyright.or.kr/eng/main.do

[2] Korea Software Copyright Committee," Establishment of the mid-and long term strategies to protect original technology of software", (2009).

[3] R. K. Gupta, "Introduction to Embedded system", ICS 212, winter, (2002).

[4] K. Lee and S. Kim, "Source Code Verification on Embedded Control Devices," International Journal of Software Engineering and Its Applications, vol. 10, no. 5, (2016), pp. 69-76.

[5] K. Lee, "Classification for Similarity Evaluation on Embedded System", The journal of the Korea Software Assessment and Valuation Society, vol. 11, no. 1, (2015), pp.47-51.

[6] Do-Hyeun Kim, KyuTae Lee, "Management of Reliability and Delivery for Software Object Material", Journal of Software Assessment and Valuation(JSAV), vol.15, No.2, pp.51-57. Dec. 2019.

[7] K. T. Lee, W. Ye, H. M. Cho, D. H. Kim and H. Ahn, "A Comparison Method of Intellectual Ownership materials", ICGHIT 2017 conference, Hangzhou China, (2017), pp. 84-85,.

[8] Do-Hyeun Kim, KyuTae Lee, "Management of Reliability and Delivery for Software Object Material", Journal of Software Assessment and Valuation(JSAV), vol.15, No.2, pp.51-57. Dec. 2019.

Exploring User Experience Components for the HSE Metaverse Evaluation Framework

Al-amin Abdulhamid [1], Khairul Shafee Kalid [1], Savita K Sugathan [1], Shamsu Abdullahi [1], Muhammad Sabo Yahaya [1], Sarah Aisyah Isnani[1]
[1] *Dept. of Computing,Universiti Teknologi PETRONAS, Perak, Malaysia*

Emails: *al-amin_24009605@utp.edu.my, khairulshafee_kalid@utp.edu.my, savitasugathan@utp.edu.my, shamsu_24001499@utp.edu.my, muhammad_24001658@utp.edu.my, sarah_21001863@utp.edu.my*

***Abstract*—Despite Metaverse potential to deliver effective Health Safety and Environment (HSE) training and awareness, there remains a gap in balancing the engaging experience with the learning aspects. User experience (UX) is the major concern as it captures users feelings while interacting with the system, which can lead to either positive or negative results. Consequently, UX evaluation is crucial for improving the quality of experience and contributes to bridging the existing gap by finding appropriate components for HSE Metaverse. This study explores the UX components for the evaluation and design of HSE Metaverse. Metaverse environment can facilitate collaborative experience while making education more accessible regardless of location or physical ability. Good UX framework for HSE Metaverse will help designers to evaluate UX and ensure the immersive virtual environments are intuitive, effective, and successful. Therefore, it is essential to identify suitable components in order to model appropriate UX framework for HSE Metaverse.**

Keywords—Metaverse, User Experience, Health, Safety, and Environment, UX Evaluation, Immersive Environments

I. INTRODUCTION

While health, safety, and environment (HSE) play a crucial role in all industries, different sectors face distinct HSE hazards that require interventions [1]. Workers in industries are exposed to a variety of possible risks, including falling equipment, working at height, and hazardous chemicals, which may result in injury [2]. Global data indicate that 2 million deaths have been documented due to workplace-related accidents and illnesses, with a total of 270 million occupational accidents that occur annually [3]. Therefore, several nations around the world have prioritized HSE as a crucial component in minimizing accidents [4].

A key goal of any HSE management system is to retain relevant information and disseminate knowledge. The realization of this goal has been accomplished through conventional training and educational approaches such as lectures, videos, and on-site demonstrations. These approaches are time consuming, costly, and often require physical participation particularly when such training must be repeated numerous times to large cohorts of workers [5]. Therefore, to deliver effective HSE training, Metaverse technology is introduced.

Metaverse provides an immersive experience that goes beyond the constraints of the physical world, emphasizing its integration of augmented reality, virtual reality, and real life [6]. A metaverse platform can transform the learning environment into a dynamic, collaborative, interesting, and hands-on experience [7]. A study by Park [8] stated that, when users can collaborate in a shared immersive space rather than just experience it alone, it can make the experience more exciting, engaging, and immersive, giving people a stronger sense of presence and enjoyment. Therefore, incorporating HSE into Metaverse expands its application possibilities into areas such as education and training.

To ensure that the HSE Metaverse is effective, the environment must be intuitive, easy to navigate, and capable of delivering a positive user experience that supports learning and skill development. As a result, a well-structured evaluation framework is essential to identify flaws, improve usability, and enhance overall environment efficiency. User experience (UX) evaluation plays a role in this process by enabling designers and developers to assess how users perceive, react and interact with the HSE Metaverse experience. Therefore, a UX evaluation framework is required for guiding iterative improvement and ensuring HSE Metaverse meets its learning objectives.

Assessing user experience in HSE Metaverse involves multiple aspects, such as the level of immersion, skill acquisition, behavioral changes, and so on. To address these needs, developers, designers and researchers must identify suitable components that can be used to construct an UX evaluation framework for the HSE Metaverse. Therefore, the main contribution of this study is the identified components that can evaluate UX in the HSE Metaverse environment. Section two discusses the background and related literature reviews. Section three describes the research methodology, while Section four presents the conclusions.

II. LITERATURE REVIEW

This section presents the background of the study and reviews the existing frameworks and models used to evaluate UX in immersive environments. It synthesizes how the frameworks have been applied and highlights the gaps that motivate the development of a dedicated UX evaluation framework for the HSE Metaverse.

A. User Experience

The notion of UX is relatively new in the domain of human-computer interaction. Don Norman coined the term “user experience” in 1995 to describe a multidimensional concept whose definition has yet to be fully determined [9]. Hassenzahl et al. reveal that, “There is currently no widely agreed-upon definition or meaning of what UX could or ought to be” [9]. UX is far from usefulness by incorporating qualities such as joy, beauty, personal progress, and enjoyment that are related to product usage. Enjoyment is the fundamental phrase in user experience [10].

ISO 9241-210 [11] defined UX as “every aspect of users’ feeling while interacting with the environment, facility, product or service”. In addition, Hassenzahl & Tractinsky [12], defined UX as “an outcome of users’ internal state (such as expectation, mood, and motivation), system attributes (such as usability, complexity, and functionality), and the environment (including social or organizational, voluntariness

of use, and usefulness of the activity)". Although usability and UX are synonymous, they are not the same. While usability refers to the interface's usefulness, efficiency, and effectiveness, UX emphasizes how a person feels while interacting with that system [10]. UX focuses on the user's well-being, not performance, as a result of contact. According to Hartson [13], UX is a complicated phenomenon due to the active engagement of interactive information about the user and product, such as social, emotional, expectations, and other aspects.

Regarding user experience assessment, the fundamental benefit is to evaluate how a product or service effectively meets the needs of its users. In addition, it seeks to understand several aspects of computer interaction far from traditional usability, including hedonics, aesthetics, emotions, and overall quality of experience [12]. UX evaluation allows developers and designers to identify design flaws and improve the design quality to fulfill user expectations [14]. In conclusion, the goal of assessing UX is to create services and products that are usable, useful, and highly desirable.

B. Metaverse

The concept of the metaverse has its roots in Neal Stephenson's novel called Snow Crash, which was initially published in 1992 [15]. Stephenson describes a shared virtual environment over the internet where individuals represented by 3D avatars can create and share their own digital realities. As computer technology continues to evolve, Metaverse transitions from science fiction to reality, becoming a shared virtual space encompassing technologies including virtual reality (VR), augmented reality (AR), and mixed reality (MR). In general, Metaverse environments include various dimensions, including human-computer interface, advanced infrastructures, special computing, and decentralization mechanisms [16].

The advantages of Metaverse include increased convenience and adaptability through business expansion, information sharing, and virtual reality engagement [17]. Metaverse significances are expanding across various industries, including e-commerce, gaming, entertainment, education, and marketing [18]. In education, it allows learners to participate in hands-on learning activities, such as virtual field trips or complex simulations, which improve knowledge and retention [19]. While Metaverse has its roots in gaming, it now encompasses a much larger spectrum of applications in various industries.

User experience is crucial in realizing the full potential of Metaverse technology, which could revolutionize traditional workflows, communication, and learning processes. The illusion in which technology replaces the user's senses is known as immersion [20]. It is regarded as an experience that happens all at once. In immersive technologies, immersion is being seen as a critical part of the user experience [21].

C. User Experience Evaluation Frameworks

Immersive technologies have recently been developed and studied, covering a wide range of industrial training and awareness application. According to [22], various evaluation frameworks and models have been used to assess UX in immersive environment, such as User Experience Questionnaire (UEQ) [23], System Usability Scale (SUS) [24], Modular Evaluation of key Components of User Experience (MeCUE) [25], and Usability Heuristics [26]. The application of these frameworks in the context of immersive learning environment is discussed below.

1) UEQ: was developed by Laugwiz et al. [23], to serve as a tool for measuring user experience of interactive products, it consists of 26 pairs of opposing terms. Users rate their experience on a seven-point scale for each pair of statements. A shortened version, UEQ-S [27], which consists of eight of the 26 items, was created to speed up UX evaluation in a situation when using the entire questionnaire is not feasible. However, [24], [25], [26] highlight framework limitations in assessing immersive technologies. Firstly, the scope of the questionnaire only covers a portion of UX, especially affective and emotional reactions, neglecting the quality of experience in an immersive environment. Moreover, its appropriateness may fluctuate among apps and users due to variation in cultural background and linguistic skills. And finally, UEQ was unable to provide much information on certain usability problems or products design flaws that need to be fixed in the products.

2) SUS: The system usability scale (SUS), was constructed by John Brooke in 1986 at Digital Equipment Corporation, serves as the tool for assessing the perceived usability of the system [24]. It comprises of 10 statements rated on a five-point scale, alternating between positive and negative wording to reduce the response bias. SUS provides a single usability score, reflecting the overall user's satisfaction, ease of use, and comfort when interacting with the product. However, despite its widespread use, Dianatfar et al. [28], reveals some SUS limitations, especially in evaluating immersive systems. The scale provides only a global usability score without distinguishing between specific dimensions. In addition, SUS does not capture key experimental elements critically for evaluating an immersive environment.

3) MECUE: is a comprehensive, multidimensional framework developed by [25], to serves as a complete instrument, containing all necessary UX elements in a set of 34 items. It divides UX into five modules, and consists of items presented in a Likert scale format, focusing on both instrumental and hedonic aspects of UX. Study by [29] highlights drawbacks while assessing the immersive setting. The instrument was originally designed for conventional digital products and may not fully account for VR experiences. Fathermore, it is a relatively long questionnaire which can increase participant fatigue, potentially affecting response reliability in extended or complex sessions.

4) U*sability Heuristics:* Usability Heuristics are a set of general principles used to evaluate the user-friendliness and design quality of interactive systems. It was initially proposed by Jakob Nielsen in 1994 [30], consisting of 10 heuristics. These heuristics serve as expert guidelines for identifying usability issues early in the design process, without requiring large-scale user testing. While Usability Heuristics offers a comprehensive means of identifying design flaws. Cassandra et al. [29], express a notable limitation, such as it relies heavily on the expertise and subjective judgment of evaluators that can lead to inconsistency and bias, and it does not quantify user satisfaction or learning impact.

In summary, these frameworks provide a valuable foundation for assessing UX. However, they were not originally designed for immersive environment context and do not capture the full spatial embodiment features that describe the HSE Metaverse environment. These limitations highlight the need to identify the suitable UX components as a basis for a HSE Metaverse evaluation framework.

III. Methodology

This section discusses the proposed research design used in constructing the UX evaluation framework called MetaHSE-UX. The methodology adopted four sequential phases: *(i) Exploration, (ii) Data Collection and preparation, (iii) Data analysis, and (iv) Validation and Documentation*. The framework evaluates UX in the HSE Metaverse. Fig. 1 illustrates the phases involved in constructing the new framework with the following outcomes.

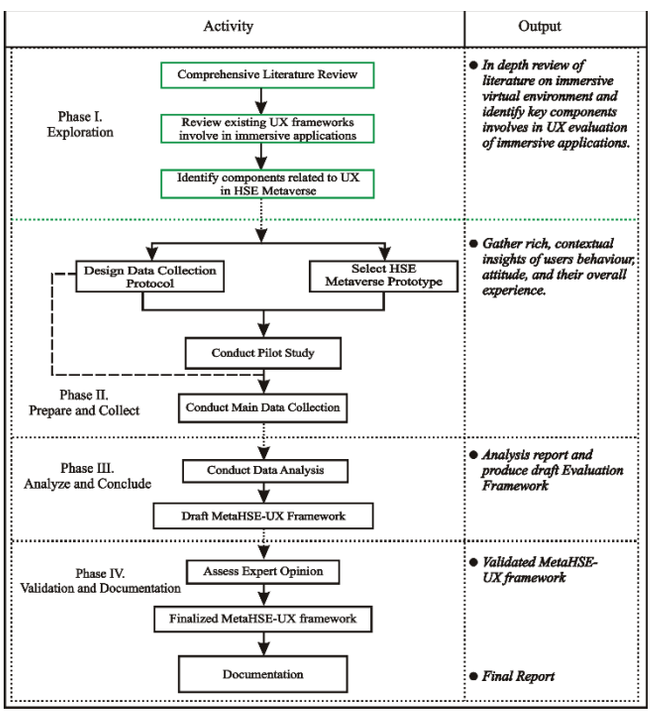


Fig. 1. Proposed Research Design

A. Phase 1: Exploration

In this phase, a comprehensive literature review on immersive technology was conducted to identify the research problems and objectives. A broad set of research papers was examined to extract various components specifically applicable for evaluating UX in the HSE Metaverse environment. In addition, various existing UX frameworks/models employed to evaluate immersive learning environments were critically reviewed. UEQ, SUS, meCUE, and Usability heuristics were frequently employed. Further discussion on applying these frameworks was explained in the previous section.

The next stage focuses on identifying the key UX components applicable for constructing MetaHSE-UX. In this stage, an in-depth analysis of immersive UX components was conducted. A content analysis approach was applied to quantify the recurrence of specific phrases, terms, and concepts in literature. Table 2 shows the identified components involved in the UX evaluation of immersive environments. Different researchers often use different terms to describe similar components. Consequently, while mapping these components, the underlying description provided in each source was used for comparison rather than relying on its titles. Additionally, the communality component is selected due to its underlying importance in designing HSE Metaverse despite its limited use in literature [35]. Communality captures shared experiences, collaboration, and a sense of being with others. This component represents the broader social characteristics of the HSE Metaverse.

TABLE I. Mapping of UX components in HSE Metaverse

UX components	Studies Involved	Initial Components
Immersion	[21], [36], [37], [38], [39], [40], [41]	✓
Presence	[21], [36], [37], [38], [39], [40], [42]	✓
Flow	[21], [36], [37], [38], [39], [42]	✓
Interactivity	[36], [40]]	
Challenge	[36]	
Usability	[37], [38], [39], [41], [42]	✓
Realism	[42]	
Loyalty	[36], [43]	
Emotion	[21], [36], [37], [39], [43]	✓
Vividness	[36]	
Learning	[36], [37], [39], [43]	✓
Judgment	[37], [39]	
Technology Adoption	[37]	
Engagement	[37], [43]	
Experience Consequences	[37], [38]	
Satisfaction	[21], [38]	
Usefulness	[21], [38]	
Visual Aesthetic	[40], [41]	
Intention	[21], [38]	
Confirmation	[21], [38]	
Communality	[41]	✓

Selecting appropriate and significant components related to HSE Metaverse will guide the subsequent stages of this research, especially the implementation and validation of the framework. Overall, the extracted components capture the core features of the HSE Metaverse that will be treated as a core UX dimension in constructing the HSE Metaverse application. Table 3 shows the description of each extracted UX component.

TABLE II. Proposed UX components for MetaHSE-UX

Components	Descriptions
Immersion	The extent to which users feel physically immersed in a virtual environment, in contrast to their awareness of the real world
Presence	The user's subjective feeling of being there in a virtual environment while located in the physical world.
Flow	The physiological state of complete absorption during a goal-oriented activity, characterized by deep focus, a sense of control, and enjoyment.
Usability	The degree to which users are able or fit to be used. A key consideration for all computer software evaluation is ease of use.
Learning	HSE design principles that facilitate users in focusing on the learning activities and skills acquisition, giving them a chance to learn.
Emotion	A holistic user's sensation of pleasure, joy, satisfaction, anxiety, etc, which users gain when interacting with the virtual environment
Communality	The user's sense of being there with others in the Metaverse and shared experience within virtual worlds.

B. Phase 2: Data Collection and Preparation

This research will adopt a qualitative exploratory case study approach by Robert [44] to investigate user experience in the HSE Metaverse context. A qualitative design was chosen because it enables the discovery of meanings that people attach to their experience. Data collection protocols will be developed to specify the study objectives, instruments, and ethical procedures. Participants will be recruited using purposive sampling, focusing on HSE practitioners with prior exposure to safety training and immersive experience. Qualitative data will be gathered during and after the interaction with HSE Metaverse through observations and semi-structured interviews, providing rich insights into how the proposed UX components are manifested in the immersive HSE training environment. The result of the data collection will contribute to designing the UX evaluation framework.

C. Data Analysis

In this phase, the data collected from the respondents will be imported into Atlast.ti for the systematic analysis. The transcripts and field notes will first be organized and coded, allowing recurrent patterns, themes, and meanings related to user experience in the HSE Metaverse to be identified. Through iterative coding, categorizing, and constant comparison, themes will be developed that reflect how users perceive immersion, presence, usability, emotion, and other relevant UX components. The results from this analysis phase will then be synthesized to inform the construction of the MetaHSE-UX framework tailored to an immersive HSE training environment.

D. Validation and Documentation

This phase involves obtaining and assessing expert opinions on the drafted MetaHSE-UX framework and subsequently finalizing the framework. The primary purpose of engaging experts is to gather expert feedback on the extracted UX components, confirm their relevance, and identify the most appropriate components for constructing the MetaHSE-UX framework. In this process, experts are asked to assess each component in terms of its suitability for evaluating UX in the HSE Metaverse. The initial draft of the proposed MetaHSE-UX framework will be submitted to these experts for validation. The expert will consist of both academic and industrial experts, including Metaverse designers. Table 3 summarizes the criteria for the expert selection.

TABLE III. CRITERIA FOR EXPERT SELECTION

Academic Expert	Industry Expert
• Have a minimum qualification of MSc or higher in computer science, IT, IS, or related discipline. • Must have prior teaching experience. • Have at least a minimum of 2-5 years' experience in HCI/UX research. • A full-time academician	• A knowledgeable Metaverse designer or immersive virtual environment developer. • Preferably some exposure to HSE training, industrial training, or risk-related education

Experts will be contacted to review, assess, and comment on the proposed MetaHSE-UX framework. Based on the feedback obtained, the framework will be refined and finalized. The final version of the framework will be the result of combining the first draft and the changes suggested by experts during the validation process. The expert amendments are critical to ensure the framework meets the objectives and requirements of the research study. Following validation, documentation will be carried out in order to organize all data, decisions, and conclusions related to the research. All relevant information will be documented in a structured report format, making sure that all processes, findings, and justifications of the research are recorded and easily accessible. As all the information in the research is clearly documented, a completely documented report form will be the expected result of this phase.

IV. CONCLUSION

In conclusion, this study identifies the suitable components applicable for the UX evaluation of the HSE Metaverse environment. The identified UX components will serve as the foundation for constructing the MetaHSE-UX framework. Besides, the framework will be developed to serve as a practical guideline for designers and practitioners to create a positive and effective UX in Metaverse-based HSE training environments. The comprehensive development and validation of the MetaHSE-UX framework are yet to be completed and are planned for future work.

ACKNOWLEDGMENT

The author would like to thank Universiti Teknologi PETRONAS and the Ministry of Higher Education Malaysia for sponsoring this research under the Fundamental Research Grant Scheme (FRGS) funding.

REFERENCES

[1] C. Benson, I. C. Obasi, D. V. Akinwande, and C. Ile, "The impact of interventions on health, safety and environment in the process industry," Heliyon, vol. 10, no. 1, p. e23604, Jan. 2024, doi: 10.1016/j.heliyon.2023.e23604.

[2] A. Babalola, P. Manu, C. Cheung, A. Yunusa-Kaltungo, and P. Bartolo, "Applications of immersive technologies for occupational safety and health training and education: A systematic review," Saf Sci, vol. 166, p. 106214, Oct. 2023, doi: 10.1016/j.ssci.2023.106214.

[3] J. TAKALA, P. HÄMÄLÄINEN, N. NENONEN, K. TAKAHASHI, O. CHIMED-OCHIR, and J. RANTANEN, "Comparative analysis of the burden of injury and illness at work in selected countries and regions," Central European Journal of Occupational and Environmental Medicine, vol. 23, pp. 7–31, 2017, doi: https://www.nnk.gov.hu/cejoem/Volume23/Vol23No1-2/23_1-2_Article_01-.pdf.

[4] S. Gendler and E. Prokhorova, "Risk-Based Methodology for Determining Priority Directions for Improving Occupational Safety in the Mining Industry of the Arctic Zone," Resources, vol. 10, no. 3, p. 20, Mar. 2021, doi: 10.3390/resources10030020.

[5] N. F. Jamaludin, R. Shamsuddin, N. A. Mohd Nor, S. F. Sedek, and M. F. Hadipornama, "Application of Virtual Reality (VR) in Health, Safety and Environment Subject for Fire Exit Training Module," Journal of Advanced Research in Technology and Innovation Management, vol. 12, no. 1, pp. 1–15, Sep. 2024, doi: 10.37934/jartim.12.1.115.

[6] E. Enamorado-Díaz, J. A. García-García, M. J. Escalona-Cuaresma, and D. Lizcano-Casas, "Metaverse Applications: Challenges, Limitations and Opportunities - A Systematic Literature Review," Jun. 01, 2025, Elsevier B.V. doi: 10.1016/j.infsof.2025.107701.

[7] E. Youssef, M. Al Malak, and M. Bayoumy, "Metaverse in Education: A Survey-Based Investigation of Collaborative Learning and Immersive Experiences," Emerging Science Journal, vol. 8, pp. 379–395, Feb. 2025, doi: 10.28991/ESJ-2024-SIED1-022.

[8] W. Park, H. Heo, S. Park, and J. Kim, "A study on the presence of immersive user interface in collaborative virtual environments application," Symmetry (Basel), vol. 11, no. 4, Apr. 2019, doi: 10.3390/sym11040476.

[9] M. Hassenzahl, A. Wiklund-Engblom, A. Bengs, S. Hägglund, and S. Diefenbach, “Experience-Oriented and Product-Oriented Evaluation: Psychological Need Fulfillment, Positive Affect, and Product Perception,” Int J Hum Comput Interact, vol. 31, no. 8, pp. 530–544, Aug. 2015, doi: 10.1080/10447318.2015.1064664.

[10] S. Borsci, S. Federici, S. Bacci, M. Gnaldi, and F. Bartolucci, “Assessing User Satisfaction in the Era of User Experience: Comparison of the SUS, UMUX, and UMUX-LITE as a Function of Product Experience,” Int J Hum Comput Interact, vol. 31, no. 8, pp. 484–495, Aug. 2015, doi: 10.1080/10447318.2015.1064648.

[11] “ISO 9241-210:2019 - Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems.” Accessed: Nov. 18, 2025. [Online]. Available: https://www.iso.org/standard/77520.html

[12] M. Hassenzahl and Noam Tractinsky, “User experience-a research agenda.,” Behaviour & information technology, vol. 25, no. 2, pp. 91–97, 2006.

[13] R. Hartson and P. S. Pyla, “The UX Book: Process and Guidelines for Ensuring a Quality User Experience - Rex Hartson, Pardha S. Pyla - Google Books,” Elsevier Science, 2012, Accessed: Nov. 18, 2025. [Online]. Available: https://books.google.com/books/about/The_UX_Book.html?id=w4I3Y64SWLoC

[14] F. Cheng, S. Yu, S. Qin, J. Chu, and J. Chen, “User experience evaluation method based on online product reviews,” Journal of Intelligent & Fuzzy Systems, vol. 41, no. 1, pp. 1791–1805, Aug. 2021, doi: 10.3233/JIFS-210564.

[15] N. B. Nasir, J. Moon, and S. B. Kim, “Metaverse in Education: Insights from South Korea and Potentials for Malaysia,” in 8th International Conference on Software Engineering and Computer Systems, ICSECS 2023, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 84–88. doi: 10.1109/ICSECS58457.2023.10256325.

[16] F. Tang, X. Chen, M. Zhao, and N. Kato, “The Roadmap of Communication and Networking in 6G for the Metaverse,” IEEE Wirel Commun, vol. 30, no. 4, pp. 72–81, Aug. 2023, doi: 10.1109/MWC.019.2100721.

[17] C. Sun, J. Cao, R. Huo, L. Du, and X. Cheng, “Metaverse Applications in Energy Internet,” in 2022 IEEE International Conference on Energy Internet (ICEI), IEEE, Dec. 2022, pp. 7–12. doi: 10.1109/ICEI57064.2022.00007.

[18] W. Li, H. Huang, T. Solomon, B. Esmaeili, and L.-F. Yu, “Synthesizing Personalized Construction Safety Training Scenarios for VR Training,” IEEE Trans Vis Comput Graph, vol. 28, no. 5, pp. 1993–2002, May 2022, doi: 10.1109/TVCG.2022.3150510.

[19] L. G. Heyn et al., “Exploring facilitation in virtual simulation in nursing education: A scoping review,” PEC Innovation, vol. 3, p. 100233, Dec. 2023, doi: 10.1016/j.pecinn.2023.100233.

[20] K. Tcha-Tokey, E. Loup-Escande, O. Christmann, and S. Richir, “A questionnaire to measure the user experience in immersive virtual environments,” in Proceedings of the 2016 Virtual Reality International Conference, New York, NY, USA: ACM, Mar. 2016, pp. 1–5. doi: 10.1145/2927929.2927955.

[21] D. Shin, “How do users experience the interaction with an immersive screen?,” Comput Human Behav, vol. 98, pp. 302–310, Sep. 2019, doi: 10.1016/j.chb.2018.11.010.

[22] M. S. Anwar, J. Yang, J. Frnda, A. Choi, N. Baghaei, and M. Ali, “Metaverse and XR for cultural heritage education: applications, standards, architecture, and technological insights for enhanced immersive experience,” Virtual Real, vol. 29, no. 2, p. 51, Mar. 2025, doi: 10.1007/s10055-025-01126-z.

[23] B. Laugwitz, T. Held, and M. Schrepp, “Construction and Evaluation of a User Experience Questionnaire,” 2008, pp. 63–76. doi: 10.1007/978-3-540-89350-9_6.

[24] P. W. Jordan, B. Thomas, I. L. McClelland, and B. Weerdmeester, Eds., Usability Evaluation In Industry. CRC Press, 1996. doi: 10.1201/9781498710411.

[25] M. Minge, M. Thüring, I. Wagner, and C. V. Kuhr, “The meCUE Questionnaire: A Modular Tool for Measuring User Experience,” 2017, pp. 115–128. doi: 10.1007/978-3-319-41685-4_11.

[26] J. Nielsen, “Ten usability heuristics,” semanticscholar.org, 2005.

[27] A. Hinderks, “Design and evaluation of a short version of the user experience questionnaire (UEQ-S).,” International Journal of Interactive Multimedia and Artificial Intelligence, 2017.

[28] V. Davidavičienė, J. Raudeliūnienė, and R. Viršilaitė, “Evaluation of user experience in augmented reality mobile applications,” Journal of Business Economics and Management, vol. 22, no. 2, pp. 467–481, Feb. 2021, doi: 10.3846/jbem.2020.13999.

[29] S. Irshad, D. R. A. Rambli, and S. Sulaiman, “Design and implementation of user experience model for augmented reality systems,” in ACM International Conference Proceeding Series, Association for Computing Machinery, Nov. 2020, pp. 48–57. doi: 10.1145/3428690.3429169.

[30] A. J. A. Alnagrat, A. S. I. Mohamed, A. A. A. Akeel, R. C. Ismail, A. M. I. Noureldeen, and S. Z. S. Idrus, “An Evaluation of User Experience in a Virtual Semiconductor Laboratory Utilizing HMDs during Covid-19 Pandemic,” in Proceeding - 2023 IEEE 3rd International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering, MI-STA 2023, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 206–211. doi: 10.1109/MI-STA57575.2023.10169117.

[31] M. Dianatfar, S. Pöysäri, J. Latokartano, N. Siltala, and M. Lanz, “Virtual Reality-based safety training in Human-Robot Collaboration scenario: User experiences testing.”

[32] S. Macias-Velasquez, H. I. Medellin-Castillo, and A. Garcia-Barrientos, “New-user experience evaluation in a semi-immersive and haptic-enabled virtual reality system for assembly operations,” International Journal of Human Computer Studies, vol. 190, Oct. 2024, doi: 10.1016/j.ijhcs.2024.103312.

[33] J. Nielsen, “ Usability engineering. ,” Morgan Kaufmann, 1994.

[34] C. Cassandra, M. N. Masrek, and F. Aman, “Evaluating metaverse platforms for educational purposes: A heuristic evaluation study,” International Journal of Advances in Applied Sciences, vol. 13, no. 4, pp. 1019–1027, Dec. 2024, doi: 10.11591/ijaas.v13.i4.pp1019-1027.

[35] Y. Tang, H. Liang, and J. Zhan, “The application of metaverse in occupational health,” Front Public Health, vol. 12, Apr. 2024, doi: 10.3389/fpubh.2024.1396878.

[36] L. K. Cheng, M. H. Chieng, and W. H. Chieng, “Measuring virtual experience in a three-dimensional virtual reality interactive simulator environment: A structural equation modeling approach,” Virtual Real, vol. 18, no. 3, pp. 173–188, 2014, doi: 10.1007/s10055-014-0244-2.

[37] K. Tcha-Tokey, O. Christmann, E. Loup-Escande, G. Loup, and S. Richir, “Towards a model of user experience in immersive virtual environments,” Advances in Human-Computer Interaction, vol. 2018, 2018, doi: 10.1155/2018/7827286.

[38] S. Dong-Hee, B. Frank, and C. Hyunseung, “Exploring the user experience of three-dimensional virtual learning environments,” Behaviour & Information Technology, 2013.

[39] E. Brucker-Kley, J. Michot, T. Keller, M. Berger, and I. Hundozi, “Evaluating User Experience and Effectiveness of an Immersive Virtual Learning Environment,” in Proceedings - 2023 14th IIAI International Congress on Advanced Applied Informatics, IIAI-AAI 2023, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 234–240. doi: 10.1109/IIAI-AAI59060.2023.00055.

[40] L. Zhang, X. Dai, X. Xu, G. Shi, and Z. Tang, “User experience research on product design evaluation based on virtual reality prototyping,” in Proceedings of 2020 IEEE International Conference on Progress in Informatics and Computing, PIC 2020, Institute of Electrical and Electronics Engineers Inc., Dec. 2020, pp. 214–218. doi: 10.1109/PIC50277.2020.9350750.

[41] S. M. Rahman, N. H. Chowdhury, J. L. H. Bowden, and J. Carlson, “Metaverse platform attributes and customer experience measurement,” Journal of Retailing and Consumer Services, vol. 83, Mar. 2025, doi: 10.1016/j.jretconser.2024.104159.

[42] I. Rhiu, Y. M. Kim, W. Kim, and M. H. Yun, “The evaluation of user experience of a human walking and a driving simulation in the virtual reality,” Int J Ind Ergon, vol. 79, Sep. 2020, doi: 10.1016/j.ergon.2020.103002.

[43] W. Zuo, Z. Wu, Z. Lu, and Q. Liu, “The Evaluation Framework of User Experience in Educational Virtual Reality Games: The Case of ilab-x Platform,” in International Conference on Virtual Rehabilitation, ICVR, Institute of Electrical and Electronics Engineers Inc., May 2021, pp. 161–165. doi: 10.1109/ICVR51878.2021.9483836.

[44] R. K. Yin, “Case study research: Design and methods (4th Ed.). Thousand Oaks, CA: Sage.,” CJAR, 2009, vol. 14, pp. 69–71, doi: 10.33524/cjar.v14i1

Prototype of Chronic Obstructive Pulmonary Disease Therapeutic Devices based on Hall Effect Sensor

Anggita Dewi Puspita[1st], Nur Afifah Zen[2nd], Dodi Zulherman[3rd]
[1st]*Departement of Biomedical Engineering*, *Telkom University*, Purwokerto, Indonesia
[2nd]*Department of Telecommunication Engineering*, *Telkom University*, Purwokerto, Indonesia
[3rd] *Center of Excellence for Healthcare and Well-Being Technologies*, *Telkom University*, Purwokerto, Indonesia
*Corresponding author: zulherman@telkomuniversity.ac.id

***Abstract*—Poor adherence to inhaler use remains a significant challenge in the treatment of Chronic Obstructive Pulmonary Disease (COPD), resulting in reduced therapy effectiveness and increased risk of exacerbations. To address this issue, this study presents an innovative inhaler prototype that integrates an ESP32 microcon-troller and a Hall-effect sensor to automatically detect inhaler actuation. Usage data is transmitted to the Blynk platform in real time, enabling activity monitoring and reminders when patients miss their therapy schedule. The system is powered by a Lithium-Polymer battery with a TP4056 charging module to support mobility. Hardware design, firmware development, and functional testing were conducted to evaluate performance. The results indicate that the prototype accurately detects inhaler activity and transmits data consistently to the monitoring application. This work demonstrates the potential of smart inhalers to improve treatment adherence among COPD patients and to assist medical professionals with remote monitoring.**

***Index Terms*—smart inhaler, ESP32, Hall effect sensor, IoT, Blynk, therapy.**

I. INTRODUCTION

Chronic Obstructive Pulmonary Disease (COPD) is a progressive respiratory disorder characterized by persis-tent airflow limitation. According to the World Health Organization (WHO), COPD is one of the leading causes of death globally and continues to show increasing prevalence every year [1]. In Indonesia, COPD ranks among the top causes of mortality and contributes signifi-cantly to the national healthcare burden due to prolonged treatment duration and recurrent hospital admissions [2]. The management of COPD involves a combination of pharmacological and non-pharmacological therapies aimed at relieving symptoms, improving lung function and preventing exacerbations. According to the Global Initiative for Chronic Obstructive Lung Disease (GOLD), pharmacological treatment primarily consists of bron-chodilators and anti-inflammatory agents delivered via various routes, including Metered Dose Inhalers (MDIs), Dry Powder Inhalers (DPIs), soft mist inhalers, and nebulizers [12]. The most commonly prescribed med-ications include Short-Acting Beta-Agonists (SABA), Long-Acting Beta-Agonists (LABA), Long-Acting Mus-carinic Antagonists (LAMA) and Inhaled Corticosteroids (ICS), either as monotherapy or in combination. Each inhaler type requires specific inhalation techniques, and improper use significantly reduces therapeutic effective-ness [13]. In more severe cases, additional interventions such as oxygen therapy, pulmonary rehabilitation, vac-cination, smoking cessation programs and, in selected patients, surgical procedures may be required [12], [13]. Inhalation therapy is the primary clinical approach for COPD management because it delivers medication di-rectly to the respiratory tract, resulting in rapid and effec-tive pharmacological action [3]. However, adherence to inhaler therapy remains low, with non-compliance rates reported between 50–80 percent [3], [4]. Poor adher-ence has been linked to incorrect inhalation techniques, lack of patient awareness, and the absence of reminder or monitoring systems, ultimately reducing treatment effectiveness and increasing the risk of exacerbations and hospitalization [3], [4]. Recent advancements in digital health and Internet of Things (IoT) technology have enabled real-time monitoring of patient medication behavior. Smart inhaler systems have demonstrated the ability to improve therapy adherence by automatically recording inhaler use, sending notifications, and provid-ing healthcare professionals with remote access to usage data [5],

[6], [7], [8]. Although smart inhalers have been successfully developed in several international studies, research and implementation in Indonesia remains very limited. To address this gap, this study proposes an innovative inhaler prototype based on an ESP32 micro-controller and a Hall-effect sensor, integrated with the Blynk IoT platform. The ESP32 serves as the processing and connectivity module for transferring inhaler usage data to the cloud [9], while the Hall-effect sensor detects magnetic-field changes during inhaler actuation [10]. The recorded data are visualized through the Blynk mobile application to support remote monitoring and improve adherence to inhalation therapy for COPD patients [11].

II. METHODOLOGY

A. Research Design

This study adopts a research-and-development approach to the design and implementation of an IoT-based ther-apeutic device. The process includes hardware assem-bly, firmware programming, IoT integration, and system evaluation. The Hall effect sensor operates by detecting changes in magnetic field strength. When a permanent magnet mounted on the inhaler actuator moves toward the sensor, the resulting voltage shift is read by the ESP32 microcontroller as a digital HIGH or LOW signal. This change indicates whether the inhaler is pressed.

B. Research Workflow

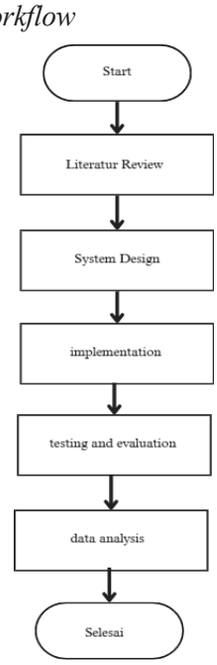


Fig. 1. Flowchart system.

Figure 1 presents this research as a development-focused study on the design and development of an innova-tive therapy device system. The stages are carried out iteratively with a hardware and software engineering approach. The methods used include:

1. Literature Study: Reviewing references related to COPD, inhalation ther-apy, and the implementation of IoMT technology in medical devices.
2. System Design: Compilation of block diagrams, hardware schematics, and software workflows.
3. Prototype Implementation: Assembling the device us-ing ESP32 and Hall effect sensors and integrating it with the IoMT platform.
4. Testing and Evaluation: Conduct functionality tests and device performance evaluations in monitoring inhaler use.
5. Data Analysis: Analyzing test data using quantitative methods to measure the effectiveness and accuracy of the device.

C. Hall Effect Sensor

The Hall effect is a physical phenomenon that occurs when an electric current flows through a thin semi-conductor material in a magnetic field perpendicular to the current. This interaction generates a transverse voltage, known as the Hall voltage, whose magnitude is proportional to the strength of the magnetic field. Hall effect sensors exploit this phenomenon to accurately detect magnetic fields and their variations, producing either digital pulses or analog signals depending on the magnetic field intensity.

In flow monitoring devices, Hall-effect sensors are commonly integrated with a magnetic element mounted on a rotating mechanical component, such as an impeller or rotor. When fluid flows through the system, the rotation of the component causes periodic variations in the magnetic field, which are detected as pulses by the sensor. The frequency of these pulses corresponds to the flow velocity, allowing the system to measure flow rates with high sensitivity and low power consumption while maintaining durability due to the absence of friction-based moving electrical contacts [10].

This operating principle can be adapted for

smart in-haler applications. The Hall-effect sensor can be used to monitor inhalation activity via two primary mechanisms. First, it can detect inhaler actuation by sensing changes in the magnetic field when the user presses or rotates a component of the inhaler equipped with a small magnet. This enables non-invasive detection of the timing and frequency of inhaler usage. Second, the sensor can be configured to estimate inhaled airflow by pairing it with a miniature mechanical element—such as an impeller or ventilated flap—that moves in response to inhaled air. The corresponding magnetic field variations can then be analyzed to determine airflow velocity and inhaled volume.

The use of Hall effect sensors in smart inhalers offers several advantages, including high durability, high mag-netic displacement sensitivity, and the ability to produce a continuous output for further signal processing. How-ever, implementing airflow measurement requires careful mechanical design and proper calibration to ensure reli-able data. When integrated with a microcontroller such as an ESP32 and an Internet of Things (IoT) platform, the Hall effect sensor enables real-time recording of inhaler usage data and automated delivery of this information to monitoring systems, supporting improved patient adher-ence to respiratory therapy.

III. Results and Discussion

Testing was conducted to ensure that the innovative ESP32-based inhaler prototype and the Hall-effect sensor functioned as intended, particularly to detect inhaler actuation, record usage by multiple respondents, and provide consistent data collection. This testing phase is still in its early stages as the device is still undergoing development and refinement.

A. Prototype Design

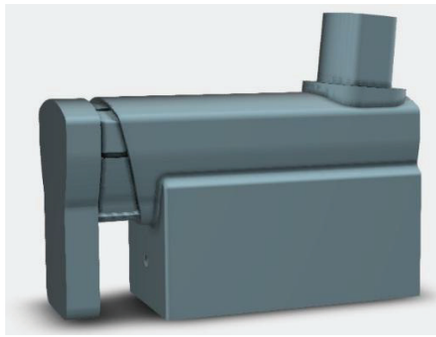


Fig. 2. Design 3D smart inhaler.

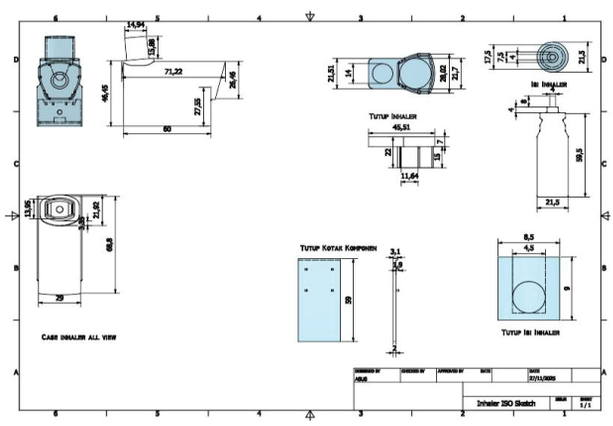


Fig. 3. ISO sketch.

As illustrated in Figs. 2 and 3, the prototype was designed with a 3D-printed casing that accommodates the ESP32, Hall-effect sensor, and inhaler magnet. The design focuses on the mechanical integration between the inhaler actuation and sensor detection. This system design integrates the Hall sensor and magnet modules into the inhaler's physical structure.

The Hall sensor is strategically placed on the back or outer side of the device that includes the mechanical actuator of the inhaler. Meanwhile, a small magnet is placed on the moving part of the inhaler (usually on the cap or the part that the user presses), so that when the user presses the inhaler (as indicated by the black arrow pointing down), the magnet will move closer to the Hall sensor and change the detected magnetic field reading. Movement is the main determinant that the inhaler is being used. When the inhaler button is pressed:

1) The magnet approaches the Hall effect sensor, producing a significant magnetic field reading.
2) The Hall sensor sends a LOW signal to the ESP32's digital pin.
3) The ESP32 processes this signal and

immediately sends usage data to the Blynk Cloud platform via the Wi-Fi network.

4) The data is displayed immediately in the Blynk app on the user's smartphone or the medical staff's device as real-time monitoring information.

B. Circuit Design

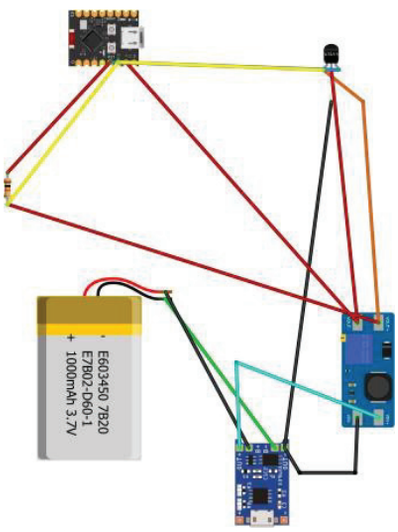


Fig. 4. Smart inhaler circuit design.

Figure 4 illustrates the Smart Inhaler Circuit Design. hardware configuration of the Smart Inhaler system, which integrates an ESP32 microcontroller, a Hall-effect sensor, a LiPo charging module, and a 3.7-V LiPo battery. The ESP32 serves as the primary processing unit responsible for acquiring sensor data and managing wire-less communication. The power The system is supplied by a 1000-mAh LiPo battery, which is connected through a dedicated charging module to ensure safe and stable power regulation.

The Hall-effect sensor detects the displacement of the magnetic actuator within the inhaler. Its output pin is interfaced with one of the ESP32's digital GPIO pins, enabling the microcontroller to register changes in magnetic field intensity during inhaler actuation. The VCC and GND terminals of the sensor are connected to the ESP32's 3.3-V and GND rails, respectively.

The LiPo charging module functions as both the battery charging interface and the power distribution unit. Its B+ and B- terminals are connected directly to the LiPo battery, while the OUT+ and OUT- terminals supply regulated power to the ESP32. This configuration allows the device to operate while charging and ensures continuous system stability.

In general, the wiring configuration ensures reliable signal acquisition from the Hall-effect sensor, stable power delivery from the LiPo battery, and seamless operation of the ESP32 as the core controller of the Smart Inhaler platform.

C. System Block Diagram

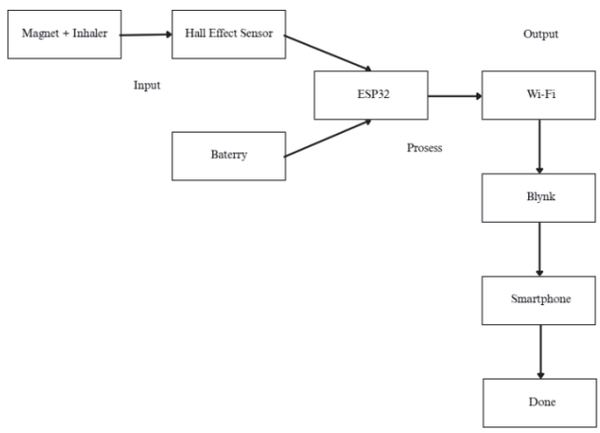


Fig. 5. Block system

The block diagram in Fig. 5 shows that the innovative inhaler system integrates hardware for actuation detec-tion and software for remote monitoring. The Hall effect sensor detects the presence of a magnetic field when the inhaler is pressed. The ESP32 processes the signal and sends the recorded event to the Blynk cloud via Wi-Fi. The Blynk application then displays notifications and usage history on the user's device.

D. Hall Effect Sensor Testing

Three respondents participated in a one-week simulation of inhaler use at 1 p.m. and 9 p.m. Respondents performed the specified number of actuations, and the detection results were recorded. At each specified time, Bylnk would send a notification to the user's smartphone to prompt immediate use of the inhaler. If at least 1 data point were not detected on virtual pin 1, it would send a notification to the user and the developer as a reminder.

E. Time Logging Consistency

In addition to the number of actuations, usage time was recorded manually for comparison with the sensor detection sequence. The results show that the system consistently represents the usage sequence based on the time

of occurrence. This is important to support data integration into monitoring platforms such as Blynk in the next stage of development.



Fig. 6. Weekly inhaler usage chart for Respondent 1 with reminder notification enabled, showing consistent adherence behavior.



Fig. 7. Weekly inhaler usage chart for Respondent 2 without reminder notification, indicating irregular inhaler usage.

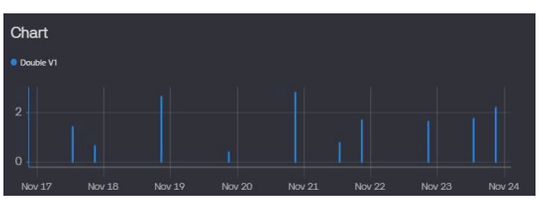


Fig. 8. Weekly inhaler usage chart for Respondent 3 with reminder notification and delayed response behavior..

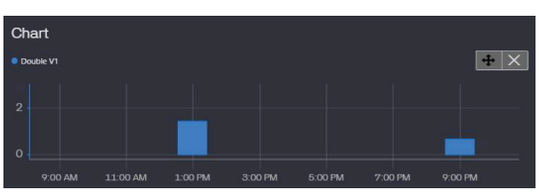


Fig. 9. Time chart usage.

As shown in Figs. 6–9, the graphs illustrate inhaler usage patterns over a one-week observation period. Respondent 1, who received reminder notifications, achieved an inhaler usage accuracy of 91.67%, indicating high adherence supported by the reminder system.

Respondent 2, who did not receive reminder notifications, demonstrated an inhaler usage accuracy of 66.67%, reflecting more irregular usage behavior and lower consistency compared to Respondent 1.

Respondent 3 achieved an inhaler usage accuracy of 71.43%. Although reminder notifications were provided, several scheduled inhalation times were missed, particularly during daytime sessions, indicating partial adherence and delayed response behavior.

Overall, these results confirm that reminder notifications improve inhaler usage adherence; however, individual user behavior remains a significant factor influencing therapy compliance.

TABLE I
INHALER USAGE DURING ONE WEEK RESPONDENT 1

Tanggal	Jam	Usage
22/11/2025	13:00:00	
22/11/2025	21:05:00	✓
23/11/2025	13:02:00	✓
23/11/2025	21:03:00	✓
24/11/2025	13:01:00	✓
24/11/2025	21:00:00	✓
25/11/2025	13:06:00	✓
25/11/2025	21:07:00	✓
26/11/2025	13:02:00	✓
26/11/2025	21:03:00	✓
27/11/2025	13:01:00	✓
27/11/2025	21:04:00	✓

TABLE II
INHALER USAGE DURING ONE WEEK RESPONDENT 2

Tanggal	Jam	Usage
22/11/2025	13:05:00	✓
22/11/2025	21:00:00	
23/11/2025	13:02:00	✓
23/11/2025	21:03:00	✓
24/11/2025	13:00:00	
24/11/2025	21:04:00	✓
25/11/2025	13:00:00	✓
25/11/2025	21:00:00	
26/11/2025	13:02:00	✓
26/11/2025	21:00:00	
27/11/2025	13:05:00	✓
27/11/2025	21:04:00	✓

TABLE III
INHALER USAGE DURING ONE WEEK RESPONDENT 3

Tanggal	Jam	Usage
18/11/2025	13:00:00	✓
18/11/2025	21:00:00	✓
19/11/2025	13:00:00	
19/11/2025	21:03:00	✓
20/11/2025	13:00:00	
20/11/2025	21:00:00	✓
21/11/2025	13:00:00	
21/11/2025	21:00:00	✓
22/11/2025	13:00:00	✓
22/11/2025	21:00:00	✓
23/11/2025	13:10:00	
23/11/2025	21:03:00	✓
24/11/2025	13:09:00	✓
24/11/2025	21:08:00	✓

Table I presents inhaler usage data for Respondent 1, who received automated reminder notifications via the Blynk application. During the one-week observation period, the respondent showed a generally consistent inhaler usage pattern. Although several scheduled inhalation times—particularly during daytime sessions—were missed, most inhaler actuations were successfully recorded. Evening usage demonstrated higher adherence compared to daytime usage. The implementation of reminder notifications contributed to improved therapy compliance, as indicated by frequent and timely inhaler usage events.

Table II summarizes inhaler usage data for Respondent 2, who did not receive any reminder notifications and served as the control subject. The data reveal a more irregular inhaler usage pattern, with multiple missed scheduled inhalation times, especially during the early phase of the observation period. Although some inhaler usage events were recorded, the absence of reminder notifications resulted in lower consistency and adherence when compared to Respondent 1. This finding emphasizes the role of reminder systems in supporting regular therapy behavior.

Table III presents inhaler usage data for Respondent 3, who received reminder notifications but did not always respond immediately. The results indicate an intermittent inhaler usage pattern, with several missed scheduled inhalation times despite the presence of reminders. This behavior represents a semi-adherent usage pattern, reflecting real-world patient conditions in which reminders may be received but not consistently acted upon.

Overall, the comparison of Tables I–III demonstrates that reminder notifications have a positive effect on inhaler usage consistency. Respondents who received reminders exhibited higher adherence levels compared to the control subject, while partial compliance was observed in the semi-intervention group. These results support the effectiveness of IoT-based reminder systems in enhancing therapy adherence and monitoring inhaler usage for COPD patients.

F. Overall System Performance

Overall, initial testing shows that the smart inhaler can detect functional inhaler use using Hall-effect sensors. The device's performance was found to be stable when tested on several individuals with varying usage habits. The results indicate that the system is ready to move to integration with the IoT platform, further testing, and the development of automatic notifications for users. However, more users are needed in the trial so the data accuracy can be compared again.

IV. CONCLUSION

This study successfully designed and evaluated a smart inhaler prototype for COPD therapy monitoring using an ESP32 microcontroller and a Hall-effect sensor. The developed system was able to reliably detect inhaler actuation events, record usage timestamps, and transmit data to an IoT-based monitoring platform in real time.

Based on weekly usage data, Respondent 1, who received reminder notifications, achieved an inhaler usage accuracy of 91.67%**,** indicating high adherence supported by the reminder system. Respondent 2, who did not receive any reminder notifications, showed a lower accuracy of 66.67%**,** reflecting more irregular and inconsistent usage behavior. Meanwhile, Respondent 3 achieved an inhaler usage accuracy of 71.43%**.** Although reminders were provided, several scheduled inhalation times were missed, indicating partial adherence and delayed response behavior.

These results demonstrate that reminder notification systems positively influence inhaler usage consistency and therapy adherence. However, individual variability in user response remains a significant factor affecting overall compliance.

Overall, the findings confirm the feasibility of the proposed smart inhaler system not only in detecting inhaler usage events but also in capturing user adherence behavior under different reminder conditions. Future work will focus on expanding the number of participants, extending the monitoring duration, optimizing power consumption, and enhancing notification strategies to improve clinical applicability for COPD patient monitoring**.**

ACKNOWLEDGMENT

The authors would like to express their sincere grat-itude to the Research and Community Service Insti-tute (LPPM) of Telkom University Purwokerto and the Biomedical Engineering Study Program for their guid-ance, facilitation, and institutional support throughout the implementation of this research.

REFERENCES

[1] World Health Organization, "Chronic obstructive pulmonary disease (COPD)," 2023.

[2] Kementerian Kesehatan Republik Indonesia, "Ayo Sehat Ke-menkes," 2018.

[3] R. D. Restrepo et al., "Medication adherence in chronic ob-structive pulmonary disease patients," International Journal of COPD, 2008.

[4] C. Yin et al., "Investigating influencing factors of medication adherence to inhalers in COPD patients," 2024.

[5] A. Nayan et al., "IoT Based Smart Inhaler for Context-Aware Service Provisioning," ICAICT, 2020.

[6] A. H. Y. Chan et al., "Digital Inhalers for Asthma or Chronic Obstructive Pulmonary Disease," Drugs and Therapy Perspec-tives, 2021.

[7] G. Greene and R. W. Costello, "Personalizing medicine – could the smart inhaler," Expert Opinion on Drug Delivery, 2019.

[8] S. J. van de Hei et al., "Effectiveness, usability and accept-ability of a smart inhaler programme," BMJ Open Respiratory Research, 2022.

[9] Espressif Systems, ESP32 Technical Reference Manual, 2022.

[10] M. Egerer et al., "Hall Effect Sensor-Based Low-Cost Flow Monitoring Device," IEEE Sensors Journal, 2024.

[11] Blynk Inc., "Introduction to Blynk," 2023.

[12] Global Initiative for Chronic Obstructive Lung Disease (GOLD), Global Strategy for the Diagnosis, Management, and Prevention of COPD, 2024 Report.

[13] F. Lavorini et al., "What to consider before prescribing inhaled medications: a practical approach for physicians," European Respiratory Journal, 2019.

Managerial Pragmatism and Tactical Responses in Requirements Traceability: An Interview-Based Study

Raditya Pratama, Auzan Taqi Syafega, Fausta Irsyad Ramadhan, Anisah Mutmainah, Feby Artwodini Muqtadiroh
Department of Information Systems, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
Emails: radityapratama2077@gmail.com, at.syafega@gmail.com, faustairsyad747@gmail.com, mutmainaha390@gmail.com, feby@its.ac.id

***Abstract*—Requirements Traceability (RT) is widely recognized as a key practice for supporting change management, impact analysis, and quality assurance in software projects. However, empirical studies repeatedly report that RT is neglected or only partially implemented in industrial settings, particularly in Agile and fast-paced environments. This paper presents an interview based qualitative study that explores how practitioners decide when and how to apply RT, what barriers they face, and how they respond tactically to these barriers in practice. We conducted in-depth semi-structured interviews with five practitioners occupying roles such as project manager, system analyst, and UI/UX designer across healthcare, government, and product-based organizations. Using a hybrid coding scheme with motivational (MOT), barrier (BAR), strategy (STRAT), and tool (TOOL) categories, and analyzing the data with ATLAS.ti, we developed a network model of RT decision-making. The findings reveal that (1) RT adoption is driven by pragmatic managerial motivations, including external audit, internal KPI pressure, and contract locking, rather than abstract quality ideals; (2) practitioners employ tactical strategies such as *finish first, track later*, administrative gaming of ticket systems, and visual validation to cope with emergencies, immature processes, and stakeholder inconsistency; and (3) the effectiveness of RT tools depends on their alignment with these motivations, with issue trackers, spreadsheets, dashboards, and visual design tools forming a heterogeneous tool ecosystem. The study contributes to a tactical response model that explains why RT is selectively sacrificed or emphasized in practice and offers implications for designing RT processes, incentives, and tool chains that better reflect real-world decision dynamics.**

***Keywords*—requirements traceability, qualitative study, decision making factors, empirical software engineering, change management**

I. Introduction

Requirements Traceability (RT), the ability to track the "life" of each requirement backward to its source and forward to design, implementation, and tests, is a foundational concept in Requirements Engineering. Prior work has established that systematic traceability supports change management, impact analysis, and requirements of verification across the software lifecycle [2], [4], [6]. In principle, RT should be even more critical in volatile and dynamic environments where requirements evolve frequently.

In practice, however, there is a persistent gap between the normative importance of RT and its actual adoption in industrial projects. Empirical and mapping studies highlight that RT is often perceived as an overhead, poorly aligned with Agile values, and fragile under continuous delivery pressures [8], [9], [11]. In Agile and DevOps contexts, teams frequently prioritize delivery speed and minimal documentation, which exacerbates the tension between maintaining trace links and meeting deadlines [5].

Existing research predominantly focuses on tools, automation techniques, and technical methods for RT, including issue-based traceability and machine learning based link recovery [1], [13]. While these works advance the "how" of traceability, there remains limited qualitative understanding of the human and organizational dynamics that shape the "why" behind RT practices: why practitioners sometimes ignore RT, why they selectively document certain decisions, and how they navigate conflicts between traceability and other project priorities.

This paper addresses this gap by examining RT as a situated managerial and tactical practice. Through semi-structured interviews with five practitioners across different industrial contexts, we investigate how decisions about RT are made, which non-technical factors influence those decisions, and what tactical responses emerge when RT conflicts with emergencies, stakeholder politics, or immaturity of organizational processes.

A. Research Questions

We structure our study around the following research questions:

- RQ1: What are the most decisive non-technical factors that influence practitioners' decisions to track or not track requirements?
- RQ2: How do project teams decide the extent to which RT is applied, especially when facing sudden requirement changes or crises?
- RQ3: What problems and barriers most often hinder effective RT implementation or lead to its failure?

- RQ4: To what extent do support tools and formal documentation help or hinder teams in making quality decisions regarding RT?

To answer these questions, we developed a coding schema that distinguishes motivational drivers (MOT), barriers (BAR), tactical decision strategies (STRAT), and tool usage patterns (TOOL). Using ATLAS.ti network views, we visualize the causal relations among these codes, resulting in a tactical response model and a map of managerial motivations and tool choices.

The main contributions of this paper are:

- A qualitative model of managerial pragmatism in RT adoption, showing how audit, KPI, and contractual considerations drive traceability behavior.
- A tactical response model that explains why practitioners sometimes delay, manipulate, or reframe RT when facing emergencies, immature SOPs, or stakeholder inconsistency.
- A characterization of tool ecosystems as reflections of organizational motivation, highlighting the complementary roles of issue trackers, spreadsheets, dashboards, and visual design tools.
- Practical implications for project managers and analysts who seek to design RT practices that are both realistic and sustainable in industrial environments.

II. Background and Related Work

A. Requirements Traceability in Practice

RT is widely discussed as a desirable property of software development processes. Classical work emphasizes trace links between requirements, architecture, code, and tests to support systematic change management and impact analysis [4]. More recent systematic reviews report that RT contributes to cost reduction and maintainability in long-lived systems [2], [6].

Despite these benefits, empirical studies consistently show that RT is difficult to sustain in practice. In Agile settings, teams often perceive RT as a non-Agile administrative workload that conflicts with lightweight documentation principles [8], [9]. Mapping studies on RT in industry reveal that many organizations maintain only partial traceability, prioritize specific artifacts (e.g., from user stories to test cases), or rely heavily on informal communication rather than explicit trace links [3], [11], [13].

B. Factors and Barriers Affecting RT

Prior research identifies human, organizational, and technological factors that influence RT practices. Human factors include practitioners' understanding of RT value, perception of its usefulness versus burden, and collaboration patterns within teams [7], [9]. Organizational factors cover management support, process maturity, and alignment with development methodologies, particularly Agile methods that emphasize minimal documentation [8], [10].

Technological factors concern the availability and integration of RT supporting tools, such as issue trackers, requirements management systems, and semi-automated trace link recovery approaches [1], [12], [13]. However, semi-automated tools often suffer from low precision or recall, requiring substantial manual confirmation, which can paradoxically increase the perceived overhead [12].

In continuous software development environments, documentation challenges amplify RT difficulties. [5] report that teams frequently sacrifice documentation quality and traceability to accelerate deployment, creating tension between speed and record keeping. Existing studies, however, rarely model the tactical decision processes through which practitioners resolve such tensions in specific situations.

III. Research Methodology

A. Research Design

This study adopts an exploratory qualitative design to capture the nuanced decision-making processes behind RT practices. Rather than attempting to measure RT coverage quantitatively, we focus on how practitioners interpret traceability, which motivations they articulate, and how they describe trade-offs in concrete project situations. This orientation aligns with prior qualitative work in empirical software engineering that seeks to explain rather than merely describe tool usage and process adherence.

B. Participants and Contexts

We recruited five practitioners using purposive and snowball sampling. Inclusion criteria required that participants hold roles such as project manager, business analyst, system analyst, or similar; have at least three years of professional experience in software-related projects; and be directly involved in requirements and traceability-related decisions.

The five informants span multiple organizational contexts:

- N1: A manager responsible for Smart Remote Services in a multinational healthcare company, focusing on technical support and VPN connectivity for medical devices.
- N2: A full-stack developer acting as project manager for government pilot projects, including a diplomatic communication system.
- N3: A software engineer who also serves as project manager for property-related marketing systems using a Scrum-based approach.
- N4: A UI/UX designer and former system analyst responsible for translating user needs into diagrams and mock-ups.
- N5: A system analyst for an Agile internal product team whose work is tightly coupled with KPI-driven ticket tracking.

Across these contexts, "traceability" ranges from tracking VPN support activities to linking requirements with design artifacts and sprint tickets. This diversity allows us to examine RT as a family of practices rather than a single uniform process.

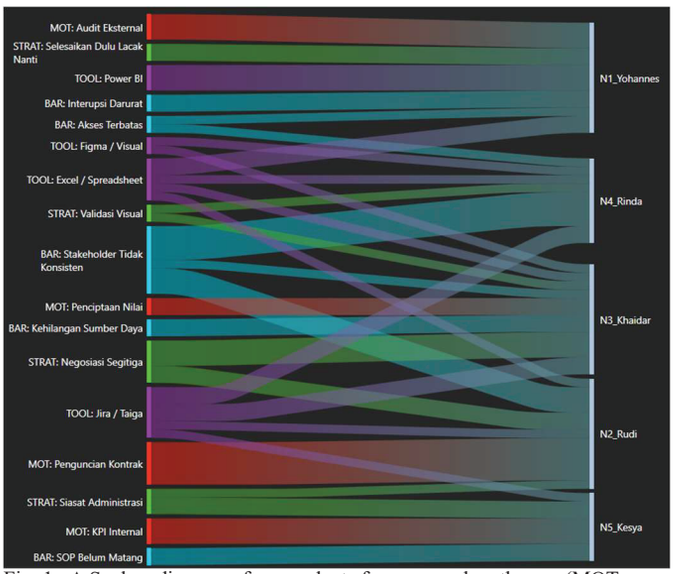


Fig. 1. A Sankey diagram of respondents focusses on key themes (MOT, BAR, STRAT, TOOL).

C. Data Collection

We conducted semi-structured interviews via online meetings and messaging channels over a period of roughly two weeks. Each interview lasted approximately 20–30 minutes. The interview guide comprised ten open-ended questions aligned with the four research questions, covering current RT processes, perceived benefits and burdens, decision-making under conflicting priorities, organizational and human barriers, and perceived effectiveness of tools.

All sessions were recorded with consent. Audio and chat logs were summarized and partially transcribed for analysis. We focused on segments where informants described concrete episodes of conflict, negotiation, or deviation from prescribed RT procedures, as these episodes provide insight into the underlying decision logics.

D. Data Analysis and Coding Scheme

Data were analyzed using ATLAS.ti Desktop through a systematic hybrid coding approach that integrated deductive structure with inductive discovery. To ensure analytical rigor and mitigate single-researcher bias, the process followed three distinct phases:

1. Deductive Coding (A Priori Framework): We established a provisional codebook derived directly from the Research Questions. This created four "container" categories which is MOT (Motivation), BAR (Barrier), STRAT (Strategy), and TOOL to ensure the analysis remained strictly aligned with the research objectives.
2. Inductive Coding (Open Coding): During the line-by-line analysis of interview transcripts, specific sub-codes were allowed to emerge inductively. For instance, while "Barrier" was a pre-set category, specific instances like "Interruption" or "Immature SOP" emerged from the participants' own descriptions. This phase captured the nuanced "tactical" behaviors that a purely deductive approach would have missed.
3. Bias Mitigation via Investigator Triangulation: To address the potential for subjective interpretation, we employed investigator triangulation. The primary author performed the initial coding. Subsequently, a second researcher reviewed a random sample of the coded segments and the generated network views to verify consistency. Discrepancies in code definitions or causal links (e.g., confirming that "Emergency" was indeed the *cause* of "Delay") were resolved through consensus discussions. The use of ATLAS.ti network views further mitigated bias by forcing a transparent audit trail, ensuring that every edge in our Tactical Response Model was backed by specific empirical quotations.

IV. Results

A. Pragmatic Motivations for RT Adoption (RQ1)

The analysis shows that practitioners rarely adopt RT solely because of abstract quality standards. Instead, RT is embraced when it directly supports pragmatic managerial goals. Three dominant motivational patterns emerge.

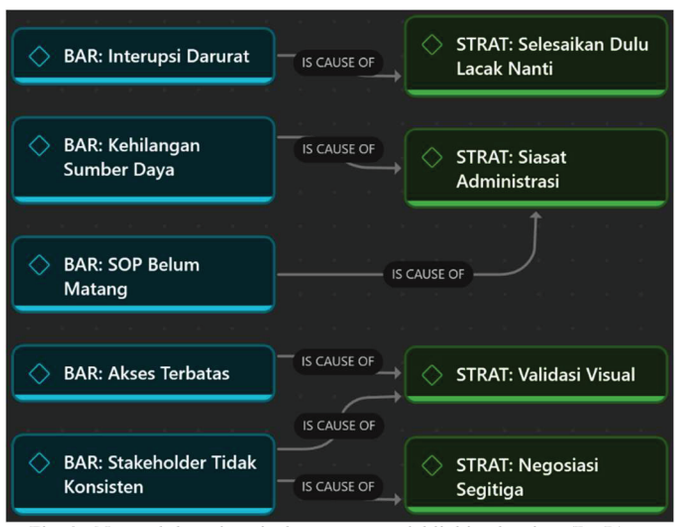


Fig. 2. Network-based tactical response model linking barriers (BAR) to strategies (STRAT).

1) Factual Accountability Under External Audit: In high-risk environments such as healthcare, RT functions as a legal and accountability mechanism. N1 emphasized that detailed logging of support activities and VPN connectivity status is essential for demonstrating factual evidence to global teams and auditors. Here, traceability is less about software artifact links and more about reconstructing a reliable history of actions for patient safety and compliance, consistent with prior observations on audit-driven RT [3].

2) Performance Mandates and KPI Pressure: In N5's Agile product team, RT is tightly coupled with individual performance metrics. Compliance with filling user stories and tasks in an issue tracker is monitored as part of KPI assessment. This configuration transforms RT into a form of self-discipline: failure to update tickets promptly has direct consequences for personal evaluation. Under such conditions, traceability is perceived as non-negotiable administrative work because it is structurally integrated into performance management.

3) Contract Locking and Scope Management: For project managers in public sector projects (N2), trace documents serve as defensive artifacts that "lock in" agreed requirements. When clients attempt to introduce additional features without adjusting budget or schedule, tracking records provide a basis for negotiating scope, time, and cost. This use of RT as a contract-protection tool complements the literature on traceability as a

means to control scope creep and manage change requests [3], [10].

B. Tactical Responses to Barriers (RQ2 and RQ3)

Barriers to RT are predominantly non-technical. Rather than simply failing to follow a prescribed process, practitioners describe deliberate tactics for navigating conflicting constraints, resulting in a tactical response model (Fig. 2).

1) Emergency-Driven Delay: Finish First, Track Later: In critical incidents, such as high-priority medical emergencies, N1 consciously suspends real-time recording in order to re- store service as quickly as possible. Trace logs are populated retrospectively once the crisis is resolved. This *finish first, track later* strategy is triggered by emergency interruptions and reflects a rational prioritization of service availability over trace completeness. From a traceability perspective, the resulting gaps are not mere negligence but calculated trade-offs under extreme time pressure.

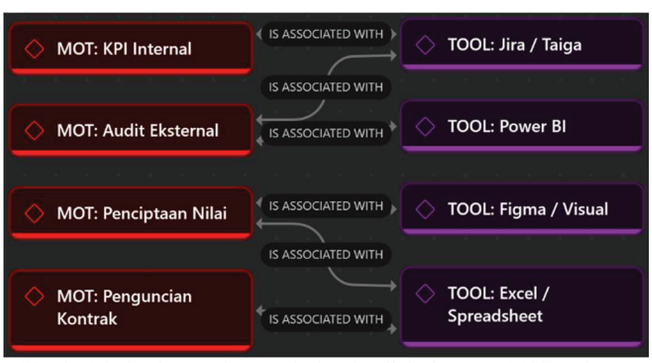


Fig. 3. Motivation–tool network showing links between MOT codes and TOOL categories.

2) Administrative Gaming Under Immature SOPs: N5 reported that immature estimation procedures made it difficult to predict realistic ticket durations. Because KPIs penalized overdue tasks, team members sometimes closed tickets prematurely and opened new ones with updated scopes or time frames. This *administrative tactic* preserves a "green" dashboard while obscuring the true history of work. This behavior illustrates how strong KPI-driven motivations, combined with immature processes, can lead to strategic manipulation of trace data.

3) Triangular Negotiation with Inconsistent Stakeholders: N2 and N3, who frequently interact with external clients, rely on trace records as negotiation instruments when stakeholders repeatedly change their minds. When confronted with new feature requests, they use documented requirements and prior agreements to argue for proportional adjustments to schedule or budget. This *triangular negotiation* strategy uses RT ar- tifacts as evidence in three-way negotiations among clients, internal teams, and management.

4) Visual Validation Under Limited Access and Ambiguity: N4 highlighted challenges in eliciting clear requirements from users who have limited time or difficulty articulating their needs. In response, she employs *visual validation* by constructing mock-ups and diagrams in tools such as Figma and using them as concrete negotiation artifacts. These visuals function as a form of traceability between abstract requirements and de- sign decisions, especially valuable when textual specifications are insufficient.

C. Tool Ecosystem as Reflection of Motivation (RQ4)

The network analysis suggests that tool choices are not neutral technical decisions but mirror underlying motivations (Fig. 3). Several patterns are noteworthy.

First, strong KPI orientation correlates with intensive use of centralized issue trackers such as Jira or Taiga. In N5's context, these tools provide granular visibility into user stories, tasks, and deadlines, enabling managers to monitor compliance and output. Here, tools are configured primarily as performance control instruments rather than as flexible documentation en- vironments.

Second, external audit motives encourage the use of analytics dashboards, such as Power BI, over raw logs. N1's organization uses aggregated visualizations to communicate connectivity status and intervention histories to global stakeholders. These tools help transform detailed trace data into audit-ready evidence, complementing classical RT links.

Third, spreadsheets remain pervasive as low-friction, adapt- able artifacts. N1, N3, and N4 use spreadsheets to manage trace matrices, summarize interactions, and support negoti- ations with clients who prefer tabular representations. This aligns with prior observations that simple, general-purpose tools often outperform sophisticated RT solutions when user familiarity and flexibility are critical [3].

Finally, visual design tools such as Figma play a crucial role in bridging misunderstandings with non-technical stakehold- ers. They support RT indirectly by anchoring discussions and decisions in shared visual artifacts, especially in early stages where formal trace matrices are not yet stabilized.

V. Discussion

A. Managerial Pragmatism and the Perceived "Burden" of RT

Our findings nuance the common claim that RT is inherently perceived as a burden in Agile and fast-paced projects [8], [9]. The interviews suggest that RT is considered overhead primar- ily when its managerial value is opaque. When practitioners can clearly see how RT supports survival, whether by passing audits, meeting KPIs, or preventing uncompensated scope increases, they are more willing to invest in it. Conversely, in contexts without such drivers, RT is easily deprioritized.

This observation complements prior work that identifies human and organizational factors as central to RT success [7], [10]. It also suggests that interventions which simply introduce new tools or prescribe more detailed trace links may fail unless accompanied by changes in incentive structures and audit practices.

B. Reframing "Neglect" as Tactical Reasoning

The tactical response model challenges interpretations of incomplete trace data as mere lack of discipline. In several cases, practitioners articulate explicit reasoning for deviat- ing from prescribed RT practices. Emergency-driven delay, administrative gaming, and triangular negotiation illustrate how practitioners weigh the costs and benefits of RT under constraints that are not captured in formal process descriptions.

These behaviors resonate with sociotechnical views of software processes, which emphasize the gap between formal methods and actual work practices. Rather than treating devi- ations as failures, organizations may benefit from examining how process definitions and tool configurations contribute to
these tactical responses.

C. Implications for RT Tool and Process Design

The The motivation–tool map highlights significant risks in adopting a "one-size-fits-all" RT solution. While issue trackers, spreadsheets, dashboards, and visual design tools all support traceability, they do so for fundamentally different managerial reasons. To be effective, RT ecosystems must be heterogeneous, allowing practitioners to combine tools that satisfy specific project profiles rather than generic best practices. For instance, in audit-heavy environments like healthcare, reliance on manual artifact linking often fails during emergencies due to the "finish first, track later" tactic. Consequently, organizations should invest in automated log analytics such as dashboards connected directly to system logs to ensure factual accountability is preserved without impeding critical service restoration.

In contrast, for client-facing or government projects driven by contract protection, the fluidity of modern issue trackers can be a liability. In these contexts, practitioners should formalize "snapshot traceability" using static spreadsheets or signed baselines at key milestones. Unlike dynamic backlogs, these static artifacts provide the indisputable evidence required for "triangular negotiation" when stakeholders introduce scope creep. Finally, for internal Agile teams, it is crucial to decouple traceability from individual KPIs. Since linking strict trace mandates to personal performance metrics frequently encourages "administrative gaming", managers should simplify trace requirements to focus on collective output rather than individual compliance, thereby reducing the incentive for developers to manipulate ticket data to meet artificial targets.

VI. Threats to Validity

As a qualitative study relying on in-depth interviews with five informants, the primary limitation of this work is the inability to statistically generalize the findings to the broader software industry. We cannot claim that a specific percentage of practitioners engage in "administrative gaming" or "triangular negotiation." However, this study aims for analytical generalization rather than statistical enumeration. By selecting a diverse set of participants spanning contrasting domains including highly regulated healthcare environments, public sector government projects, and fast-paced internal product teams we identified consistent behavioral patterns that transcend specific organizational cultures.

The recurrence of tactical responses, such as "Finish First, Track Later" and "Administrative Gaming", across these varied contexts suggests that these behaviors are not isolated incidents but structural responses to universal tensions between fluid Agile work and rigid managerial controls (KPIs, Audits). Therefore, while the sample size is small, the transferability of the proposed Tactical Response Model is high for other industrial settings characterized by similar high-pressure and compliance-driven constraints. The depth of the interview data allowed us to saturate the specific categories of the model (MOT, BAR, STRAT), providing sufficient "information power" to characterize the decision-making mechanisms, even if the prevalence of these mechanisms requires future quantitative validation.

VII. Conclusion and Future Work

This paper has examined RT as a managerial and tactical practice rather than solely as a technical method. Through interviews with five practitioners and hybrid coding supported by ATLAS.ti, we developed a tactical response model that explains how practitioners navigate non-technical barriers and why RT is selectively emphasized or deferred.

The study demonstrates that RT adoption is driven by pragmatic motivations such as external audits, KPI pressure, and contract protection; that practitioners employ specific tactics, including emergency-driven delay, administrative gaming, and visual validation, to reconcile RT with operational constraints; and that tool ecosystems reflect organizational motivations more than universal best practices.

Future research could extend this work in several direc- tions. First, large-scale surveys or mixed-method studies could test whether the identified barriers and strategies such as emergency-driven delay or administrative gaming are preva- lent across industries and cultures. Second, in-depth case stud- ies could track how tactical strategies evolve as organizations mature their SOPs and reconfigure RT tools. Finally, design research could explore interventions that realign incentives, processes, and tools to support RT practices that are both realistic and ethically robust.

References

[1] A. A. Al-Subaihi, A. Al-Badwi, and W. Al-Marri, “Machine learning approaches for automated software traceability: A systematic review,” *Journal of King Saud University–Computer and Information Sciences*, vol. 34, no. 9, pp. 6806–6825, 2022.

[2] N. Ali, L. A. Zaina, and W. Al-Marri, “The impact of traceability on software maintenance and evolution: A systematic literature review,” *Journal of Software: Evolution and Process*, vol. 35, no. 5, e2534, 2023.

[3] A. Bouaouad and M. Gzara, “Requirements traceability: A systematic review and industry case study,” *Procedia Computer Science*, vol. 184, pp. 111–118, 2021.

[4] O. Gotel, P. M”ader, and M. Zallio, “Modeling requirements traceability: A review of the past, present, and future,” in *Proc. 1st Int. Workshop on Twin Peaks of Requirements and Architecture*, 2012, pp. 43–52.

[5] M. L. Kery, M. Yin, and A. Horvath, “A mapping study on documentation in continuous software development,” in *Proc. 29th IEEE/ACM Int. Conf. on Program Comprehension (ICPC)*, 2021, pp. 172–183.

[6] Y. Li, Y. Wang, M. Li, and G. Li, “A systematic literature review of pre-requirements specification traceability,” *IEEE Access*, vol. 9, pp. 79 549–79 558, 2021.

[7] M. Panis, M. Verdonk, and S. De Haes, “Collaborative traceability management: A literature review,” in *Proc. 20th Int. Conf. on Enterprise Information Systems (ICEIS)*, 2018, pp. 237–246.

[8] H. Raharjo, A. Wibowo, and A. Nurdin, “Requirements traceability in agile software development: A systematic literature review,” *Telkomnika*, vol. 18, no. 5, pp. 2561–2568, 2020.

[9] R. P. D. Santos, N. Niu, and D. Viana, “Requirements traceability in agile software development: An empirical study,” in *Proc. 26th IEEE Int. Requirements Engineering Conf. (RE)*, 2018, pp. 360–365.

[10] R. Torkar, T. Gorschek, and R. Feldt, “Traceability maintenance factors and guidelines: A case study,” in *Proc. 3rd Int. Conf. on Software Engineering Advances (ICSEA)*, 2009, pp. 401–406.

[11] M. Tufail, A. A. Khan, and M. Riaz, “Empirical studies on software traceability: A mapping study,” *IEEE Access*, vol. 9, pp. 131 923–

131 945, 2021.
[12] S. Winkler and J. Pilgrim, “What have we learnt from the challenges of semi-automated requirements traceability?” in *Proc. 5th Int. Conf. on Global Software Engineering (ICGSE)*, 2010, pp. 209–218.
[13] T. Zhang, J. Li, H. Zhao, and Z. Liu, “A systematic literature review of issue-based requirement traceability,” in *Proc. 27th Asia-Pacific Software Engineering Conf. (APSEC)*, 2020, pp. 461–470.

Optimized Differential Ring VCO Architecture Featuring Wide Tuning Range, Low Phase Noise, and Enhanced Linearity in CMOS Technology

1st Oddepelli Rakesh
Dept. of ECE
NIT Warangal
Warangal, India
or24ecr2r03@student.nitw.ac.in

2nd Prof Patri Sreehari Rao
Dept. of ECE
NIT Warangal
Warangal, India
patri@nitw.ac.in

***Abstract*—The design and analysis presented in this research provide a three-stage differential ring-voltage controlled oscillator (DR-VCO) with a high tuning range (TR) of 90.9% and a non-linearity error of 0.4% in TSMC BCD 180 nm technology. This proposed structure exhibits efficient phase noise characteristics, achieving -100.51 dBc/Hz at an offset frequency of 1 MHz, indicating its robustness and suitability for a precision frequency synthesizer. The figure of merit (FOM) calculations provide a comprehensive evaluation of the VCO performance, yielding a result of -177.1 dBc/Hz. This study shows that the proposed VCO could improve the performance of frequency synthesis applications, such as Phase-Locked Loop (PLL) and communication systems, particularly in IoT devices, where linearity is significant. With a tuning frequency of 1.20 GHz to 3.20 GHz, it offers exceptional adaptability for a variety of applications.**

***Index Terms*—DR-VCO, Duty cycle, FoM, Non-Linearity error, Phase Locked Loop (PLL), Phase noise, Tuning Range (TR).**

I. Introduction

In the modern interconnected world, there are billions of edge devices connected via wireless communication, including IoT devices. The voltage-controlled oscillator (VCO) uses amplification, feedback, and a resonance circuit to generate a repeating voltage waveform at a specific frequency. VCO is utilized in a variety of disciplines, including communications, frequency synthesis, and clock synthesis [1][2].

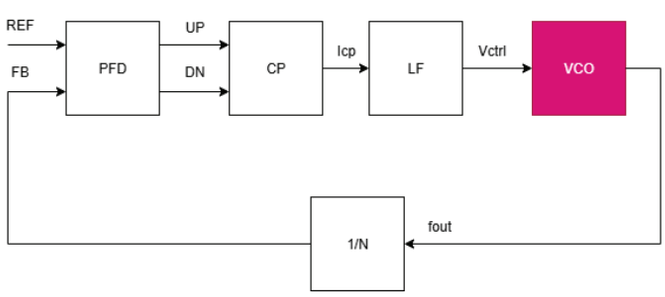


Fig. 1. Block diagram of PLL

The performance of the Voltage-Controlled Oscillator has a direct impact on the overall efficiency of the Phase-Locked Loop circuit. The two primary types of VCO are the ring VCO and the LC-VCO. LC-VCO circuits are frequently used in high-frequency applications because of their exceptional phase noise performance. LC-VCO is not ideal for systems that require a minimal area and low power, since the presence of an inductor increases the overall size and power consumption of the circuit. The ring oscillator performs better than an LC oscillator in several areas, such as tuning range, power consumption, and area measurement, although its phase noise is relatively high [3]. As a result, the type of oscillator that is selected should meet specific design requirements.

In CMOS implementation, the Differential Ring VCO [4], which uses the Dual Delay Path Technique along with the Pre-Charger and Pre-Discharger Structure. It also utilizes a composite load as an active inductor, which expands the tuning range but fails to achieve satisfactory linearity, phase noise, and FOM. The four-stage differential ring VCO [5] is based on a proportional to absolute-temperature (PTAT) reference circuit having an integrated linear regulator that dynamically compensates the VCO frequency against changes in the power source and the surrounding temperature for frequency stability for GPS applications, but it fails to achieve a wide tuning range. To reduce power consumption, the current starved topology with DR-VCO and a 0.5V power supply is used. However, its linearity and phase noise are low [8]. The fully differential four-stage ring topology with active inductors consumes more power, and the tuning range is about 20.4%, with a minimal phase noise of -86.7 dBc/Hz [9], which is not suitable for wide tuning applications. Differential cascode voltage switch logic with a resistive enhancement technique is employed to reduce delay, but it fails to achieve linearity and tuning range [10].

II. Conventional Ring Oscillator

A traditional differential ring VCO is a widely used type of voltage-controlled oscillator that consists of a ring of differential delay cells connected in a feedback loop, generating oscillations with a frequency controlled by an input voltage. Modern communication and signal synthesis systems utilize it due to its fundamental principles and distinct advantages over single-ended ring oscillators.

A common oscillator circuit used in many different applications is shown in Fig.2. The ring oscillator is mostly

used to provide clock signals and other periodic waveforms. An even or odd number of consecutive inverter stages coupled in a closed loop to form a ring oscillator. CMOS technology is commonly used to create pairs of transistors that form each inverter stage. To complete a closed-loop circuit, the signal produced by the last inverter stage is looped back and injected as input.

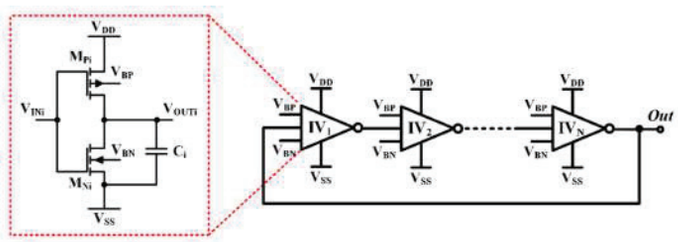


Fig. 2. Block diagram of basic ring oscillator

There are two types of delay cells used in ring VCOs: single-ended and differential. In single-ended designs, there is an odd number of delay cells, and in differential designs, oscillations are possible with an even or an odd number of delay cells. It is a popular kind of VCO that produces oscillations with control voltage. Modern communication and signal synthesis systems utilize it due to its fundamental principles and distinct advantages over single-ended ring oscillators.

The DR-VCO comprises differential delay cells that are connected in a ring. Each stage creates a delay and an inversion to maintain the oscillation. The overall phase shift around the loop must meet the Barkhausen criterian.. The number of stages (N) and the propagation delay per stage (t_d) are the main factors that affect the oscillation frequency. The frequency of oscillation is approximately given by:

$$f_{osc} = \frac{1}{2Nt_d} \tag{1}$$

III. Proposed differential ring VCO

The delay cell design is a crucial component that greatly affects the oscillator's behavior and operation of a differential ring VCO. An odd or even number of delay cell stages can be used to create the DR-VCO. Fig.3 shows a delay cell with the suggested configuration. Two, three, or four stages of delay cells are the most often used numbers. The choice of three delay cells balances speed, power efficiency, and startup conditions. Compared to the two-stage designs, it avoids latch-up issues common in even-stage rings, it consumes less power, and offers faster oscillation than four-stage designs.

A three-stage delay cell is chosen in this case because it offers benefits over two or four-stage counterparts, such as reduced propagation delay and improved power efficiency. To use the VCO block in any wireless communication system integrated into a PLL, the design goal is to produce a broad range of oscillating frequencies with linearity. In addition to reduced area, phase noise, improved linearity, and wide tuning range are taken into account. Negative feedback requires higher resistance to maintain the same speed, which is less desirable in integrated circuits. Therefore, positive feedback is utilized to increase speed.

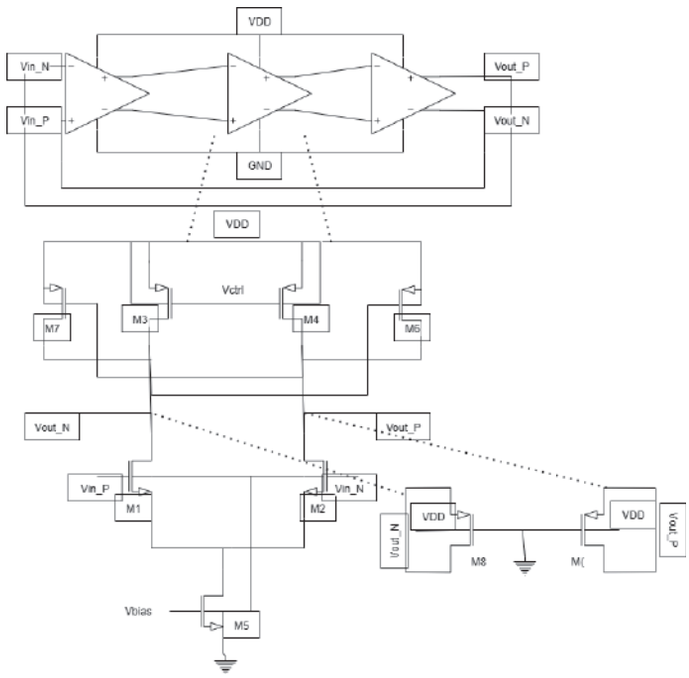


Fig. 3. Schematic diagram of proposed DR-VCO

Vin_P and Vin_N denote the differential inputs in this structure, and $Vout_P$ and $Vout_N$ denote the differential outputs. Two NMOS transistors, M1 and M2, are coupled to the input voltage. The control voltage (V_{ctrl}), connected to the gates of the PMOS transistors M3 and M4, modulates the output frequency. The oscillator receives positive feedback from a pair of cross-coupled PMOS transistors, M6 and M7. The bias voltage (Vbias) tunes the NMOS transistor M5, which serves as the source of tail current. To achieve fine-tuning, improved linearity, and reduced phase noise, all transistors should operate in the saturation region. Operating any transistor outside the saturation region can cause current mismatches, leading to stability issues and increased signal phase noise.

To maintain stability and reduce current mismatch, all transistors are designed to operate in the saturation region, resulting in an extremely low non-linearity error of 0.4%. Additionally, the differential structure naturally rejects common-mode noise, improving phase noise and signal integrity.

The differential output terminals of a delay cell are connected to the source and drain terminals of M8 and M9 PMOS Transistors. The gates of M8 and M9 are grounded to

TABLE I
PROPOSED DR-VCO PERFORMANCE COMPARISION

Reference	**[2]**	**[4]**	**[13]**	**[6]**	**[11]**	**[12]**	**This work**
Technology (nm)	130	180	180	180	180	180	**180**
Supply voltage (V)	1.5	1.8	1.8	1.8	1.8	1.8	**1.8**
No of stages	3	4	4	3	4	3	**3**
Oscillation frequency (GHz)	1.1-1.62	0.853-3.178	1.67-3.82	1.72-1.92	5.19-5.93	2.5-5.2	**1.2 - 3.2**
Tuning Range(%)	19.1	115.3	78.2	10.9	13.3	70	**90.9**
Control voltage (V)	0.4-1.3	0.7-1.8	0-1.4	0-1.6	0.3-1.8	-	**0.41-1.1**
DC Power consumption (mW)	3	1.495	3.82	13	80	17	**18.6**
Phase Noise (dBc/Hz)	-78.8	-87.02	-91.2	-102	-99.5	-90	**-100.5**
Linearity error (%)	-	-	-	-	-	-	**0.4**
Duty cycle (%)	-	-	-	-	-	-	**50**
Offset frequency (MHz)	1	1	1	1	1	1	**1**
FoM (dBc/Hz)	-143.9	-176.5	-174.9	-157.3	-158.4	-168.9	**-177.1**
Area (mm^2)	-	-	-	-	-	0.24	**0.0035**

adjust the circuit's capacitance and improve the controllability of the output frequency at lower voltages. Because of this combination, it can be easily integrated into PLL applications and wireless communication systems.

IV. RESULTS AND DISCUSSSIONS

The proposed VCO is designed using Cadence Virtuoso on TSMC BCD 180 nm technology, and the results are shown in Table I. The VCO design enables frequency variation from 1.20 GHz to 3.20 GHz, as shown in Fig.5. The frequency of oscillation, depicted with control voltage of 0.41 V to 1.1 V, achieved a tuning range of 90.9%. The phase noise of -100.5 dBc/Hz is shown in Fig.4 at the offset frequency of 1 MHz, and the duty cycle achieved is about 50%. The figure-of-merit (FOM), a statistic that measures the effectiveness of the VCO performance, is derived using the data from the phase noise, power dissipation, and tuning range measurements at the oscillation frequency. At the offset frequency of 1 MHz, the performance parameter FOM is -177.1 dBc/Hz.

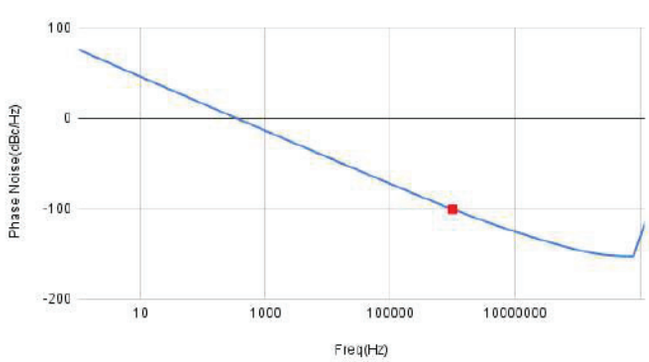


Fig. 4. VCO's phase noise at offset frequency of 1 MHz

$$FoM = L(\delta f) - 20\log\frac{f_o * TR}{\delta f * 10} + 10\log\frac{PD}{1mw} \quad (2)$$

where:

L(δf) = Phase noise

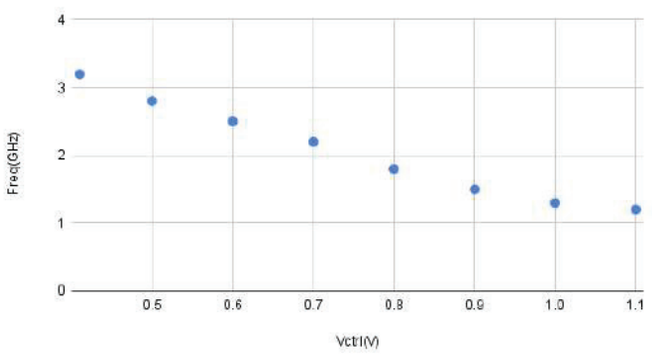


Fig. 5. Output frequency variations with Vctrl

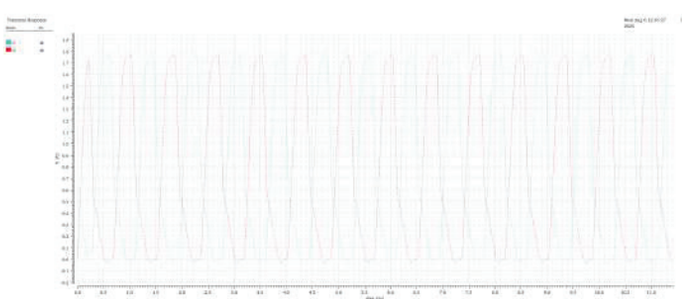


Fig. 6. Transient response of VCO

δf = Offset frequency
f_o = Output frequency
PD = Power dissipation
TR = Tuning range

VCO non-linearity error is the non-ideal, non-linear relationship between the input voltage and output frequency, which can degrade the system performance, where precise frequency control is required. The presented DR-VCO achieves a non-linearity error of 0.4%.

$$Non - linearity\ error(\%) = \frac{f_{actual} - f_{ideal}}{f_{actual}} * 100 \quad (3)$$

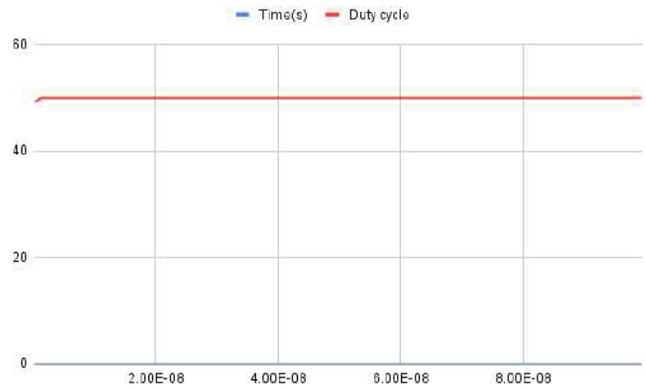


Fig. 7. Duty cycle

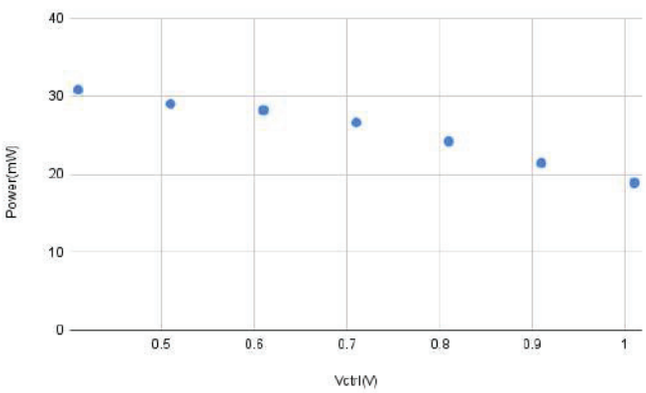


Fig. 8. Variations in power with Vctrl

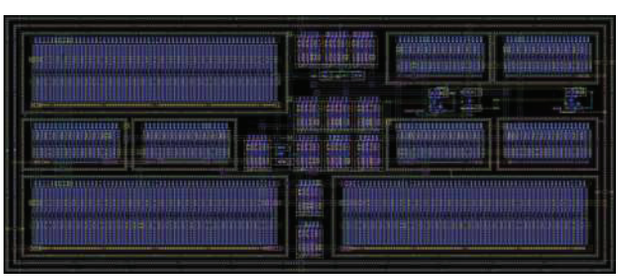


Fig. 9. Layout

$$TR\% = \frac{f_{max} - f_{min}}{f_{center}} \tag{4}$$

$$f_{center} = \frac{f_{max} + f_{min}}{2} \tag{5}$$

Phase noise, die area, linearity, power consumption, and tuning range are all trade-offs that are balanced in the delay cell design. A compact, power-efficient, and high-performance VCO that is suitable for the Internet of Things and wireless communication applications is produced by the ideal combination of the selected three-stage differential topology. This differential ring VCO is an ideal choice for integrating into a phase-locked loop because of its linear frequency control and wide tuning range. For communication systems to achieve uniform frequency synthesis, this integration is necessary.

V. CONCLUSION

The proposed DR-VCO can be tuned from 1.2 GHz to 3.2 GHz, achieving a tuning range of 90.9%, which allows it to be highly adaptable for different types of frequency bands in modern communication systems. The design achieves phase noise between -95.8 dBc/Hz and -100.51 dBc/Hz at a 1 MHz offset, which ensures better signal stability and performance in PLL-based systems. Power consumption is in the range of 18.6 mW to 30.6 mW, making it efficient enough for low-power applications. The core occupies a die area of 0.0035 mm² (70 µm × 50 µm), meaning it can be easily integrated into compact devices with space limitations. The combination of wide tuning range, low phase noise, good non-linearity error of 0.4% and compact size makes it a highly efficient, robust, and versatile VCO for modern electronic systems. Although designed for PLL and wireless communication applications, the architecture presented in this paper may also offer advantages for precision high-frequency systems due to its balanced performance.

REFERENCES

[1] A. M. A. Ali et al., “A 0.13um low phase noise and fast locking pll,” in 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), vol. 1, 2019, pp. 1468–1471

[2] X. Chen, L. Cai, Y. Xu, J. Ou, and J. Liao,”A low jitter low power and wide tuning range differential ring oscillator topology in 28nm cmos technology for clock synthesizer applications,” in 2018 29th Irish Signals and Systems Conference (ISSC), 2018, pp. 1–4.

[3] S. Min, T. Copani, S. Kiaei, and B. Bakkaloglu,”“A 90-nm cmos 5-ghz ring-oscillator pll with delay-discriminator-based active phase-noise cancellation,” IEEE Journal of Solid-State Circuits, vol. 48, no. 5, pp. 1151–1160, 2013.

[4] V. K. Viaee, M. Alijani and A. Abrishamifar, ”A Differential Ring VCO with a New Structure for the Pre-Charger and Pre-Discharger Method,” 2024 6th Iranian International Conference on Microelectronics (IICM), Tabriz, Iran, Islamic Republic of, 2024, pp. 1-5, doi: 10.1109/IICM65053.2024.10824623.

[5] I. -C. Hwang, ”A 0.236 mm 2 , 3.99 mW Fully Integrated 90 nm CMOS L1/L5 GPS Frequency Synthesizer Using a Regulated Ring VCO,” in IEEE Microwave and Wireless Components Letters, vol. 22, no. 6, pp. 324-326, June 2012, doi: 10.1109/LMWC.2012.2194996.

[6] Z. -Z. Chen and T. -C. Lee, ”The Design and Analysis of Dual-Delay-Path Ring Oscillators,” in IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 58, no. 3, pp. 470-478, March 2011, doi: 10.1109/TCSI.2010.2072390.

[7] B.Razavi, Design of Analog CMOS Integrated Circuits, Edition-2001.

[8] P. Srikram, P. Ambalathankandy, M. Motomura and M. Ikebe, ”A 0.5 V Modified Pseudo-Differential Current-Starved Ring-VCO with Linearity Improvement for IoT Devices,” 2023 International Electrical Engineering Congress (iEECON), Krabi, Thailand, 2023, pp. 219-223, doi: 10.1109/iEECON56657.2023.10126919.

[9] C. Zhang, Z. Li, J. Fang, J. Zhao, Y. Guo and J. Chen, ”A novel high-speed CMOS fully-differentical ring VCO,” 2014 12th IEEE International Conference on Solid-State and Integrated Circuit Technology (ICSICT), Guilin, China, 2014, pp. 1-3, doi: 10.1109/ICSICT.2014.7021580.

[10] D. Z. Turker, S. P. Khatri and E. Sanchez-Sinencio, ”A DCVSL Delay Cell for Fast Low Power Frequency Synthesis Applications,” in IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 58, no. 6, pp. 1225-1238, June 2011, doi: 10.1109/TCSI.2010.2103170.

[11] Y. A. Eken and J. P. Uyemura, ”A 5.9-GHz voltage-controlled ring oscillator in 0.18-/spl mu/m CMOS,” in IEEE Journal of Solid-State Circuits, vol. 39, no. 1, pp. 230-233, Jan. 2004, doi: 10.1109/JSSC.2003.820869.
[12] Wei-Hsuan Tu, Jyh-Yih Yeh, Hung-Chieh Tsai and Chorng-Kuang Wang, ”A 1.8V 2.5-5.2 GHz CMOS dual-input two-stage ring VCO,” Proceedings of 2004 IEEE Asia-Pacific Conference on Advanced System Integrated Circuits, Fukuoka, Japan, 2004, pp. 134-137, doi: 10.1109/APASIC.2004.1349429.
[13] Alijani M, Javanmardi M, Abrishamifar A. A wide tuning range CMOS differential ring VCO using an active inductor for wireless applications. Int J Circ Theor Appl. 2025; 53(3): 1171-1183. doi:10.1002/cta.4155

Designing an Integrated Digital Governance Architecture for Disaster-Resilient Villages

Astuti Purwaningsih[1], Lukito Edi Nugroho[2], Rudy Hartanto[3]
Department of Electrical and Information Engineering
Faculty of Engineering, Universitas Gadjah Mada
Yogyakarta, Indonesia
[1]astuti.purwaningsih@mail.ugm.ac.id, [2]lukito@ugm.ac.id, [3]rudy@ugm.ac.id

***Abstract*—Indonesia's Disaster-Resilient Village (Destana) program faces critical challenges in digital governance, characterized by fragmented applications, reactive budgeting, and disconnected risk data. While the national standard SNI 8357:2024 defines functional indicators for resilience, a coherent architectural blueprint to translate these into interoperable digital systems remains absent. This study designs an integrated Enterprise Architecture (EA) for village disaster governance aligned with the national SPBE framework. Using a standards-driven derivation method, 55 SNI indicators were decomposed into 10 core business processes and synthesized into a multi-layered architecture. The architecture employs a Service-Oriented Architecture (SOA) and API-first strategy to realize a collaborative ecosystem supported by four modular applications, hosted on a hybrid-cloud infrastructure to ensure data sovereignty. Key results highlight a system-enforced interlocking mechanism where village development planning is mandatorily linked to digital risk assessment payloads. This framework facilitates a fundamental shift in local governance from reactive emergency response to proactive, evidence-based mitigation, providing a strategic blueprint to eliminate sectoral silos. (*Abstract*)**

Keywords—Disaster-Resilient Village, Enterprise Architecture, Service-Oriented Architecture (SOA), System Interoperability, SPBE, SNI 8357:2024. (*key words****)***

I. INTRODUCTION

Indonesia's geographical position at the convergence of four major tectonic plates creates exceptional disaster vulnerability, a fact underscored by the World Risk Report 2024 which identifies it as one of the most disaster-prone nations globally [1][2][3]. In response, the Sendai Framework 2015-2030 has mandated a shift toward localized capacity building [4]. Aligning with this, the National Disaster Management Agency (BNPB) has prioritized community resilience as the cornerstone of the 2025-2029 National Disaster Management Plan, positioning the village as the forefront of mitigation through the Disaster-Resilient Village (Destana) program [5][6].

Despite this strategic prioritization, implementation is impeded by structural fragmentation. A critical disconnect exists between disaster management planning and village development planning. Research indicates that disaster risk documents are often treated as supplementary rather than mandatory inputs for the Village Medium-Term Development Plan (RPJMDes), leading to funding discontinuities where mitigation activities are excluded from the village budget [7][8]. Furthermore, the governance approach remains largely elite-centric, failing to institutionalize knowledge beyond the village apparatus [9]. This fragmentation is exacerbated by chaotic digital governance; village governments are currently overwhelmed by non-integrated applications from supra-village authorities [10]. This lack of integration results in acute data silos, hindering the real-time coordination and efficient reporting required for effective disaster risk reduction [11].

To address these interoperability challenges, Enterprise Architecture (EA) offers a strategic instrument for aligning organizational policies and digital systems [12][13]. While frameworks like TOGAF have been adapted for municipal levels [14][15][16], applying them to the rural context presents challenges. Empirical evidence suggests standard EA methodologies assume a level of organizational maturity and IT infrastructure rarely present in rural Indonesia, where internet penetration averages only 43% [17]. Recent studies on "Smart Village" EA [18][19][20] predominantly address administrative or healthcare domains, leaving the specific, time-critical requirements of Disaster Risk Reduction (DRR) largely untouched. This gap is significant because the national Electronic-Based Government System (SPBE) framework lacks operational guidance for integrating disaster resilience at the village level [21], and the functional indicators of SNI 8357:2024 remain high-level metrics without an architectural blueprint for digital implementation [22].

This study addresses these critical gaps by developing an integrated Enterprise Architecture for village-level disaster governance that aligns SNI 8357:2024 with the SPBE framework. A key distinction of this research is the recognition that village governments primarily function as service controllers and end-users rather than system developers. Therefore, the proposed architecture serves as a strategic blueprint for supra-village authorities to provide an integrated digital ecosystem. Specifically, the technical contributions of this study are: (i) standards-to-requirements derivation by transforming 55 SNI 8357:2024 indicators into structured business processes; (ii) a modular SOA blueprint and integration matrix to resolve application fragmentation within the SPBE structure; (iii) a hybrid-cloud and security baseline formulation for a resilient deployment model; and (iv) a risk-to-budget enforcement mechanism introducing system-based control logic to interlock assessment with village planning.

II. RELATED WORKS

A. Barriers in Community-Based DRR Implementation

Village Resilience Assessment is a vital tool for mapping baseline capacity, yet it requires rigorous data validation to be effective [11]. Practical implementation often faces structural hurdles where governance remains elite-centric, failing to institutionalize disaster knowledge among residents [9]. Furthermore, a critical disconnect exists between disaster management planning and the Village Medium-Term Development Plan (RPJMDes), leading to funding discontinuities and neglected infrastructure maintenance [7][8]. These challenges are exacerbated by sluggish digital

transformation and the burden of managing multiple non-integrated information systems [10].

B. EA and Smart Village Digital Transformation

Enterprise Architecture (EA) is a strategic instrument for aligning business processes with IT infrastructure and national regulations [12][13]. While frameworks like TOGAF and Zachman have been adapted for national and municipal levels [14][15][16], their complexity often exceeds the organizational maturity of rural units. Current "Smart Village' EA initiatives focus predominantly on general administration, population services, or healthcare [18][19][20]. However, these models leave the time-critical and multi-agency requirements of Disaster Risk Reduction (DRR) largely untouched.

C. Technical Research Gap

The absence of an architectural blueprint that integrates the functional indicators of SNI 8357:2024 within the national SPBE framework represents a significant gap, as summarized in Table I. Current studies do not address the specific coordination logic required for village-level disaster governance. This study fills this gap by translating normative standards into a coherent, engineering-based digital architecture.

TABLE I. STATE OF THE ART SUMMARY AND RESEARCH GAP

Research Focus	Method	Supporting Studies	Key Limitations/ Research Gap
Destana Implementation (Village)	Qualitative/ Descriptive	[7][8][9] [10][11]	Lacks technical architectural blueprints; focuses on policy barriers
Public Sector EA (National/ Regency)	TOGAF, Zachman, FEAF	[12][13] [14][15] [16]	Assumes high maturity; not tailored for rural resource constraints
Smart Village EA (Village)	TOGAF ADM	[18][19] [20]	Domain-specific (admin, health); does not address DRR requirements
Proposed Study: Integrated Destana EA	*BPE + Service-Oriented Architecture*	*This Study*	*Integrates SNI 8357 indicators for DRR within a system-based SPBE structure.*

III. METHODOLOGY

This study employed a Design Science Research (DSR) approach to construct an Enterprise Architecture (EA) as a digital governance artifact. The research framework was aligned with the National SPBE Architecture (Perpres No. 132/2022) [23] to ensure interoperability and compliance.

A. Data Source and Tools

Primary data were gathered through semi-structured interviews and observations in Kulon Progo Regency, involving purposively selected informants from the Regional Disaster Management Agency (BPBD) to define operational requirements, the Department of Communication and Informatics (Diskominfo) to specify technical standards, and three villages representing Pratama, Madya, and Utama resilience levels. All informant identities and institutional data were anonymized to ensure data privacy. The dataset comprised 55 functional indicators derived from SNI 8357:2024 [22]. Architectural modeling was executed using Archi 5.6.0 software with ArchiMate 3.2 notation to ensure technical rigor and standardized visualization.

B. Research Procedure

The study followed a systematic three-stage procedure to transform normative standards into a technical architecture:

1. Business Process Engineering
   The business process domain was established using a Standards-Driven Derivation method to transform qualitative SNI 8357:2024 indicators into structured digital workflows. This involved Functional Decomposition to break down indicators into discrete activity units, followed by Semantic Clustering to group activities based on shared business objects. These clusters were then synthesized into end-to-end business processes using Maturity-Based Lifecycle Modeling, ensuring the digital workflows reflect the village's resilience evolution from initiation (Pratama) to optimization (Utama).
2. Integrated SPBE Architecture Design
   The derived processes were mapped into the remaining SPBE domains using a systems engineering approach:
   - Actor & Role Mapping: A stakeholder analysis was conducted to define governance roles based on the division of concurrent government affairs. This step assigned specific responsibilities to internal agencies and external partners;
   - Service, Data, and Application Domains: These domains were realized through a Service-Oriented Architecture (SOA) and API-first strategy. This approach encapsulated internal workflows into integrated services, ensured data interoperability using "One Data" principles (e.g., integration with InaRISK and OpenSID), and resolved application fragmentation through a modular structure;
   - Infrastructure & Security: A Hybrid-Cloud Deployment model was formulated to balance data sovereignty and high availability. The security domain was integrated through defense-in-depth measures based on the CIA Triad (Confidentiality, Integrity, Availability) to ensure system resilience against both digital and physical disaster risks.
3. Evaluation and Validation
   The proposed architecture underwent a two-fold evaluation: (i) a gap analysis benchmarking the 'To-Be' design against the 'As-Is' status of the three case villages, and (ii) a structured expert review by a senior specialist in Information Systems Architecture and Smart Village/City frameworks to ensure technical feasibility.

IV. RESULTS AND DISCUSSION

A. The Proposed Enterprise Architecture Model

The primary result of this study is a comprehensive Enterprise Architecture (EA) blueprint designed to transform disaster management in Indonesian villages into a digitally governed ecosystem. The comprehensive Layered View of the architecture is presented in Fig. 1, illustrating the vertical alignment between the Business, Application, and Technology layers.

The Business Layer defines the "Collaboration Logic" between actors, which is realized by the Application Layer through decoupled modular services, and finally supported by the Technology Layer via a Hybrid-Cloud infrastructure to ensure high availability.

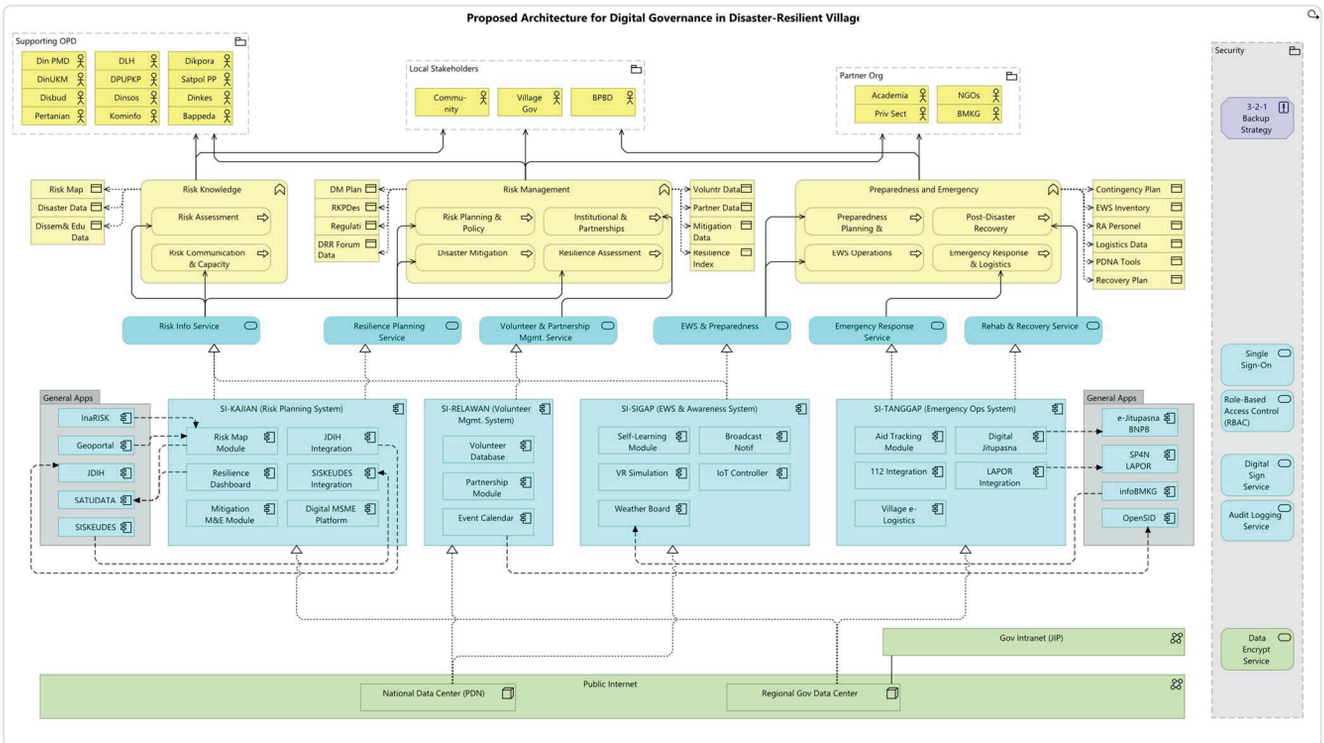


Fig. 1. Proposed Architecture for Digital Governance in Disaster-Resilient Village (Layered View)

B. Business Process Engineering

To transform the qualitative compliance indicators of SNI3 8357:2024 into a structured system, this study applied a Standards-Driven Derivation approach. The 55 standard indicators were processed through Functional Decomposition to break down indicators into activity units and Semantic Clustering to group activities sharing similar business objects. The derivation results are summarized in Table II.

TABLE II. RECAPITULATION OF SNI INDICATOR DERIVATION INTO BUSINESS PROCESS

ID	Business Process	Related Indicator Code (SNI3 8357:2024)
BP-01	Risk Assessment	5.2.1.1.1, 5.2.2.1.2, 5.2.3.1.2
BP-02	Risk Communication & Capacity	5.2.1.1.2, 5.2.2.1.1, 5.2.3.1.1, 5.2.3.1.3
BP-03	Risk Planning & Policy	5.2.1.2.1, 5.2.2.2.1, 5.2.2.3.2, 5.2.3.2.1, 5.2.3.2.7
BP-04	Institutional & Partnerships	5.2.1.2.2, 5.2.1.2.3, 5.2.2.2.2, 5.2.2.2.6, 5.2.3.2.2, 5.2.3.2.3, 5.2.3.2.6, 5.2.3.2.8
BP-05	Disaster Mitigation	5.2.1.2.4, 5.2.1.3.1, 5.2.2.2.4, 5.2.2.3.1, 5.2.3.2.4, 5.2.3.3.3
BP-06	Resilience Assessment	5.2.1.2.5, 5.2.2.2.5, 5.2.3.2.5
BP-07	Preparedness Planning & Simulation	5.2.1.4.1, 5.2.1.4.3, 5.2.2.4.1, 5.2.2.4.3, 5.2.2.4.6, 5.2.3.4.1, 5.2.3.4.3, 5.2.3.4.5
BP-08	EWS Operations	5.2.1.4.2, 5.2.2.4.2, 5.2.3.4.2
BP-09	Emergency Response & Logistics	5.2.2.4.4, 5.2.2.4.5, 5.2.2.4.7, 5.2.3.4.4, 5.2.3.4.7
BP-10	Post-Disaster Recovery	5.2.2.4.8, 5.2.2.4.9, 5.2.3.4.6, 5.2.3.4.8, 5.2.3.4.9, 5.2.3.4.10

This engineering process converts normative regulations into specific system requirements, ensuring that the resulting digital workflows are inherently compliant with national standards. By formalizing "intangible" processes into distinct system workflows, the architecture mandates activities such as Risk Assessment (BP-01) to be executed as functional prerequisites. Using Maturity-Based Lifecycle Modeling, these processes reflect the evolutionary path of village capacity, preventing them from being bypassed in the village planning cycle and ensuring that downstream physical mitigation is strictly evidence-based.

C. Collaborative Governance Model

From a systems engineering perspective, the collaborative governance model is translated into a Role-Based Access Control (RBAC) mechanism within the architecture. The specific roles and responsibilities are mapped in Table III.

TABLE III. MATRIX OF MULTI-STAKEHOLDER COLLABORATION

Actor Group / Specific Actors	Primary Role in Architecture	Main Business Function
Local Stakeholders • Village Gov (Pemdes) • Community • BPBD	Executor & Command: Primary implementers at the local level and holders of command authority during emergencies.	All Business Processes (BP-01 to BP-10)
Supporting Gov Dept. (OPD) • PMD, Bappeda • Social Affairs Office, Health Office, Civil Service Police Unit • Environment Office, Public Works & Housing Office, Agriculture Office • Cooperatives & MSME Office, Cultural Office	Enabler & Regulator: Providers of regulations, technical standards, secondary data, and supporting infrastructure.	• Risk Knowledge • Risk Management

Actor Group Specific Actors	Primary Role in Architecture	Main Business Function
• Education Office, Communications & Informatics Office		
Partner Organization • Academia • NGOs • Private Sector • BMKG	Collaborator: Capacity-building partners, weather and hazard data providers (BMKG), and contributors of additional resources.	• Risk Management • Preparedness &Emergency

The mapping defines distinct user roles and access privileges for Local Stakeholders (Executors), Supporting OPDs (Regulators), and Partner Organizations (Collaborators). This technical segregation ensures data integrity and clear accountability across the ecosystem. By technically integrating roles for the Community and NGOs within the SI-RELAWAN and SI-SIGAP modules, the system institutionalizes the Pentahelix model, ensuring that community members are active users with defined rights to input data and receive alerts.

D. Application and Data Integration Strategy

The core of the IT solution lies in the Service-Oriented Architecture and API-First strategy detailed in Table IV. This approach shifts the paradigm from building isolated tools to creating a decoupled ecosystem of shared services.

TABLE IV. INTEGRATION MATRIX OF SERVICES, DATA, AND APPLICATION

Service	Business Process	Data	Application
Risk Info Service	• Risk Assessment • Risk Comm. & Capacity • Resilience Assessment	• Risk Map • Disaster Data • Dissemination & Edu Data • Resilience Index	SI-KAJIAN
Resilience Planning Service	• Risk Planning & Policy • Disaster Mitigation	• Disaster Management Plan & RKPDes • Regulation (Perdes) • Mitigation Data	SI-KAJIAN
Volunteer & Partnership Management Service	Institutional & Partnership	• DRR Forum Data • Volunteer Data • Partner Data	SI-RELAWAN
EWS & Preparedness	• EWS Operations • Preparedness & Planning	• EWS Operations Preparedness Planning	SI-SIGAP
Emergency Response Service	Emergency Response & Logistics	• Logistics Data Contingency plan	SI-TANGGAP
Rehab & Recovery Service	Post-Disaster Recovery	• PDNA Tools Recovery Plan	SI-TANGGAP

Instead of creating a monolithic application, the architecture decomposes functionalities into four modular applications (SI-KAJIAN, SI-SIGAP, SI-TANGGAP, SI-RELAWAN) that communicate via standardized RESTful APIs. This strategy effectively resolves the fragmentation issues by enforcing the "One Data" principle through integration with national systems—pulling risk maps from InaRISK and population data from OpenSID.

By utilizing Shared Services and a centralized API Gateway, the architecture eliminates point-to-point (P2P) integration complexity and ensures data produced at the village level is immediately valid at higher governance levels. The integration logic is illustrated in Fig. 2, which demonstrates the Service Orchestration and data payloads between national platforms (InaRISK), regional servers, and village-level applications. This mechanism ensures that risk intelligence is synchronized in real-time while maintaining strict data validation constraints before any administrative record is persisted in the database.

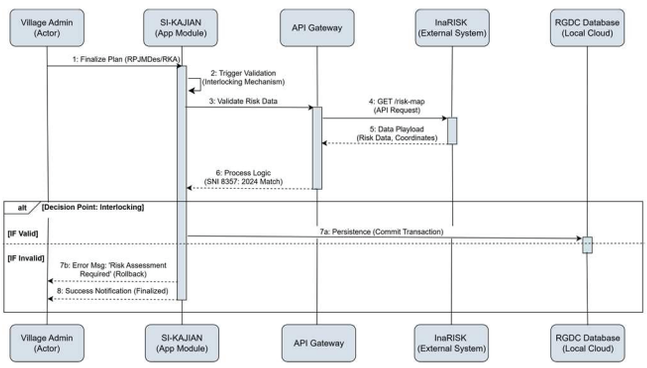


Fig. 2. Sequence Diagram of the Service Orchestration and System Interlocking Mechanism for Evidence-Based Planning.

E. Technical Infrastructure and Security Design

The infrastructure design is engineered for Resilience and Availability, as outlined in Table V.

TABLE V. INFRASTRUCTURE AND SECURITY SPECIFICATIONS

Compo-nent	Diagram Specification	Justification
Hosting Strategy	1. National Data Center : Hosting for SI-RELAWAN & SI-SIGAP. 2. Regional Gov Data Center: Hosting for SI-KAJIAN & SI-TANGGAP.	Implements a Hybrid Cloud strategy. High-exposure public services are placed in the central cloud for scalability, while sensitive/administrative applications are hosted in district government servers to ensure data sovereignty and compliance with government data regulations.
Network Strategy	1. Public Internet 2. Government Intranet (JIP): Secure access for village gov to regional servers.	Using the Government Intranet (JIP) ensures secure G2G data exchange, while the Public Internet guarantees broad G2C accessibility for citizen-facing services.
Security Principles	3-2-1 Backup Strategy	Ensures data availability and protection against physical disaster risks by maintaining off-site backup copies.
Security Services	1. Single Sign-On (SSO) 2. Digital Signature Service 3. Role-Based Access Control (RBAC) 4. Audit Logging Service 5. Data Encryption Service	• SSO: Centralized, seamless authentication. • Digital Signature: Ensures legal validity of digital documents. • RBAC / Audit Logging / Encryption: Protects confidentiality and integrity of sensitive data (victim data, logistics records).

The adoption of a Hybrid-Cloud Strategy—hosting public-facing services (SI-SIGAP) on the National Data Center (PDN) and administrative services (SI-KAJIAN) on the Regional Data Center—is justified by an analysis of

Recovery Time Objective (RTO) and Data Sovereignty. This ensures that critical warning systems remain operational on the central cloud even if local village infrastructure is compromised.

Security is implemented as a Baseline Architectural Requirement using a Defense-in-Depth strategy. Following the CIA Triad (Confidentiality, Integrity, Availability), the architecture incorporates AES-256 encryption, digital signatures (TTE) for legal non-repudiation, and a 3-2-1 Backup Strategy to protect against physical data loss in disaster-prone areas.

F. Gap Analysis Results

The gap analysis provides a critical evaluation of the divergence between current practices and the proposed architectural standards. This comparison is summarized in Table VI.

TABLE VI. GAP ANALYSIS MATRIX

Aspect	Current Implementation Status (As-Is)	Proposed Architectural State (To-Be)	Key Gap & Pain Point
Governance	Single-Agency Driven: Disaster management is perceived solely as BPBD's responsibility; limited involvement from other agencies (OPDs)	Orchestrated Collaboration: Institutionalized Pentahelix model where the Regional Secretary (Sekda) orchestrates cross-agency support	Holistic Resilience: Eliminates sectoral egos and ensures multi-dimensional support for villages
Planning Strategy	Reactive & Project-Based: Budgeting focuses on emergency response. Plans (RKPDes) are disconnected from risk maps	Evidence-Based & Systemic: Planning is mandatorily linked to risk data. The system enforces risk assessment before budget approval	Cost Effectiveness : Shift from "spending on response" to "investing in prevention."
Data Management	Passive Consumer: Villages rely on central data. High data redundancy (re-inputting resident data multiple times)	Active Producer (Data Sovereignty): Villages update their own data. "One Data" integration with OpenSID prevents redundancy	Accuracy & Ownership: Local data is more accurate and up-to-date than top-down national data
Digital Ecosystem	Fragmented Tools: Coordination via chat apps (WhatsApp) or standalone project apps that are often abandoned	Integrated Ecosystem: A modular platform with SSO and shared services hosted on Hybrid Cloud	Sustainability : Reduces IT maintenance burden and ensures system longevity

A detailed analysis of the Village Government Work Plan and Budget (RKA Desa) in Kulon Progo Regency reveals a significant architectural gap: current disaster programs are limited to reactive funding and ad-hoc responses, mapping to only a fraction of the proposed business processes. This fragmentation is exacerbated by governance silos where disaster management is isolated within BPBD, lacking the cross-sectoral orchestration from the Regional Secretary (Sekda) defined in the architecture's Business Layer.

The proposed architecture addresses this structural deficiency by functioning as a System Control Mechanism. In the As-Is state, village budgeting remains reactive because there is no technical constraint preventing the approval of budgets without comprehensive risk analysis. The To-Be architecture introduces an automated interlocking mechanism within the SI-KAJIAN module. Technically, this is implemented as a hard constraint in the application logic, where the Resilience Planning Service requires a validated risk-data payload from the Risk Map Module before a planning document can be persisted in the database. This represents a shift from a "Moral Appeal" to "System-Enforced Validation", effectively ensuring data integrity and procedural compliance through the technical architecture. Furthermore, the transformation from manual administration to a Knowledge Management System ensures that risk intelligence is persistent, directly tackling sustainability issues through data persistence.

To ensure technical viability, the proposed architecture was further validated through a structured review by a senior specialist in Information Systems Architecture. The validation confirmed the model's logical consistency and its alignment with national SPBE standards, specifically praising the Hybrid-Cloud strategy as a pragmatically sound solution for maintaining high availability in disaster-prone regions without compromising local data jurisdiction. This expert endorsement, combined with the gap analysis results, reinforces the architecture's readiness for pilot-scale deployment.

G. Implications and Limitations

The findings of this study offer significant contributions to the domain of digital public administration. Theoretically, this research contributes a Technical Interoperability Pattern, demonstrating that the Enterprise Architecture framework can be effectively scaled down to the micro-governance level by using a standards-driven derivation method. Practically, the proposed architecture serves as a strategic instrument for regional leadership to dismantle the sectoral silos identified in the gap analysis. By explicitly mapping the roles of supporting agencies within the Business Layer, the architecture provides a technical mandate for the Regional Secretary to orchestrate cross-agency collaboration via decoupled shared services.

Furthermore, the architecture addresses the critical issue of reactive budgeting by introducing a system-based control mechanism. By technically integrating the risk assessment module with the planning module through API-first principles, the system enforces a logic where village development plans must be justified by valid risk data. This implies a necessary transformation in village fiscal policy, moving toward evidence-based mitigation investment. Additionally, the proposed hybrid-cloud architecture empowers villages as data producers, ensuring they maintain their own resilience data through a distributed data ownership model independent of external project cycles.

Despite its comprehensive design, this study has limitations. First, the architecture acts as a logical enabler; its efficacy depends on the political will to enforce cross-agency collaboration. Second, the current model requires further Front-End Engineering and usability testing to address

varying digital literacy levels in rural settings. Lastly, while the DSR-based methodology is generalizable, the architectural artifacts are contextual to Indonesia's decentralization model. Future research should focus on longitudinal field testing and the prototyping of these modules to validate user engagement in diverse rural infrastructures.

V. CONCLUSION

This study presents a comprehensive Enterprise Architecture for the Destana program, providing a structured framework to bridge the gap between normative national standards and grassroots digital implementation. By transforming 55 qualitative indicators from SNI 8357:2024 into 10 structured business processes, this research provides a unified ecosystem that aligns local disaster management with Indonesia's SPBE framework.

The primary contribution of this work lies in the introduction of a system-enforced interlocking mechanism within the SI-KAJIAN module. This technical constraint compels a fundamental shift in village governance: from "a moral appeal" for risk-aware planning to a "systemic requirement", where valid risk data becomes a functional prerequisite for budget approval. Furthermore, the adoption of an API-first strategy and a hybrid-cloud infrastructure successfully addresses long-standing issues of data fragmentation and sectoral silos. This model empowers villages with data sovereignty while ensuring real-time synchronization with national systems like InaRISK and OpenSID, thereby institutionalizing the Pentahelix Collaboration model into a digital reality.

While this research provides a comprehensive logical blueprint validated by architectural experts, it is currently limited to the design phase. Future research should focus on pilot-scale prototyping and usability testing to evaluate the architecture's performance in diverse rural infrastructures. Longitudinal studies are also recommended to measure the impact of this "architecture-as-control" approach on the actual resilience levels and fiscal efficiency of Indonesian villages.

ACKNOWLEDGMENT

The authors receive financial support from Ministry of Communication and Digital Affairs of Indonesia for the submitted work.

REFERENCES

[1] Badan Nasional Penanggulangan Bencana, "IRBI Indeks Risiko Bencana Indonesia Tahun 2024," Badan Nasional Penanggulangan Bencana, Jakarta, Jan. 2025.

[2] Bündnis Entwicklung Hilft and IFHV, "WorldRiskReport 2024 Focus: Multiple Crises," Bündnis Entwicklung Hilft, Berlin, 2024.

[3] Badan Nasional Penanggulangan Bencana, *Data Bencana Indonesia 2024*. BNPB, 2024.

[4] U. Nations Office for Disaster Risk Reduction, "Sendai Framework for Disaster Risk Reduction 2015 - 2030."

[5] Badan Nasional Penanggulangan Bencana, "Peraturan BNPB Nomor 1 Tahun 2025 tentang Rencana Nasional Penanggulangan Bencana Tahun 2025-2029," Jakarta, 2025.

[6] BNPB, "Peraturan Kepala Badan Nasional Penanggulangan Bencana Nomor 1 Tahun 2012 tentang Pedoman Umum Desa/Kelurahan Tangguh Bencana," 2012.

[7] Supratiwi, Yuwanto, and Kushandajani, "Community participation in disaster management in Gondoriyo village administration, Ngaliyan sub-district, Semarang city," in *IOP Conference Series: Earth and Environmental Science*, Institute of Physics, 2022. doi: 10.1088/1755-1315/1041/1/012032.

[8] M. Fuady, M. A. Kevin, M. R. Farrel, Buraida, and A. Triaputri, "Effectiveness and Challenges of the Resilient Village Program in Indonesia," *IOP Conf Ser Earth Environ Sci*, vol. 1479, no. 1, p. 012025, Apr. 2025, doi: 10.1088/1755-1315/1479/1/012025.

[9] S. Arifin, S. S. Wicaksono, S. Sumarto, M. Martitah, and D. Sulistianingsih, "Disaster resilient village-based approach to disaster risk reduction policy in Indonesia: A regulatory analysis," *Jàmbá Journal of Disaster Risk Studies*, vol. 13, no. 1, Jun. 2021, doi: 10.4102/jamba.v13i1.1021.

[10] M. Faozanudin, A. Rokhman, L. Sri Sulistiani, and M. Yamin, "Investigating Key Factors Affecting the Slow Digital Transformation in Village Government Public Service Improvement in Indonesia," *International Research Journal of Multidisciplinary Scope*, vol. 06, no. 04, pp. 654–667, 2025, doi: 10.47857/irjms.2025.v06i04.07116.

[11] V. F. Nurmala, B. S. Pelita Panjaitan, and A. Subiyanto, "The Importance of Village Resilience Assessment in Village Based Programs to Support Regional Resilience," *IOP Conf Ser Earth Environ Sci*, vol. 1173, no. 1, p. 012059, May 2023, doi: 10.1088/1755-1315/1173/1/012059.

[12] H. Guo and S. Gao, "Enterprise Architectures in E-Governments Studies: Why, What and How?," 2020, pp. 3–14. doi: 10.1007/978-3-030-45002-1_1.

[13] R. Trimanadi and D. Indra Sensuse, "Understanding Government Enterprise Architecture: A Review and Case Study," *IEEE Access*, vol. 13, pp. 90088–90108, 2025, doi: 10.1109/ACCESS.2025.3571742.

[14] A. H. Fahlevi, R. A. Nugraha, Falahah, and A. M. S, "Enterprise Architecture Design on Data and Application Domain Using TOGAF Framework (Case Study: Communication and Informatics Office of Bandung City Government)," in *2023 International Conference on Digital Business and Technology Management (ICONDBTM)*, IEEE, Aug. 2023, pp. 1–7. doi: 10.1109/ICONDBTM59210.2023.10326923.

[15] A. A. Saiya and A. A. Arman, "Indonesian Enterprise Architecture Framework: A Platform for Integrated and Connected Government," in *2018 International Conference on ICT for Smart Society (ICISS)*, IEEE, Oct. 2018, pp. 1–6. doi: 10.1109/ICTSS.2018.8549990.

[16] A. K. Darmawan, F. Masykur, M. Muhsi, B. A. Umam, and R. Rofiuddin, "Proposing Enterprise Architecture for Smart Regencies in Indonesia: A Perspective of Zachman Framework (ZF)," in *9th International Conference on ICT for Smart Society: Recover Together, Recover Stronger and Smarter Smartization, Governance and Collaboration, ICISS 2022 - Proceeding*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/ICISS55894.2022.9915118.

[17] "Survei Penetrasi Internet dan Perilaku Penggunaan Internet 2025," 2025.

[18] M. Ilham Alhari and A. A. Nur Fajrillah, "Enterprise Architecture: A Strategy to Achieve e-Government Dimension of Smart Village Using TOGAF ADM 9.2," *JOIV : International Journal on Informatics Visualization*, vol. 6, no. 2–2, p. 540, Aug. 2022, doi: 10.30630/joiv.6.2-2.1147.

[19] D. R. M. Imanudin, B. M. Izzati, and F. Dewi, "Design Of Enterprise Architecture for Smart Village in Sumur Bandung Sub-District (Case Study: Governance Dimension)," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 7, no. 3, pp. 869–879, Aug. 2022, doi: 10.29100/jipi.v7i3.3171.

[20] K. N. Ikrima, A. A. Nur Fajrillah, and W. A. Nurtrisha, "Enterprise Architecture as Smart Village Development Strategy in the Health Services Dimension," *JITSI : Jurnal Ilmiah Teknologi Sistem Informasi*, vol. 4, no. 3, pp. 101–109, Sep. 2023, doi: 10.62527/jitsi.4.3.166.

[21] "Peraturan Presiden Nomor 95 Tahun 2018 tentang Sistem Pemerintahan Berbasis Elektronik," Jakarta, 2018.

[22] Badan Standarisasi Nasional, "SNI3 8357:2024 Desa tangguh bencana," 2024.

[23] "Peraturan Presiden Nomor 132 Tahun 2022 tentang Arsitektur Sistem Pemerintahan Berbasis Elektronik Nasional," Jakarta, 2022.

Design of a High-Rate Modulation Code for 4-ary Holographic Data Storage Systems

Khanh Long Tran
Department of Software Engineering
FPT University
Ha Noi, Viet Nam
Khanhtlhe176617@fpt.edu.vn

Chi Dinh Nguyen*
Faculty of Computing Fundamentals
FPT University
Ha Noi, Viet Nam
Chind3@fe.edu.vn

***Abstract*—Holographic data storage (HDS) is regarded as a promising solution for future high-capacity optical storage, owing to its ability to record and reconstruct entire pages of data in parallel. However, when using multilevel symbols to increase density, the system becomes very sensitive to two-dimensional (2D) intersymbol interference (ISI), especially when the lowest-level and highest-level symbols are adjacent. In this study, we propose a high-rate 4-ary modulation code designed to suppress harmful forbidden patterns. On a realistic HDS channel with misalignment, optical blur, and noise, the coding yields better page balance and consistently outperforms a prior-art 4/6-rate code at the same page density. The simulation results show that the proposal achieves about $1\,\mathrm{dB}$ gain improvement over the 4/6-rate prior code at bit error rate 10^{-9}.**

***Index Terms*—Holographic data storage, multilevel recording, four-level modulation code, two-dimensional intersymbol interference, high-rate constrained coding**

I. Introduction

The global datasphere is rapidly expanding, driving demand for denser, faster, and more energy-efficient storage systems [4]. Holographic data storage (HDS) addresses the challenges by recording full two-dimensional (2D) pages in a volume medium for high volumetric density and page-parallel I/O, using angular/shift/wavelength multiplexing techniques to superimpose many holograms in nearly the same region [1], [3], [5], [7]. For 4-ary page recording in HDS, where each pixel represents two bits mapped as symbols 0, 1, 2, 3, the blur from the point-spread function produces 2D intersymbol interference (ISI) that becomes more severe under misalignment and interpage crosstalk [5], [6], [8]. The most harmful cases are adjacent 0–3 (or 3–0), which persist even with well-tuned equalization/detection, motivating constrained modulation codes to avoid such patterns [2], [9], [13].

Prior HDS modulation codes demonstrate the central trade-off between robustness and rate [11]–[15], [18], [19]. Low-rate schemes (e.g., 2/3 or 4/6) strongly suppress 2D ISI and fatal adjacencies but sacrifice space efficiency [11]. Higher-rate options (e.g., 8/10 or 14/16) recover density but relax constraints and thus offer weaker protection against worst-case patterns [12], [18]. This gap motivates the codes that maintain a high rate while shaping page characteristics to stabilize symbol histograms and mitigate 2D ISI.

We present a 10/12-rate 4-ary modulation code for the multilevel HDS. A 4×3 block layout suppresses harmful 0–3/3–0 adjacencies. We evaluate on a realistic HDS channel with a 2D partial-response equalizer and a soft-output detector [5], [10], under misalignment, optical blur, and additive noise. At equal page density, the code yields better page balance and tighter histograms, and achieves about $1\,\mathrm{dB}$ signal-to-noise ratio (SNR) improvement over the previous proposed 4/6-rate code at bit error rate (BER) 10^{-9}. The advantage persists under misalignment and increases with blur.

The remainder of the paper is organized as follows. Section II reviews the related works and provides essential background. Section III formulates the problem and introduces the proposed coding design. Section IV describes the channel model and the detector employed for performance evaluation. Section V presents the experimental results and discussions, and Section VI concludes the paper.

II. Related Work

In multilevel HDS systems, the short-range constraints are used to curb 2D ISI and reduce interpage interference (IPI). Blur and misalignment reduce the effective separation between symbol levels, with interactions between levels 0 and 3 being particularly detrimental. In practice, symbol histograms often become unbalanced and level usage may be skewed, leading to increased cluster overlap and DC bias, which in turn elevate error rates and promote error propagation. Modulation codes are therefore designed to suppress such extreme interactions and prevent page layouts that exacerbate overlap or imbalance.

Early studies constrained small pixel neighborhoods using tiled templates to suppress 0-3 adjacencies and mitigate ISI. A 3×1 tile (A,B,C) with level sets $A\in\{0,1,2\}$, $B\in\{1,2\}$, $C\in\{1,2,3\}$ was mirrored on alternate raster lines to block horizontal or vertical 0-3 pairs, biasing usage toward mid levels and tightening received histograms [11]. A later $6/9$ variant adopted a 3×3 tile with three free sites and six limited to $\{1,2\}$, achieving similar adjacency control and improved histogram separation [12]. The approach evolved into a $4/6$ mapping using a 3×2 tile with two free sites ($\{0,1,2,3\}$) and four restricted to $\{1,2\}$, completely eliminating $(0,3)$ interactions [14]. The $4/6$ design maintained a lower BER and remained robust under lateral misalignment.

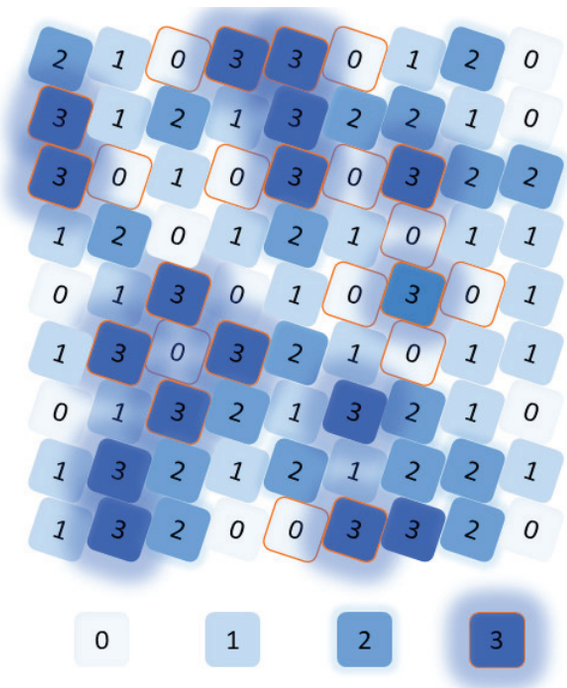


Fig. 1: Fatal interference (0-3 interactions) and 2D ISI.

To further enhance density, high-rate designs have been proposed that either slightly relax constraints or extend the codeword length. For instance, a 14/16 scheme based on a 4×4 tile allows a few 0-3 cases but resolves them by a four-neighbor majority rule (ties to 0) to limit IPI and stabilize page intensity [13]. Similarly, a 4/5-rate code preserves adjacency control while increasing the minimum inter-codeword distance to approximately ≈ 0.47 in the chosen intensity metric, at equal page density, achieving about $1\,\mathrm{dB}$ better than a 6/9 baseline at BER 10^{-9} [15].

In summary, low-rate codes (e.g., $2/3$-rate, $4/6$-rate) effectively suppress harmful 0–3 adjacencies but suffer from reduced coding efficiency, whereas higher-rate schemes (e.g., $4/5$-rate, $14/16$-rate) improve density at the expense of reintroducing forbidden patterns and greater sensitivity to blur and misalignment. Motivated by these trade-offs, this study introduces a high-rate coding scheme that maintains adjacency suppression while achieving reliable BER performance.

III. Proposed 10/12 4-ary Modulation Code

A. Problem Formulation

In general, an m-ary page stores $\log_2 m$ bits per pixel. We focus on a 4-ary HDS system, where each pixel carries two bits and the bit pairs {00, 01, 10, 11} are represented by the symbols {0, 1, 2, 3}. In HDS systems, 2D ISI is considered as the dominant impairment that degrades overall system performance. This challenge becomes even more critical when multilevel signaling is employed in HDS systems. The most critical interaction occurs when the smallest and largest symbols are adjacent (i.e., 0 next to 3 or 3 next to 0), forming what are known as fatal-interference (FI) patterns, as illustrated in Fig. 1. The objective of this paper is to design a high-rate modulation code that simultaneously limits FI interactions within blocks and across block borders, while maintaining a balanced proportion of mid-level symbols 1, 2 to maintain overall page intensity. These two objectives, when combined, tighten the received histogram, stabilise detection thresholds under blur and misalignment, and reduce the raw BER. The following sections present a 10/12-rate 4-ary modulation code based on these design principles.

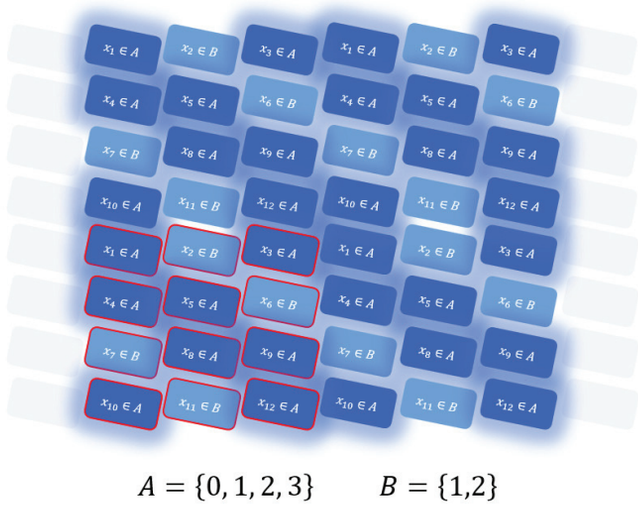


Fig. 2: 4×3 codeword layout shown in part of the page.

B. Proposed Modulation Code

The encoder maps a length–10 input $(a_1, \ldots, a_{10}) \in \{0,1,2,3\}^{10}$ to a 4×3 codeword and tiles such blocks across a page. Let

$$C_{12} \equiv \begin{bmatrix} x_1 & x_2 & x_3 \\ x_4 & x_5 & x_6 \\ x_7 & x_8 & x_9 \\ x_{10} & x_{11} & x_{12} \end{bmatrix}$$

Four cells are constrained to the mid–level set $B = \{1,2\}$, namely x_2, x_6, x_7, and $x_{11} \in B$, while the remaining eight cells take values from $A = \{0,1,2,3\}$. By constraining these positions (anchor cells) within each 4×3 block, page tiling significantly reduces the occurrence of $(0,3)$ or $(3,0)$ interactions, although a few residual cases may still appear. This design also ensures a one-to-one mapping, since the total number of possible input combinations equals the total number of codewords that can be generated under this construction, i.e., $2^4 \cdot 4^8 = 4^{10}$. This tiling structure is illustrated in Fig. 2.

Encoding. The encoder maps ten 4-ary input symbols $(a_1, a_2, \ldots, a_{10})$ into a 2D 4×3 output block $(x_1, x_2, ..., x_{12})$. Eight symbols are placed directly into unconstrained cells. Whereas, the remaining two input symbols, a_5 and a_{10}, are decomposed into four mid-level anchor cells x_2, x_6, x_7, x_{11}, whose values are restricted to the mid-level set $B = \{1,2\}$. This mapping ensures that the four constrained positions

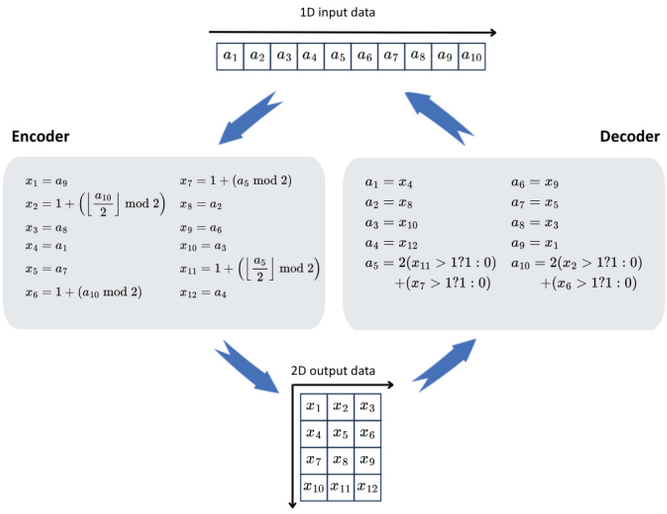


Fig. 3: Encoding and decoding schemes.

always belong to $\{1, 2\}$, effectively reducing the probability of 0–3 adjacencies across block borders.

Decoding. During decoding, the 12 received symbols $(x_1, \ldots, x_{12})$ are used to reconstruct the ten input symbols. Eight symbols $(a_1, a_2, a_3, a_4, a_6, a_7, a_8, a_9)$ are recovered directly from their corresponding positions. The remaining two a_5, a_{10} are reconstructed from the binary indicators of the constrained cells. The values are determined by combining the binary states $(x_i > 1, i = \{2, 6, 7, 11\})$ of its corresponding anchor pair, which reverses the modulo mapping applied in the encoder. The overall process preserves a one-to-one mapping, ensuring lossless reconstruction. Fig. 3 illustrates a full design of the encoding and decoding process.

IV. Holographic Channel Model and Detection Framework

Fig. 4 illustrates the end-to-end detection pipeline for a multilevel HDS system. The one-dimensional (1D) input sequence $a[k]$ is first processed by a 2D modulation encoder, which maps it into page patterns $d[p, q] \in \{0, 1, 2, 3\}$ that are more robust to channel distortion. The coded page $d[p, q]$ is then passed through the HDS channel. From a physical perspective, this channel is often modeled as a 4-f optical system, where the finite aperture at the Fourier plane acts as a low-pass filter. This aperture constraint represents the diffraction limit of the system, the root cause of 2D ISI, and defines the channel's characteristic point spread function (PSF). The continuous-domain PSF [16], $h(x, y)$, is commonly modeled using a sinc function incorporating blur, via σ_b, and potential misalignments (m_x, m_y),

$$h(x, y) = \frac{1}{\sigma_b^2} \mathrm{sinc}^2 \left(\frac{x - m_x}{\sigma_b}, \frac{y - m_y}{\sigma_b} \right), \tag{1}$$

where $\mathrm{sinc}(x, y) = (\sin(\pi x)/\pi x) \cdot (\sin(\pi y)/\pi y)$.

For digital processing, this continuous PSF must be sampled by the pixels of the detector. The discrete PSF, $h[p, q]$, is thereby computed by integrating the continuous PSF over the area of a detector pixel located at (p, q),

$$h[p, q] = \int_{q-1/2}^{q+1/2} \int_{p-1/2}^{p+1/2} h(x, y)\, dx\, dy. \tag{2}$$

This discretization is essential because it converts the physical optics of the recording medium into a form usable by digital estimation and equalization. The discrete PSF is typically truncated to a 5×5 window to model 2D ISI with tractable complexity.

With the discrete PSF, the received signal at pixel (p, q) is the 2D convolution of the original data page with the channel impulse response,

$$\begin{aligned} r[p, q] &= (d \otimes h)[p, q] + w[p, q] \\ &= \sum_{k,l} d[k, l]\, h[p - k, q - l] + w[p, q]. \end{aligned} \tag{3}$$

where, $r[p, q]$ is the measured charge-coupled device (CCD) intensity, and $w[p, q]$ is modeled as AWGN, capturing stochastic sources such as shot noise (from the quantum nature of photons) and thermal noise (from the detector electronics). The channel quality can be quantified by the SNR,

$$SNR = 10 \log_{10} \left(\frac{1}{\sigma_w^2} \right), \tag{4}$$

where σ_w^2 is the Gaussian noise variance (the signal power normalized to 1).

To manage the resulting 2D ISI, the received signal $r[p, q]$ is passed through a 2D partial response (PR) equalizer. This equalizer does not attempt to eliminate ISI, instead, it reshapes the received signal into a predefined, simpler PR target t, which is more easily handled by the detector. The resulting equalized signal is,

$$e[p, q] \;=\; r[p, q] \otimes g[p, q] \tag{5}$$

where $g[p, q]$ represent for the equalizer coefficients. A commonly used PR target is a 3×3 kernel that keeps only the center pixel and its four orthogonal neighbors. We assume that the center has weight M, each immediate neighbor (up, down, left, right) has weight 1, and all other positions are 0,

$$t \;=\; \begin{bmatrix} 0 & 1 & 0 \\ 1 & M & 1 \\ 0 & 1 & 0 \end{bmatrix},$$

Here, M controls how strongly the current pixel contributes, while 1s indicate the amount of influence (coupling) on the surrounding pixels, and 0s mean no coupling. The equalizer coefficients are updated using the Least Mean Squared (LMS) algorithm, based on the difference between the actual equalizer output $e[p, q]$ and the ideal PR-shaped reference $z[p, q]$.

The equalized signal is subsequently processed by a soft-output Viterbi algorithm (SOVA) detector [17], [20]. Detection is performed on a trellis structure defined by the PR target, enabling maximum-likelihood (ML) sequence estimation with soft outputs. The SOVA detector computes path metrics for

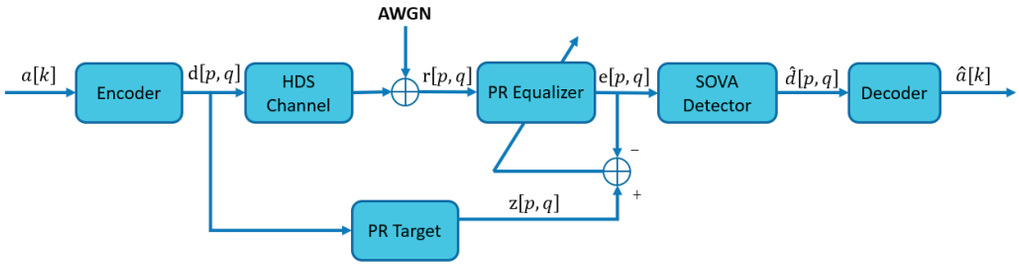


Fig. 4: Detection pipeline with 2D PR equalizer and SOVA.

all trellis states using soft decisions, and provides reliability information in the form of log-likelihood ratios (LLRs) for the output bits. At each state, the path metric (PM) is updated by adding the corresponding branch metric (BM), which quantifies the local mismatch between the received equalized symbol and the expected symbol predicted by the trellis. For a state transition, the BM is computed as the squared Euclidean distance between the equalizer output and the estimated level as follows,

$$\lambda(s_j, s_k) = \Big\{ e[p,q] - \Big(\hat{a}_{p,q+1}(s_j) + M \cdot \hat{a}_{p,q}(s_j) + a_{p,q-1}(s_k) \Big) \Big\}^2 \tag{6}$$

where s_j and s_k is the current state and the next state; and $\hat{a}(s_j)$ and $a(s_k)$ are decisions in these states, respectively. The SOVA decoder selects the trellis path minimizing the accumulated branch metric and outputs $\hat{d}[p,q]$. The decoder then applies the inverse modulation mapping to $\hat{d}[p,q]$, reconstructing the estimated user bitstream $\hat{a}[k]$.

V. Simulation Results and Discussion

To evaluate the performance of the proposed 10/12-rate modulation code, we conducted simulations over the 4-ary HDS channel model described in Section IV. We simulated about 10^9 bits in total, where each data page was configured to a size of 1024×1024 bits, with more than 1000 pages simulated. The performance of the proposed code is benchmarked against the prior-art 4/6-rate modulation code.

We first sweep $M \in [2, 18]$ and plot the BERs of the 4/6-rate and 10/12-rate codes. The simulation results are shown in Fig. 5. For very small $M = 2$, the center value has almost the same influence as its neighbors (each neighbor having a weight of 1). The target pixel is prone to errors caused by significant interference from adjacent pixels, which may become even more severe when aggregated. This drives the detector's signal-to-interference ratio very low, resulting in a variable BER. As M increases, the BER drops sharply for both schemes, reflecting the effectiveness of the detector in suppressing the ISI. Moreover, beyond $M = 6$, the BER improvement becomes marginal, indicating that the predictor has reached an optimal range. Throughout the entire range of M, the proposed 10/12-rate code consistently outperforms the 4/6-rate code, confirming that its improved code structure

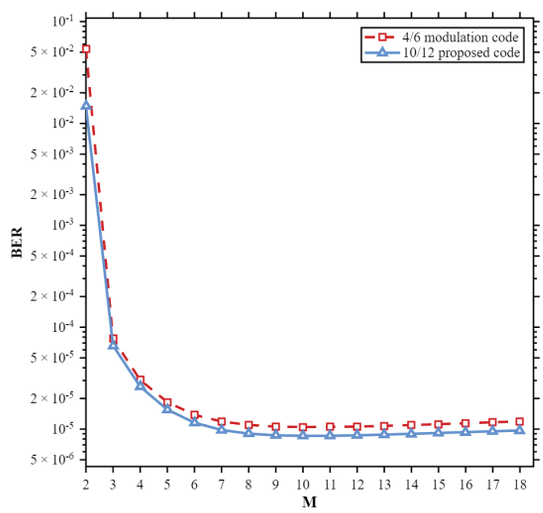


Fig. 5: BER as a function of the PR target's central coefficient M (at 12 dB SNR).

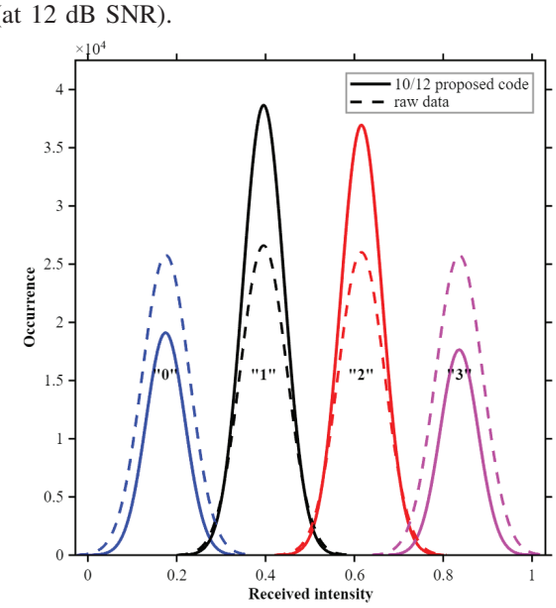


Fig. 6: Received intensity distributions.

enhances noise robustness and ISI tolerance. In the remaining simulations, we used $M = 10$.

We analyzed the histograms of the received intensities shown in Fig. 6. For the raw (uncoded) data (dashed lines), the

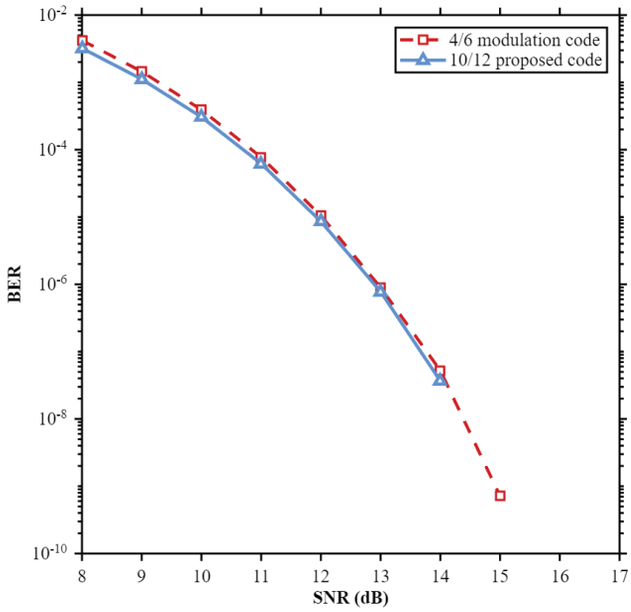


Fig. 7: BER across SNR (0% misalignment, $\sigma_b = 1.0$).

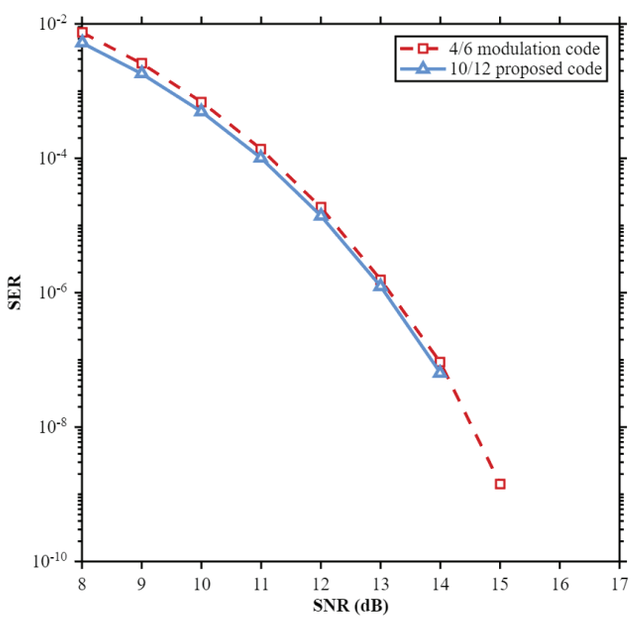


Fig. 8: SER performance across SNR (0% misalignment, $\sigma_b = 1.0$).

distributions for symbols 0, 1, 2, and 3 are broad and exhibit significant overlap. This indicates severe 2D ISI, making it difficult to distinguish between adjacent levels. In contrast, the 10/12-rate code produces a more distinct symbol distribution (solid line) while exhibiting a notably better balance between the (1, 2) and (0, 3) pairs. The reduction in overlap shows that the code's structural constraints successfully suppress the most detrimental interference patterns, yielding a cleaner, more distinguishable received waveform.

Fig. 7 and Fig. 8 present the BER and symbol error rate

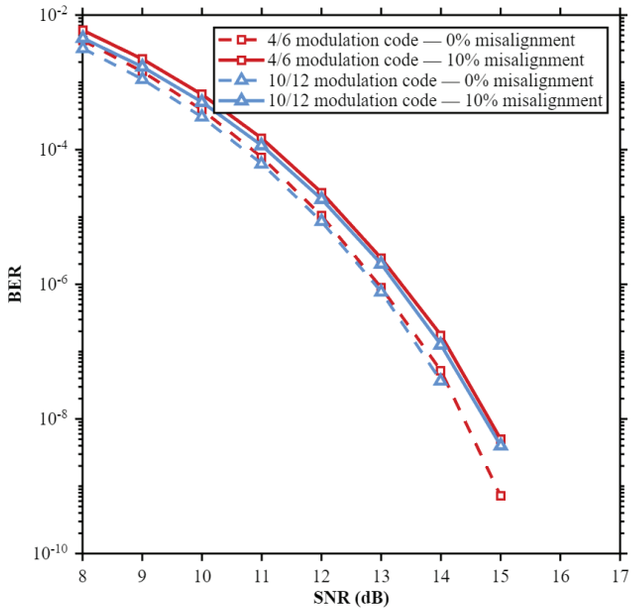


Fig. 9: BER performance under 10% misalignment ($\sigma_b = 1.0$).

(SER) performance as a function of SNR. The proposed 10/12-rate code consistently outperforms the 4/6-rate code. At BER 10^{-9}, the proposed code requires 1.0 dB less SNR. The performance improvement of the proposed code can be attributed to its better data balancing across each storage page. The balancing is physically important because the holographic storage records the page in the Fourier domain. When the symbol distribution is unbalanced, energy tends to concentrate in the zero-frequency component, known as the DC peak. This accumulation can saturate the dynamic range of the recording medium, where the interference patterns are stored, thereby attenuating the high-frequency components that actually carry the data information. As a result, the reconstructed data becomes blurrier, and the overall BER can raise significantly. By spreading the symbol energy more uniformly in space, the 10/12-rate code can naturally suppress the DC peak and preserve more optical dynamic range for data-carrying high-frequency components. Consequently, the channel observed by the detector becomes more linear and stable, resulting in more accurate signal reconstruction.

Fig. 9 evaluates the robustness of the proposed model under 10% misalignment. Both codes degrade, but the 10/12-rate code incurs a significantly smaller penalty. Fig. 10 shows the BER as a function of blur magnitude. Across all blur levels, the 10/12-rate design consistently outperforms the 4/6-rate code. Thus, the proposed code design delivers strong performance under normal channel conditions and maintains reliable quality under more severe channel. Finally, the superior performance aspects of the proposed code compared with the 4/6-rate code are briefly summarized in Table I.

The proposed 10/12-rate code achieves an effective balance between high code rate and strong suppression of 2D ISI, overcoming the limitations of prior approaches. The key

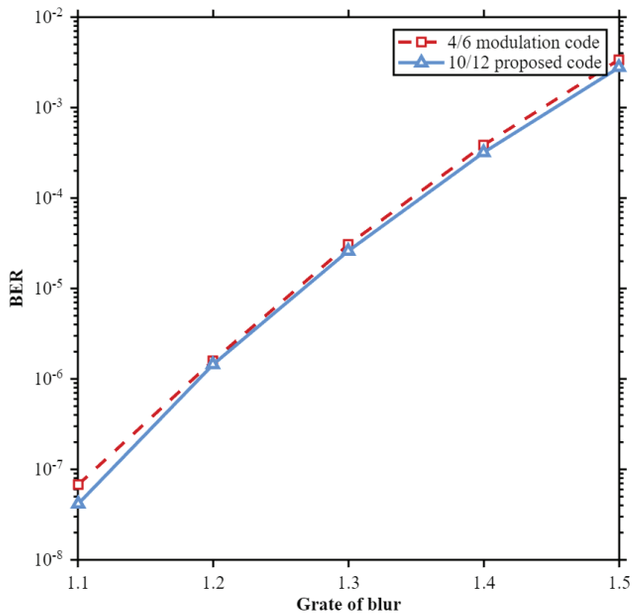


Fig. 10: BER performance across blur magnitude (at 15 dB, 0% misalignment).

TABLE I: Summary of performance gain of the proposed 10/12 modulation code

Item	Gain
Code rate	The proposed 10/12 modulation code achieves a 25% higher code rate (0.83 vs. 0.67) compared to the 4/6-rate code, enabling improved data efficiency without sacrificing robustness.
Low-SNR region (below 10 dB)	The proposed code achieves lower BER under the same SNR conditions, demonstrating improved ISI resistance.
High-SNR region (10 dB and above)	The advantage persists across most of this range; at BER = 10^{-9}, the proposed code requires about 1 dB lower SNR.
Overall performance	The proposed coding design achieves a low BER and offers a more favorable balance between efficiency and reliability.

innovation of the proposed coding lies in its 4×3 block layout featuring four anchor cells. This specific placement ensures that, even when blocks are tiled across a page, the probability of fatal interference is significantly reduced. The proposed scheme is also designed for high-speed parallel I/O, emphasizing minimal computational overhead. The encoder employs a direct mapping strategy for most of the data: eight symbols are mapped directly, and only two require simple modulo-2 operations and bit shifting. This design requires very few logic gates and no complex look-up tables. The decoding process is equally efficient and remains lossless. This logical comparison is computationally light, ensuring that decoding does not become a bottleneck in the retrieval pipeline.

VI. Conclusion

This paper presented a high-rate 10/12 4-ary modulation code for multilevel holographic data storage systems. The proposed code achieves higher data efficiency with reduced redundancy, while maintaining reliable BER performance under severe HDS channel impairments. The simulation results show that the proposed code scheme achieves about 1 dB SNR gain at BER = 10^{-9}. The proposal also delivers better performances than the conventional 4/6-rate code at 10% misalignment.

References

[1] K. Curtis, L. Dhar, A. Hill, W. Wilson, and M. Ayres, *Holographic Data Storage: From Theory to Practical Systems*. Hoboken, NJ, USA: John Wiley & Sons, 2010. doi: 10.1002/9780470666531.

[2] K. A. S. Immink, *Codes for Mass Data Storage Systems*, 2nd ed. Eindhoven, The Netherlands: Shannon Foundation Publishers, 2004. ISBN: 978-9074249270.

[3] H. J. Coufal, D. Psaltis, and G. T. Sincerbox, Eds., *Holographic Data Storage*. Springer Series in Optical Sciences, vol. 76. Berlin, Heidelberg: Springer, 2000. doi: 10.1007/978-3-540-47864-5.

[4] D. Reinsel, J. Gantz, and J. Rydning, "The Digitization of the World: From Edge to Core," IDC white paper sponsored by Seagate, 2018.

[5] L. Hesselink, S. S. Orlov, and M. C. Bashaw, "Holographic Data Storage Systems," *Proc. IEEE*, vol. 92, no. 8, pp. 1231–1280, 2004. doi: 10.1109/JPROC.2004.831212.

[6] J. W. Goodman, *Introduction to Fourier Optics*, 3rd ed. Greenwood Village, CO, USA: Roberts and Company, 2005.

[7] K. Rastani *et al.*, "Storage capacity and cross talk in angularly multiplexed holograms: two case studies," *Appl. Opt.*, vol. 32, no. 20, pp. 3772–3778, 1993. doi: 10.1364/AO.32.003772.

[8] T. Muroi, N. Kinoshita, N. Ishii, K. Kamijo, H. Kikuchi, Y. Kawata, and N. Shimidzu, "Optical compensation of hologram distortion avoiding interpage crosstalk on reconstructed image in angle-multiplexed holograms," *Appl. Opt.*, vol. 50, no. 29, pp. 5700–5709, 2011. doi: 10.1364/AO.50.005700.

[9] J. F. Heanue, M. C. Bashaw, and L. Hesselink, "Channel codes for digital holographic data storage," *J. Opt. Soc. Am. A*, vol. 12, no. 11, pp. 2432–2439, 1995. doi: 10.1364/JOSAA.12.002432.

[10] K. Koo, S.-Y. Kim, J. J. Jeong, and S.-W. Kim, "Data page reconstruction method based on two-dimensional soft output Viterbi algorithm with self reference for holographic data storage," *Opt. Rev.*, vol. 21, no. 5, pp. 591–596, 2014. doi: 10.1007/s10043-014-0094-x.

[11] S. Y. Kim and J. Lee, "A Simple 2/3 Modulation Code for Multi-Level Holographic Data Storage," *Jpn. J. Appl. Phys.*, vol. 52, no. 9, p. 09LE04, 2013.

[12] B. Kim, K. Park, and J. Lee, "4-level 6/9 Modulation Code for Holographic Data Storage," *J. Korea Inf. Commun. Soc.*, vol. 39A, no. 10, pp. 574–578, 2014. doi: 10.7840/kics.2014.39A.10.574.

[13] J. Kim, S. Jeong, and J. Lee, "4-ary 14/16 modulation code for reducing two-dimensional inter-symbol interference," *IET Commun.*, vol. 14, no. 8, pp. 1335–1339, 2020. doi: 10.1049/iet-com.2018.5710.

[14] C. D. Nguyen, N. X. Pham, C. C. Duong, and L. C. Nguyen, "Multilevel Modulation Coding for Four-Level Holographic Data Storage Systems," in *Proc. 2020 Int. Conf. Advanced Technologies for Communications (ATC)*, 2020, pp. 30–33.

[15] C. D. Nguyen, "Rate-4/5 4-ary modulation code for 4-level holographic data storage systems," *Cogent Eng.*, vol. 10, p. 2224140, 2023. doi: 10.1080/23311916.2023.2224140.

[16] M. Keskinoz and B. V. K. Vijaya Kumar, "Efficient modeling of volume holographic storage channels (VHSC)," in *Proc. SPIE Optical Data Storage 2000*, vol. 4090, pp. 205–210, Sep. 2000. doi: 10.1117/12.399348.

[17] C. D. Nguyen and J. Lee, "Iterative detection supported by extrinsic information for holographic data storage systems," *Electron. Lett.*, vol. 51, no. 23, pp. 1857–1859, Nov. 2015. doi: 10.1049/el.2015.2308.

[18] C. D. Nguyen and J. Lee, "8/10 two-dimensional modulation code for holographic data storage systems," *Electron. Lett.*, vol. 52, no. 9, pp. 710–712, 2016. doi: 10.1049/el.2016.0354.

[19] C. D. Nguyen, K. V. Vu, and K. Cai, "Two-dimensional weight-constrained codes for crossbar resistive memory arrays," *IEEE Commun. Lett.*, vol. 25, no. 5, pp. 1435–1438, 2021. doi: 10.1109/LCOMM.2021.3050159.

[20] A. J. Viterbi, "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm," *IEEE Trans. Inf. Theory*, vol. 13, no. 2, pp. 260–269, Apr. 1967. doi: 10.1109/TIT.1967.1054010.

Augmentation Effect on Indonesian Sign Language Image Classification

Stefani Siahaan
Biomedical Eng. Study Program
Telkom University
Purwokerto, Indonesia
2211108026@ittelkom-pwt.ac.id

Sevia Indah Purnama
Biomedical Eng. Study Program
Telkom University
Purwokerto, Indonesia
seviaindah@telkomuniversity.ac.id

Eko Fajar Cahyadi
Telecomm. Eng. Study Program
Telkom University
Purwokerto, Indonesia
ekofajarcahyadi@telkomuniversity.ac.id

Dodi Zulherman
Biomedical Eng. Study Program
CoE Healthcare and Well-Being Tech.
Telkom University
Purwokerto, Indonesia
zulherman@telkomuniversity.ac.id

***Abstract*—This study evaluates the effect of data augmentation on the performance of the EfficientNetB0 model to classify *Sistem Isyarat Bahasa Indonesia (SIBI)* gestures. A total of 1,035 images from six gesture classes were trained using transfer learning in two scenarios: without and with augmentation. The augmented model applied rotations, shifts, zooms, and brightness adjustments. The results show that data augmentation improved overall accuracy from 91.93% to 96.89%, with higher precision, recall, and F1-scores across all classes. The confusion matrix also showed a reduction in misclassification, especially for visually similar gestures. These findings confirm that applying data augmentation effectively enhances model generalization and robustness for limited sign language datasets.**

***Index Terms*—data augmentation, EfficientNetB0, transfer learning, CNN, Indonesian Sign Language**

I. Introduction

Hearing impairment is one of the most common disabilities globally, significantly affecting communication for millions of people. According to the World Health Organization (WHO), approximately 1.57 billion people experience hearing loss, with 430 million requiring rehabilitation services [1]. In Indonesia, the 2023 Indonesian Health Survey (SKI) reported that 0.4% of people aged one year and over have hearing disabilities, yet only 4.1% use hearing aids [2]. Furthermore, data from the Ministry of Social Affairs indicate that deaf individuals make up 7.03% of the total population of persons with disabilities [3]. These statistics underscore the urgent need for technological innovations that promote inclusive communication.

Sign language is the central visual-manual communication system for deaf people, using hand gestures, facial expressions, and body postures to convey meaning [4]. It is essential for participation in education, employment, and daily social interaction [5]. However, many hearing individuals are unfamiliar with sign language, resulting in communication gaps in public spaces such as schools, hospitals, and government offices [6]. This challenge motivates the development of automatic sign recognition systems, which are based on computer vision and deep learning.

Among deep learning architectures, Convolutional Neural Networks (CNNs) have been widely used for image-based classification due to their ability to capture spatial features [7], [8]. A study by Kumar *et al.* [9] showed that CNNs combined with Mediapipe achieved 99.95% accuracy for real-time American Sign Language (ASL) recognition. However, CNN-based systems still struggle to distinguish visually similar gestures, where slight variations in finger position or hand orientation can lead to misclassification [10], [11]. Gestures that require temporal information, such as ”J” or ”Z,” are also difficult to identify from static images [12]. In the Indonesian context, Lemmuela *et al.* [13] showed that variations in lighting and viewpoint reduce the performance of CNNs in gesture recognition.

For local sign language systems such as the Sistem Isyarat Bahasa Indonesia (SIBI), the available datasets are still small and limited. A study by Mayang Sari and Jamzuri only achieved 59% accuracy using Inception V3 with Random Forest, indicating that the limitations of the dataset limit CNN performance [14]. To mitigate this, transfer learning using pre-trained CNN architectures, such as VGG16, ResNet50, or EfficientNet-B0, trained on large datasets like ImageNet, allows for faster convergence and improved accuracy on small-scale datasets. Studies have shown that EfficientNet-B0 outperforms lighter models such as MobileNet in limited data scenarios, demonstrating its strong potential for visual classification tasks, including sign language recognition [15]. However, Juodelyte *et al.* [16] found that the performance of the model still heavily depends on the diversity and quality of the dataset.

To improve model generalization, data augmentation is used to synthetically expand the dataset using transformations such

as rotation, brightness, shift, and zoom [12]. A study by Perea-Trigo *et al.* [17] reported an accuracy increase of 11.7%, a study by Park and Sohn [18] increased accuracy from 41.9% to 96.2%, while Fadillah *et al.* [19] and Islam *et al.* [20] both confirmed substantial improvements in hand gesture recognition through augmentation.

Based on these findings, this study evaluates the effect of data augmentation on transfer learning-based CNN models, specifically EfficientNet-B0, for static SIBI gesture classification. The dataset used is a self-collected dataset containing six gestures with a high degree of similarity, used in two controlled scenarios—without and with augmentation—utilizing identical training parameters. The results are expected to reveal the contribution of data augmentation to the accuracy and robustness of the model, providing information to develop a vision-based assistive communication system for the deaf community in Indonesia.

II. Related Work

Several previous studies have explored the use of Convolutional Neural Networks (CNNs) and transfer learning in the recognition of Indonesian sign language (SIBI and BISINDO). Jollyta *et al.* [21] analyzed six pretrained models—VGG16, VGG19, ResNet50, DenseNet121, InceptionV3, and MobileNetV2—on the SIBI alphabet dataset and found that MobileNetV2 achieved the highest accuracy of 78%. Although transfer learning improves performance on limited datasets, the absence of data augmentation limits model generalizability.

Sihananto *et al.* [22] employed a conventional CNN without transfer learning, achieving an accuracy of 93.29% for SIBI and 82.32% for BISINDO. This study emphasized that dataset consistency and data preprocessing have a significant impact on accuracy, although performance decreases under uncontrolled conditions.

Al Fajri *et al.* [23] used MobileNetV2 with transfer learning and achieved 95.45% accuracy for SIBI recognition, demonstrating that pre-trained models can effectively address data scarcity by leveraging ImageNet features. Similarly, Amelia *et al.* [24] used the Xception architecture for BISINDO recognition that achieved high accuracy at a lower computational cost, with the potential of modern lightweight architectures. Regarding data augmentation, Mujahid *et al.* [25] showed that applying basic augmentation (rotation, flip, translation, and zoom) increased CNN accuracy by more than 10%, proving that increasing image diversity improves model robustness.

III. Methodology

This section discusses the research flow, dataset, data preprocessing, data augmentation, transfer learning model, and evaluation metric.

A. Research Flow

As shown in Fig. 1, the research procedure for this study is organized into multiple steps. Dataset collections, preprocessing, data augmentation, transfer learning model construction, and performance evaluation are among the phases. Each stage is intended to ensure that the model can accurately classify the SIBI hand gestures.

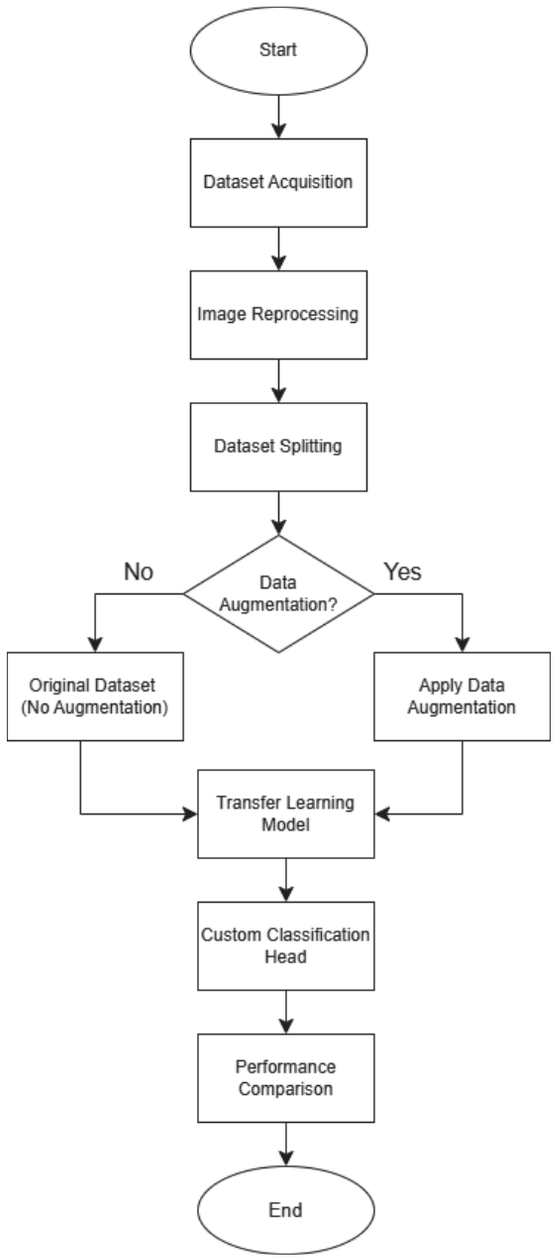


Fig. 1: Flowchart of research.

B. Dataset

The dataset used in this study consists of six static hand gesture classes from the SIBI, grouped into three pairs of gestures with a high degree of similarity: ”MAU” is visually identical to ”TERIMA”, ”JAGA” is visually identical to ”KASIH”, and ”ORANG” is visually identical to ”TETANGGA”. These six gestures were chosen for their visual similarity, which provides a challenge for the model in distinguishing subtle hand variations. All gesture images were collected from a single subject

using a 12-megapixel smartphone camera. The data acquisition process incorporated three clothing colors—red, white, and green—and two background types (plain and patterned) to introduce moderate visual diversity.

A total of 1,035 images were acquired, with approximately 160–170 images per class. Each image was resized to 224 × 224 pixels and normalized to a pixel value range of [0, 1]. The dataset was divided into 70% for training, 15% for validation, and 15% for testing to ensure a balanced and representative evaluation across gesture classes.



Fig. 2: Augmentation results.

C. Data Preprocessing

During the image preparation stage, all images are first resized to 224 × 224 pixels, and the pixel values are normalized to fall between 0 and 1. This process ensures consistent input dimensions and alignment with the EfficientNetB0 architecture, while also accelerating model convergence.

Subsequently, two experimental setups were defined: without augmentation and with augmentation. In the non-augmented setup, only the original images were used, while in augmented setup, new images were synthetically generated using data augmentation techniques to improve model generation.

D. Data Augmentation

Data augmentation was performed to artificially expand the dataset and improve generalization performance. The following augmentation techniques were applied:

- Rotation: 10°
- Width and height shift: 0.05
- Zoom Range: 0.05
- Brightness Range: [0.8 - 1.2]

These augmentations enable the model adapt to diverse gesture orientations and lighting variations. Fig. 2 shows examples of the augmented gesture images.

E. Transfer Learning Model

This study used EfficientNetB0 as the transfer learning backbone due to its efficiency and high performance through compound scaling, which balances network depth, width, and resolution. The model, pre-trained on ImageNet, was used as a frozen feature extractor to capture visual patterns of SIBI gesture images.

A custom classification head was added, consisting of a GlobalAveragePooling2D layer, followed by Dense layers (256 and 128 neurons, ReLU activation) with BatchNormalization and Dropout (0.5) to improve generalization. The final Softmax layer produced six gesture class probabilities. The model was trained using the Adam optimizer (learning rate = 1e-4) and Categorical Crossentropy loss, under identical configurations for both non-augmented and augmented datasets to fairly evaluate augmentation effects.

F. Evaluation Metrics

Model performance was assessed using several classification metrics, including accuracy, precision, recall, and F1-score. These metrics are calculated based on the confusion matrix and are defined as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

where TP, FP, and FN denote true positives, false positives, and false negatives, respectively. The results of these metrics provide a comprehensive understanding of the model's ability to classify the SIBI gestures effectively.

IV. Results and Discussion

The performance of the EfficientNetB0 model was evaluated under two conditions: training without data augmentation and with data augmentation. As shown in Fig. 3, the model trained without augmentation achieved a training accuracy of approximately 75% and a validation accuracy of approximately 92% after 45 epochs. With data augmentation applied, the accuracy improved, with training accuracy reaching 90% and validation accuracy stabilizing at 94–95%.

The loss curves also showed faster convergence. As shown in Fig. 4, with the training loss decreasing from 1.1 to 0.4, and the validation loss decreasing from 0.6 to 0.2. These results indicate that data augmentation effectively improves generalization and reduces overfitting, allowing the model to learn more robust gesture representations.

Further evaluation is presented in Fig. 5 and Fig. 6, which summarizes the classification results for both models. Before augmentation, the model achieved an overall accuracy of 91.93%, with a perfect F1-score (1.00) for the "JAGA" and

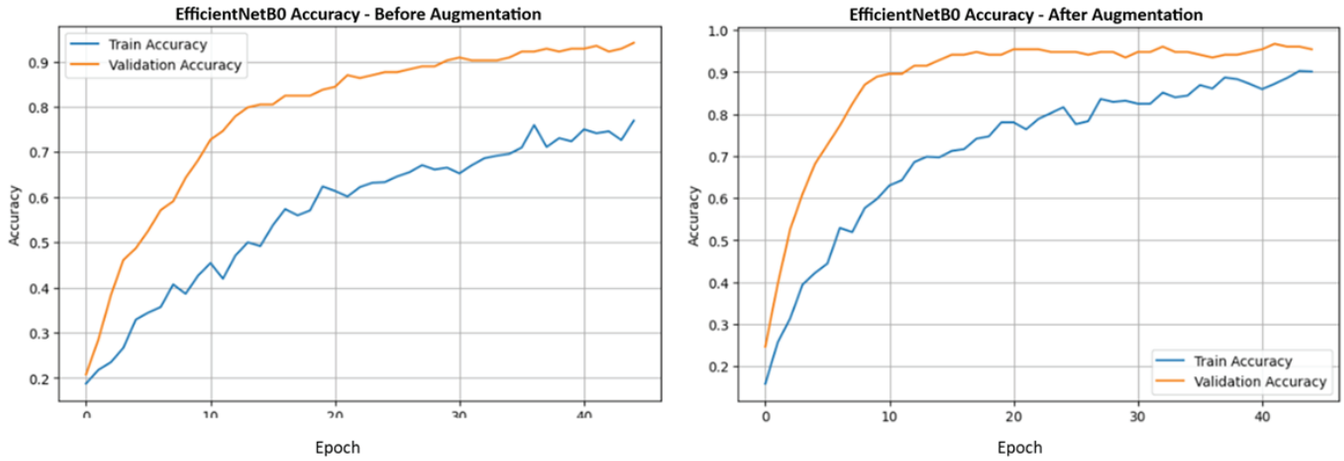


Fig. 3: Accuracy before and after augmentation.

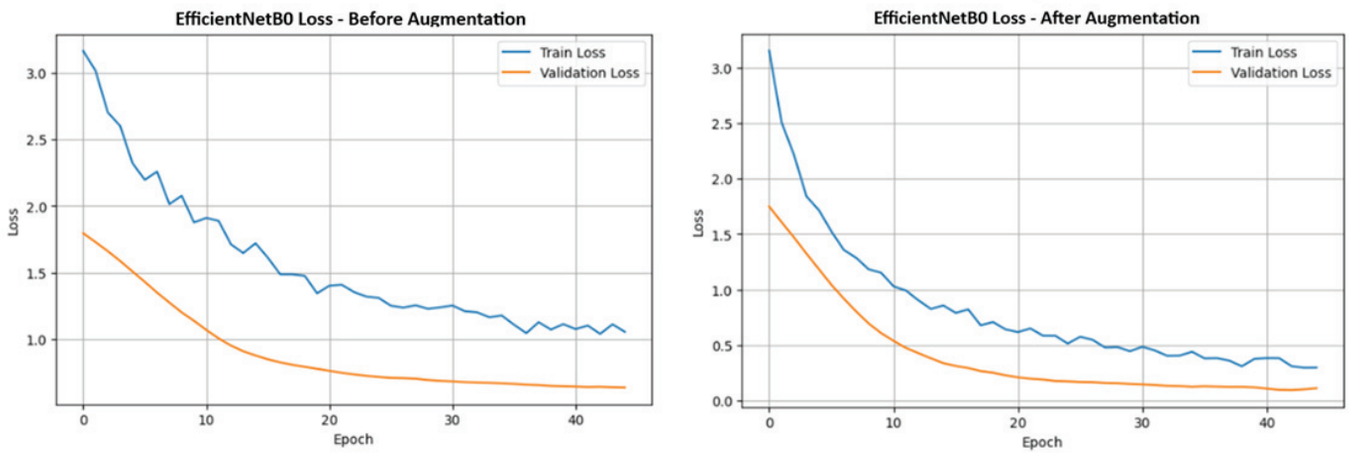


Fig. 4: Loss before and after augmentation.

```
============================================================
1. EfficientNetB0 Classification Report (Sebelum Augmentasi)
------------------------------------------------------------
              precision    recall  f1-score   support

        JAGA       1.00      1.00      1.00        27
       KASIH       0.87      1.00      0.93        26
         MAU       1.00      1.00      1.00        27
       ORANG       0.91      0.74      0.82        27
      TERIMA       1.00      0.89      0.94        27
   TETANGGA        0.77      0.89      0.83        27

    accuracy                           0.92       161
   macro avg       0.92      0.92      0.92       161
weighted avg       0.93      0.92      0.92       161

Akurasi Sebelum Augmentasi: 91.93%
============================================================
```

Fig. 5: Classification report before augmentation.

```
============================================================
2. EfficientNetB0 Classification Report (Sesudah Augmentasi)
------------------------------------------------------------
              precision    recall  f1-score   support

        JAGA       1.00      1.00      1.00        27
       KASIH       1.00      1.00      1.00        26
         MAU       1.00      1.00      1.00        27
       ORANG       0.93      0.93      0.93        27
      TERIMA       0.96      1.00      0.98        27
   TETANGGA        0.92      0.89      0.91        27

    accuracy                           0.97       161
   macro avg       0.97      0.97      0.97       161
weighted avg       0.97      0.97      0.97       161

Akurasi Sesudah Augmentasi: 96.89%
============================================================
```

Fig. 6: Classification report after augmentation.

”MAU” classes, but a lower performance for the ”ORANG” (F1 = 0.82) and ”TETANGGA” (F1 = 0.83) classes. After augmentation, all performance metrics improved, achieving an overall accuracy of 96.89%. Precision, recall, and F1-score averaged 0.97, with the ”JAGA”, ”KASIH”, and ”MAU” classes showing excellent classification results. This indicates that the augmented data helped the model recognize gesture variations more effectively, especially in classes that previously showed confusion.

The confusion matrix shown in Fig. 7 further validates this improvement. Before augmentation, most misclassifications occurred for visually similar gestures such as ”ORANG” and ”TETANGGA”. After augmentation, these errors decreased substantially, especially for ”ORANG”, whose recognition rate

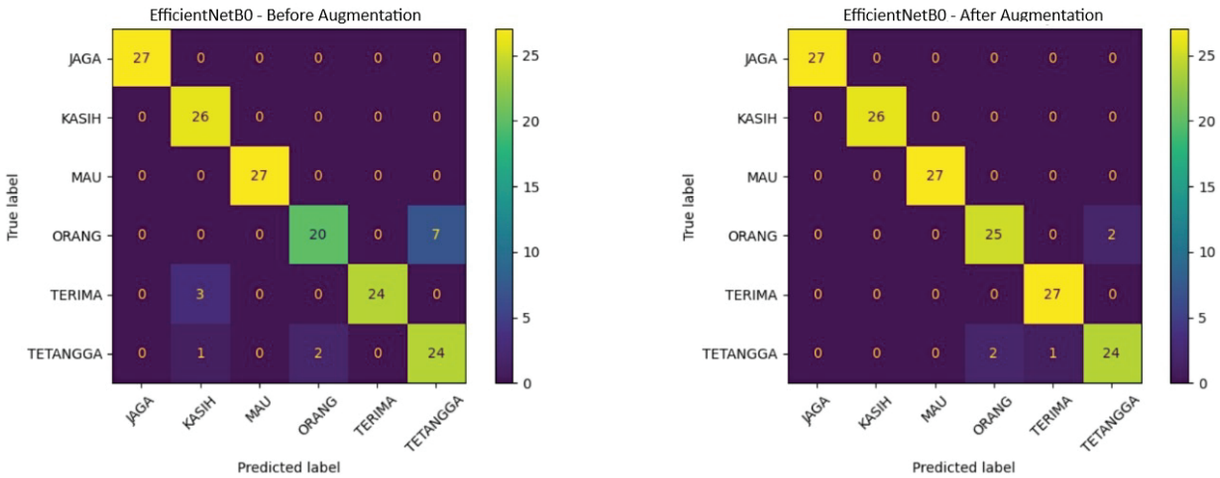


Fig. 7: Confusion matrix before and after augmentation.

increased from 74% to 93%. Meanwhile, ”JAGA”, ”KASIH”, and ”MAU” achieved perfect predictions with no misclassifications. Overall, the model generalizability improved by nearly 5%, indicating that integrating geometric and brightness variations during training effectively improves the classification performance on small-scale gesture datasets.

This study confirms that the combination of transfer learning with EfficientNetB0 and on-the-fly data augmentation improves the robustness and consistency of the model in recognizing SIBI gestures. Improvements across evaluation metrics demonstrate the augmentation's efficiency in addressing data limitations and visual variation issues commonly found in sign language datasets.

V. Conclusions and Future Scope

This study demonstrates that data augmentation substantially improves the performance of the EfficientNetB0 model in recognizing SIBI gestures. By applying changes such as rotating, shifting, zooming, and adjusting brightness, the model learned to recognize more varied features. This reduced the tendency of the model to overfit the training data. Because of this, the model's overall accuracy increased from 91.93% to 96.89%, along with improvements in precision, recall, and F1-score. Looking at the confusion matrix also showed that mistakes between similar gestures became much less common after using data augmentation. This shows that combining transfer learning with good data augmentation can make models more reliable and adaptable, especially when working with smaller gesture datasets.

For future work, the study suggests trying more advanced data augmentation methods, such as random cropping or color changes, increasing the variety of the data set, and testing other transfer learning models, such as ResNet or MobileNet for comparison. In addition, building the system to recognize signs in real-time or translate sign language from videos could open up more possibilities for helping the deaf community in Indonesia through assistive communication tools.

Acknowledgment

This study is fully supported by Telkom University.

References

[1] World Health Organization, “Deafness and hearing loss,” [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss, 2024, accessed: Oct. 14, 2025.

[2] Kementerian Kesehatan Republik Indonesia, “Kemenkes ingatkan bahaya gangguan pendengaran: Deteksi dini dan pencegahannya,” [Online]. Available: https://teropongbisnis.id/kemenkes-ingatkan-bahaya-gangguan-pendengaran/, 2023, accessed: Oct. 14, 2025.

[3] Kementerian Sosial Republik Indonesia, “Sistem informasi manajemen penyandang disabilitas: Data penyandang tuna rungu,” Kementerian Sosial Republik Indonesia, Indonesia, Tech. Rep., 2024, accessed: Oct. 14, 2025.

[4] W. C. Stokoe Jr, “Sign language structure: An outline of the visual communication systems of the american deaf,” *Journal of Deaf Studies and Deaf Education*, vol. 10, no. 1, pp. 3–37, 2005.

[5] S. Isnaniah, T. Agustina, Islahuddin, and F. Annisa, “The use of sign language in deaf indonesian classrooms in surakarta,” *Jurnal Keilmuan Bahasa, Sastra, dan Pengajarannya*, vol. 9, no. 2, pp. 468–481, 2023.

[6] B. R. Fajri and G. Kusumastuti, “Perceptions of hearing people on sign language learning,” in *5th International Conference on Education and Technology (ICET 2020)*. Atlantis Press, 2020, pp. 364–367.

[7] S. Tan, N. Khan, Z. An, Y. Ando, R. Kawakami, and K. Nakadai, “A review of deep learning-based approaches to sign language processing,” *Advanced Robotics*, vol. 38, no. 23, pp. 1649–1667, 2024.

[8] X. Zhao, L. Wang, Y. Zhang, X. Han, M. Deveci, and M. Parmar, “A review of convolutional neural networks in computer vision,” *Artificial Intelligence Review*, vol. 57, no. 4, p. 99, 2024.

[9] R. Kumar, A. Bajpai, and A. Sinha, “Mediapipe and cnns for real-time asl gesture recognition,” *arXiv preprint arXiv:2305.05296*, 2023.

[10] D. Kumari and R. S. Anand, “Isolated video-based sign language recognition using a hybrid cnn-lstm framework based on attention mechanism,” *Electronics*, vol. 13, no. 7, p. 1229, 2024.

[11] K. Nasreddine and A. Benzinou, “Shape geodesics for robust sign language recognition,” *IET Image Processing*, vol. 13, no. 5, pp. 825–832, 2020.

[12] C. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” *Journal of Big Data*, vol. 6, no. 1, pp. 1–48, 2020.

[13] I. V. Lemmuela, M. Ayub, and O. Karnalim, “Dynamic sign language recognition in bahasa using mediapipe, long short-term memory, and convolutional neural network,” *Journal of Information Systems Engineering and Business Intelligence*, vol. 11, no. 1, pp. 17–29, 2025.

[14] M. Sari and E. R. Jamzuri, "Hand sign recognition of indonesian sign language system sibi using inception v3 image embedding and random forest," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 9, no. 2, pp. 258–265, 2025.

[15] H. Pu and K. Yi, "A comparative analysis of efficientnet and mobilenet models' performance on limited datasets: an example of american sign language alphabet detection," *Highlights Sci Eng Technol*, vol. 94, pp. 558–64, 2024.

[16] D. Juodelyte, A. Jiménez-Sánchez, and V. Cheplygina, "Revisiting hidden representations in transfer learning for medical imaging," *arXiv preprint arXiv:2302.08272*, 2023.

[17] M. Perea-Trigo, E. J. López-Ortiz, L. M. Soria-Morillo, J. A. Álvarez-García, and J. Vegas-Olmos, "Impact of face swapping and data augmentation on sign language recognition," *Universal Access in the Information Society*, vol. 24, no. 2, pp. 1283–1294, 2025.

[18] C.-I. Park and C.-B. Sohn, "Data augmentation for human keypoint estimation deep learning based sign language translation," *Electronics*, vol. 9, no. 8, p. 1257, 2020.

[19] R. Z. Fadillah, A. Irawan, and M. Susanty, "Data augmentasi untuk mengatasi keterbatasan data pada model penerjemah bahasa isyarat indonesia (bisindo)," *Jurnal Informatika*, vol. 8, no. 2, pp. 208–214, 2021.

[20] M. Z. Islam, M. S. Hossain, R. Ul Islam, and K. Andersson, "Static hand gesture recognition using convolutional neural network with data augmentation," in *2020 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2020 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*. IEEE, 2020, pp. 324–329.

[21] D. Jollyta, P. Prihandoko, J. Johan, W. Ramdhan, and E. Santoso, "Transfer learning model evaluation on cnn algorithm: Indonesian sign language system (sibi)," *Journal of Applied Business and Technology*, vol. 6, no. 2, pp. 83–92, 2025.

[22] A. N. Sihananto, E. M. Safitri, Y. Maulana, F. Fakhruddin, and M. E. Yudistira, "Indonesian sign language image detection using convolutional neural network (cnn) method," *Inspiration: Jurnal Teknologi Informasi dan Komunikasi*, vol. 13, no. 1, pp. 13–21, 2023.

[23] F. P. Al Fajri, A. R. Adriansyah, and S. Munir, "Model deep learning untuk penerjemah bahasa isyarat sibi dengan arsitektur transfer learning," *Journal of Digital Business and Technology Innovation*, vol. 2, no. 1, pp. 133–139, 2025.

[24] M. V. Amelia, W. S. J. Saputra, K. M. Hindrayani, and P. A. Riyantoko, "Indonesian sign language (bisindo) classification using xception transfer learning architecture," *International Journal of Advances in Data and Information Systems*, vol. 6, no. 2, pp. 320–331, 2025.

[25] A. Mujahid, M. J. Awan, A. Yasin, M. A. Mohammed, R. Damaševičius, R. Maskeliūnas, and K. H. Abdulkareem, "Real-time hand gesture recognition based on deep learning yolov3 model," *Applied Sciences*, vol. 11, no. 9, p. 4164, 2021.

Development and Evaluation of a Low-Cost EMG-Based System for Muscle Weakness Detection

1st Kezia Hening Ata *, 2nd Sevia Indah Purnama †, 3rd Fikra Titan Syifa ‡,

* Department of Biomedical Engineering, Telkom University, Purwokerto 53147, Jawa Tengah, Indonesia

†Department of Electrical Engineering, Telkom University, Purwokerto 53147, Jawa Tengah, Indonesia

‡Department of Biomedical Engineering, Telkom University, Purwokerto 53147, Jawa Tengah, Indonesia

Emails: 1st keziaheningata@student.telkomuniversity.ac.id, 2nd seviaindah@telkomuniversity.ac.id, 3rd fikras@telkomuniversity.ac.id

***Abstract*—** Stroke remains a leading cause of long-term motor disability, often resulting in muscle weakness that affects both contralesional and ipsilesional limbs. Objective monitoring of muscle strength is essential for effective rehabilitation and progress evaluation. This study presents the development of a prototype EMG-based muscle weakness assessment system designed for potential application in post-stroke evaluation. The system integrates a MyoWare EMG sensor with an ESP32 microcontroller to capture muscle activity signals and compute Root Mean Square (RMS) values as quantitative indicators of muscle strength. Testing results demonstrate clear differentiation between relaxation and contraction states, with RMS values of 121.15 ADC units during relaxation and 2629.72 ADC units during contraction. The prototype exhibits stable signal acquisition, adequate sensitivity, and real-time processing capabilities suitable for preliminary muscle activity assessment. While current validation is limited to normal muscle simulations, the system shows promising feasibility as a low-cost, portable monitoring tool that could support rehabilitation decision-making and provide objective feedback in post-stroke therapy sessions.

***Keywords*—** *electromyography, muscle weakness assessment, post-stroke rehabilitation, root mean square, EMG signal processing, neuromuscular monitoring*

I. INTRODUCTION

Stroke remains one of the leading causes of mortality and long-term disability worldwide. The World Health Organization (WHO) reports that approximately 15 million people experience stroke annually, with 5 million deaths and another 5 million individuals living with permanent disabilities [1]. In Indonesia, the 2020 National Health Profile recorded 1,789,261 people who had experienced or were currently living with stroke, illustrating the significant national burden of this condition [2]. Post-stroke impairments commonly manifest as motor deficits such as hemiparesis and reduced muscle strength, both of which greatly impact daily functional activities.

These motor impairments are not limited to the contralesional side but may also occur on the ipsilesional limb due to the bilateral contribution of both cerebral hemispheres in motor control. As a result, stroke survivors may experience diminished force production and altered muscle activation patterns on both sides of the body, contributing to functional limitations [3]. Objective monitoring of muscle strength is therefore essential in rehabilitation to evaluate patient progress and identify potential asymmetries in neuromuscular activation.

Surface electromyography (sEMG) is widely used to assess muscle activation by recording electrical signals produced during contraction. The Root Mean Square (RMS) parameter is commonly used because it represents the energy of the EMG signal and reflects the degree of motor unit activation within a specific time window [4]. RMS is useful for quantifying muscle strength, assessing neuromuscular performance, and detecting asymmetry between limbs, particularly in conditions involving motor impairment such as stroke [5].

Advancements in low-cost sensors and embedded systems enable the development of simple EMG-based devices that can serve as preliminary monitoring tools in various settings. This study presents a prototype EMG-based muscle weakness assessment system capable of capturing EMG amplitude and generating real-time RMS output. Although this stage of research does not involve stroke patients, the prototype is designed as an initial step toward post-stroke evaluation applications, focusing on signal acquisition quality and system feasibility for future clinical validation.

II. RELATED WORK

A. EMG in Clinical and Rehabilitation Contexts

Electromyography (EMG) has long been used to examine neuromuscular function in both clinical diagnostics and rehabilitation research. Studies consistently show that surface EMG provides reliable information on muscle activation patterns, enabling the evaluation of motor performance and strength deficits in various conditions, including stroke-related impairments [5][6]. RMS is one of the most widely

applied amplitude features because it reflects signal energy and correlates strongly with muscle force generation during voluntary contraction [8].

B. Post-Stroke Muscle Weakness and Activation Patterns

Previous research indicates that stroke affects motor control on both sides of the body, challenging the earlier assumption that the ipsilesional limb remains fully intact. Several studies have documented measurable strength deficits and altered activation patterns in the limb on the same side as the brain lesion due to bilateral motor representation in the central nervous system [7]. These findings underscore the importance of monitoring both limbs to identify asymmetric activation, which is a common indicator of neuromuscular dysfunction in post-stroke patients [9].

C. Technological Approaches for EMG-Based Assessment

A variety of technological solutions have been proposed for EMG-based monitoring, ranging from laboratory-grade acquisition systems to low-cost microcontroller-based prototypes. Many designs incorporate surface electrodes, basic filtering stages, and digital processing algorithms to extract key parameters such as RMS or mean absolute value [6][8]. Despite their usefulness, numerous existing systems still rely on external hardware, high-resolution ADCs, or complex computational steps that may reduce practicality, especially in low-resource clinical environments.

D. Research Gaps and Motivation for the Proposed System

Although prior studies have highlighted the relevance of EMG in detecting post-stroke muscle weakness, there remains a need for systems that are low-cost, easy to operate, portable, and capable of providing immediate quantitative feedback [7]. Few studies combine simple hardware, real-time RMS computation, and direct visualization suitable for preliminary stroke screening. This gap motivates the development of the proposed EMG-based muscle weakness assessment prototype, which aims to support early evaluation and potentially enhance rehabilitation decision-making [9].

III. METODOLOGY

A. Process Flow

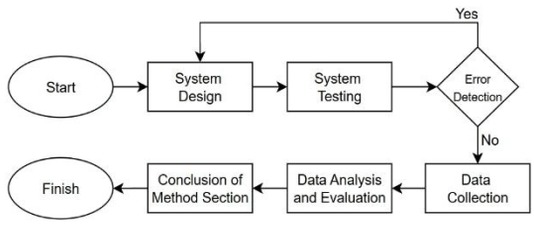


Figure 1. Block Diagram fot proposed methodology

The system design included defining the architecture and integrating the EMG sensor with the processing module to ensure stable signal acquisition. The system was then tested to verify its functionality and identify whether any errors occurred during operation. After confirming that the device worked properly, data were collected according to the measurement protocol and later analyzed to assess signal quality and muscle weakness indicators. The hardware components were carefully selected to meet the specifications required for accurate EMG signal detection, including appropriate electrode types, amplifier circuits with sufficient gain, and analog-to-digital converters with adequate sampling rates. The integration phase involved establishing proper communication protocols between the sensor unit and the processing module to minimize signal latency and ensure data integrity throughout the acquisition pipeline. During the testing phase, multiple validation procedures were conducted under controlled conditions to evaluate system stability, repeatability, and responsiveness across different muscle activation levels. The data collection protocol was designed to capture both baseline relaxation signals and various contraction intensities, providing a comprehensive dataset for subsequent analysis. Signal processing techniques, including filtering, feature extraction, and statistical analysis, were systematically applied to quantify key parameters such as amplitude, root mean square (RMS) values, and frequency domain characteristics. The results of this analysis were used to evaluate overall system performance and was also used as a basis for drawing overall system performance weakness assessment.

B. Hardware Components

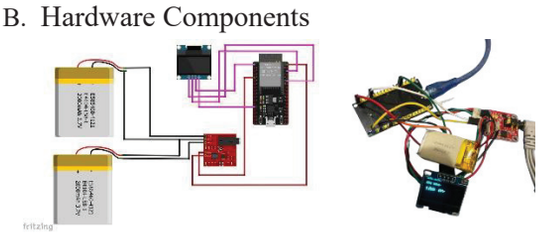


Figure 2. Hardware Wiring and Prototype Assembly

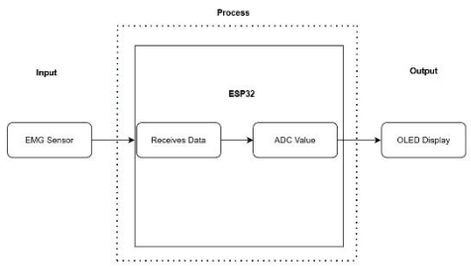


Figure 3. Block Diagram of EMG Data Processing Using ESP32

The system's hardware implementation operates as a stand-alone unit designed for the local acquisition and processing of muscle activity data. The flow begins with the MyoWare EMG Sensor, which serves as the signal acquisition unit, capturing the muscle contraction's analog signals. This conditioned signal is then fed into the ADC pin (Analog-to-Digital Converter) on the ESP32 Microcontroller. The ESP32 acts as the core processing unit, responsible for digitizing the signal, performing the Root Mean Square (RMS) Calculation to quantify muscle strength, and controlling the overall system. The resulting RMS data is then transmitted via the I2C communication interface from the ESP32 to the OLED Display, which serves as the visual output interface for real-time display to the user.

C. Software

The software architecture manages the acquisition and processing of the EMG signal using the Arduino IDE environment. The program continuously reads the analog input from the EMG sensor, performs basic signal stabilization, and calculates the Root Mean Square (RMS) as the primary indicator of muscle activity. The RMS value is then compared with predefined thresholds to evaluate muscle activation levels, and the results are displayed in real time on the OLED screen. This workflow ensures consistent operation from data acquisition to on-device visualization.

IV. RESULT AND DICUSSION

A. System Testing Results



Figure 4. EMG Signal During Relaxation Using ESP32 Serial Plotter

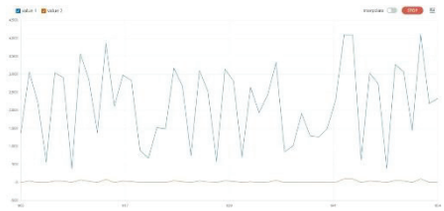


Figure 5. EMG Signal During Contraction Using ESP32 Serial Plotter

Table 1 ADC Raw Data and RMS Calculation

Time(s)	1	1.5	2	2.5	3	RMS (ADC)
Relax	103	89	142	173	69	121.15
Contraction	1573	2093	2864	2371	3728	2629.72

Figure 4 illustrates the EMG signal during the relaxation phase, where the captured waveform exhibits a highly stable, low-amplitude baseline, reflecting the minimal electrical activity, or quiescence, of the muscle. Conversely, Figure 5 depicts the dynamic contraction phase, showing a pronounced and rapid increase in amplitude attributable to significantly heightened neuromuscular activation and maximal motor unit recruitment. The comparison of raw numerical data from both conditions, detailed in Table 1, shows a stark contrast: ADC values during the relaxed state range minimally between 69 and 173, indicative of the system's inherent noise level. However, during contraction, the ADC values dramatically escalate to a maximum range of 1,573 to 3,728. This substantial difference in signal intensity is quantitatively validated by the Root Mean Square (RMS) calculation, which is the primary metric for muscle strength. The RMS value stabilizes at approximately 121.15 ADC units for relaxation and sharply rises to 2,629.72 ADC units for contraction. This significant disparity confirms that the developed prototype possesses both the sensitivity and resolution to accurately differentiate between low-level and high-level muscle activation states. The ability of the system to effectively capture the transition between inactive and fully activated conditions, evidenced by the high reliability of the RMS metric, strongly supports its potential utility in the objective and quantitative monitoring of post-stroke motor function recovery.

B. Technical Performance Analysis

Based on the EMG signal testing results, the system demonstrates stable technical performance and characteristics consistent with typical EMG signals. In Figure 4, the relaxation signal exhibits low and relatively

constant amplitude, indicating an inactive muscle state and minimal noise interference. This baseline stability also indicates proper electrode placement and optimal filtering system performance in suppressing electromagnetic interference. Conversely, in Figure 5, the contraction signal shows a significant increase with higher and more fluctuating amplitude patterns, demonstrating motor unit activation consistent with EMG physiological theory. This amplitude increase results from motor unit recruitment during voluntary contraction, where stronger contractions activate more motor units. No abnormal signal spikes were detected during testing, indicating that the system can responsively record changes in muscle activity without significant technical interference. The favorable signal-to-noise ratio is evident from the clear distinction between both conditions, validating that the amplifier gain has been properly configured. Thus, the signal stability and recorded activation patterns prove that this device functions effectively and can accurately represent both relaxation and contraction muscle conditions.

C. Feasibility and Limitations

Based on the EMG signal testing results demonstrating the system's capability to distinguish between relaxation and contraction muscle conditions through differences in amplitude values and RMS calculations, this device is considered to have potential for use as a muscle strength monitoring system in post-stroke patients. The device can capture changes in muscle activity in real-time, thus enabling its utilization in rehabilitation process monitoring. The system's responsiveness in detecting transitions from relaxation to contraction conditions also indicates that this device can provide direct feedback to patients and therapists during therapy sessions. Nevertheless, the system still has several limitations, such as sensitivity to noise caused by electrode movement or environmental interference, as well as potential reading errors if the sensor position is not optimally attached. Moreover, testing was only conducted on normal muscle simulations, therefore clinical validity in stroke patients cannot be comprehensively confirmed. The characteristics of EMG signals in stroke patients, which generally exhibit lower amplitude and irregular activation patterns compared to normal muscles, need to be considered in further development. For further implementation, calibration processes, testing with larger sample sizes, and integration of signal filtering mechanisms are required to improve the accuracy and stability of the device. Implementation of adaptive filtering algorithms and standardized electrode placement techniques can also help minimize measurement variability. Comparative studies with commercially validated clinical EMG devices need to be conducted to ensure system reliability. Furthermore, trials across various stroke severity levels and diverse age groups will provide more comprehensive data regarding the device's effectiveness in actual medical rehabilitation contexts.

D. Discussion

The results indicate that the developed EMG monitoring prototype successfully differentiates muscle activation levels, as demonstrated by the distinct separation between relaxation and contraction phases. The recorded ADC values show minimal activity during relaxation and significantly higher amplitudes during contraction, which is further validated through RMS calculations presented in Table 1. These findings confirm that the system provides adequate measurement sensitivity and stability for basic EMG evaluation. However, while the current setup demonstrates reliable technical performance, improvements are needed in signal filtering, calibration, and validation using clinical stroke patient data to enhance accuracy and applicability. Overall, these results support the feasibility of using the system as a preliminary muscle activity assessment tool and provide a foundation for further development toward clinical rehabilitation monitoring.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the supervising lecturer for the guidance, valuable insights, and continuous support provided throughout the research process, particularly in the development and implementation of the IoT-based system. Appreciation is also extended to the authors' families and colleagues for their encouragement and moral support, which greatly contributed to the completion of this work.

REFERENCES

[1] M. R. Oktavian, A. S. Wibowo, and S. D. Sari, "Rancang Bangun Sistem Monitoring Rehabilitasi Lutut Menggunakan Sensor EMG dan Accelerometer," Jurnal Fisika dan Aplikasinya, vol. 19, no. 2, pp. 48–57, 2023

[2] R. R. Rachman, D. Sudarmaji, and H. Haryanto, "Pengukuran Aktivitas Otot Menggunakan Sensor EMG Sebagai Sistem Monitoring Latihan Pada Pasien Stroke," Jurnal ELINVO (Elektronika, Informatika, dan Vokasional), vol. 7, no. 1, pp. 1–7, May 2022.

[3] A. Pradana, R. N. Ismanto and I. R. Santoso, "Design and Implementation of a Real Time EMG Signal Classification System Based on Arduino for Rehabilitation Applications," Int. J. Electr. Comput. Eng. (IJECE), vol. 24, no. 3, pp. 1515–1522, 2024

[4] N. F. Azzahra, "Pengembangan Sistem Berbasis Elektromiografi (EMG) dan Inersia untuk Menilai Tingkat Kelemahan Otot Pasien Stroke," Skripsi, Universitas Airlangga, Surabaya, 2025.

[5] H. Winarno, L. Embeded, K. Keilmuan, and E. Biomedik, "Analisa Sinyal EMG Dengan Metode Continuous Wavelet Transform pada Pasien Pasca Stroke," 2017.

[6] H. Araiza Lizarde and D. Lorias-Espinoza, "Analysis of Surface Electromyographic Signals

(Semg) Using Labview's Biomedical Toolkit," Jul. 2024

[7] L. Pellegrino, M. Coscia, P. Giannoni, L. Marinelli, and M. Casadio, "Stroke impairs the control of isometric forces and muscle activations in the ipsilesional arm," Scientific Reports, vol. 11, no. 1, Sep. 2021

[8] K. Watanabe, M. Kouzaki, M. Ogawa, H. Akima, and T. Moritani, "Relationships between muscle strength and multi-channel surface EMG parameters in eighty-eight elderly," European Review of Aging and Physical Activity, vol. 15, no. 1, Apr. 2018

[9] J. Tchimino, J. L. Dideriksen, and S. Dosen, "EMG feedback improves grasping of compliant objects using a myoelectric prosthesis," Journal of NeuroEngineering and Rehabilitation, vol. 20, no. 1, Sep. 2023

Realtime 3D Animation Capture Using Pose Estimation and Video Segmentation with YOLOv11n-Seg model

1st Steven *, 2nd Andry Chowanda †
BINUS Graduate Program – Master Of Computer Science, Jakarta, Indonesia
Emails: *steven122@binus.ac.id, 2nd achowanda@binus.edu*

***Abstract*— Human Pose Estimation (HPE) is a foundational task in computer vision, critical for fields such as game development, augmented reality, and healthcare. While many state-of-the-art models demonstrate high accuracy on standard datasets, adapting them for real-time applications remains computationally challenging. This research proposes YOLO11-OpenGCN, a unified lightweight cascade pipeline designed for efficient 3D animation capture using a custom 25-keypoint skeleton. The proposed architecture consists of three stages: 1 a pre-trained YOLOv11-Nano Segmentation model for human detection; 2 a lightweight OpenPose model with a MobileNetV3 backbone for 2D pose estimation; and 3 a Semantic Graph Convolutional Network (SemGCN) for 2D-to-3D pose lifting. The model was fine-tuned on the CHI3D dataset (50,000–100,000 images) using multi-view data and extensive augmentation to ensure robustness. Experimental results demonstrate that YOLO11-OpenGCN achieves significant improvements in speed over existing baselines, reaching an inference speed of 15.74 FPS this represent five-fold improvement compared to 3.15 FPS for comparable frameworks. The model achieved an Average Precision (AP) of 80.2% on the validation set, offering a viable solution for real-time motion capture on consumer hardware.**

Keywords—3D Human Pose Estimation, Real-time Animation, YOLOv11, OpenPose, Graph Convolutional Networks (GCN), Lightweight Architecture, Cascade Pipeline

I. INTRODUCTION

Animation plays a crucial role in game development, significantly impacting player interaction in genres such as puzzles, RPGs, action, sports, and strategy games [1]. Motion capture technology is essential in this domain to ensure in-game actions appear realistic and immersive [2].

Other previous research, has explored and developed their own framework using a pre-trained YOLOv5 model for human detection, combined with a High-Resolution Network (HRNet) to detect 2D poses, and integrated with a Temporal Convolutional Model (TCM) to detect 3D poses. By using the Human3.6M and COCO datasets, they achieved high accuracy for assessment in sports[3].

Other research by utilized the G2O-Pose algorithm for 3D Pose estimation, which employs reverse joint correction using the Human3.6M dataset. This reverse joint correction algorithm uses a heuristic search to detect, correct, and suppress unnatural poses, with an average correction accuracy rate reaching over 80 [4].

Furthermore, other research conducted model improvements for a Deep Convolutional Neural Network (D-CNN) using pose extraction. A Region-based CNN (R-CNN) was used for the initial detection of the human body from images. Subsequently, an enhanced D-CNN

combining several convolutional and pooling layers was used to precisely extract feature points from the pose. Finally, a pose recognition algorithm based on feature point matching was used to reconstruct the 3D pose. [5]

There is also innovative research in the field of motion

generation, which directly uses generative models. A prominent example is the research by[6], who developed a Transformer-based model, similar to GPT-2, to synthesize realistic human movements. In this study, they introduced two models, AnimGPT and DenoisAnimGPT.

Both models were trained on the CMU Motion Capture dataset, a database rich in complex and diverse human motion sequences. Unlike pose estimation methods that reconstruct motion from video, AnimGPT and DenoisAnimGPT learn to understand the patterns, rhythms, and transitions in motion data to generate new, previously unseen movement sequences.

Human Pose Estimation and video segmentation are among the most active research areas in computer vision. These technologies are applied in diverse fields including healthcare, sports analysis, and activity recognition [7]. Specifically, combining HPE with segmentation allows for high-fidelity motion capture, which is critical for realistic action synthesis. HPE is defined as the process of localizing human joints (keypoints) in images or videos. There are two primary approaches: 2D HPE and 3D HPE [8], where 3D estimation results are often highly dependent on the accuracy of the intermediate 2D pose.

In this paper, we focus on both 2D and 3D pose estimation to achieve high-quality animation for the game development industry. We proposed a lightweight cascaded pipeline named YOLO11-OpenGCN. This pipeline utilizes YOLOv11-Nano Segmentation [19] for human detection, isolating the subject from the background. Subsequently, a lightweight OpenPose model [8] with a MobileNetV3 backbone [18] is used for 2D pose estimation. Finally, a Semantic Graph Convolutional Network (SemGCN) [15] performs 2D-to-3D lifting. We fine-tuned this system on the CHI3D dataset [20] to robustly detect a custom 25-keypoint skeleton across 700,000 frames from multiple camera views.

II. Problem Statement

Despite significant progress in human pose estimation, two major challenges persist for practical applications: computational efficiency and skeletal flexibility. Many state-of-the-art models, while accurate, are computationally too intensive for real-time performance on consumer devices [10]. Furthermore, existing models are often designed for standard academic skeletons (e.g., COCO keypoints) [11], making it difficult to adapt them to custom animation rigs or datasets that utilize a different number of joints.

In this paper, we directly address these issues by proposing an efficient, cascaded pipeline for 3D pose estimation specifically tailored for a custom 25-keypoint skeleton with custom dataset CHI3D. Our core objective is to bridge the gap between high-accuracy methods and real-time applications.

III. Methodology

To formally describe our method, we define the following
key concepts:

- Person Detection(Segmentation): A detected person is represented by a mask or bounding box in the 2D image space. This isolates the region of interest containing the subject[12].
- 2D Human pose(J_{2D}): A 2D human pose is defined as the set of K keypoint coordinates in 2D image space[11] Where x and y is pixel coordinate of $i-th$ joint and K is the total number of joints. In our case $K = 25$
- 3D Human pose(J_{3D}): A 3D human pose is corresponding set of K keypoints in a 3D world or camera coordinate system[13].
- Cascade pipeline $(.)$: our proposed method is a sequential function that map from input image to final 3D poseJ_{3D}. this pipeline comprises three distinct functions[12]

A. Loss Function

1) 2D Pose Estimation

the lightweight OpenPose model is trained using mean squared error (MSE). This loss compares the model predicted heatmap for each joint $(H_k)$to a ground truth heatmap (H_k)[14] generated a gaussian kernel. Formula (1) is to calculate the Mean Square error of 2D keypoint position

$$(1)\ L_{2D} = \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K} ||\ M_{n,k}(\widehat{H}_{n,k} - H_{n,k})||^2$$

Where:

- N is the batch size.
- K is the number of keypoints (25).
- $H_{n,k}$ and $\widehat{H}_{n,k}$ represent the ground-truth and predicted heatmaps, respectively, for the $k-th$ keypoint of the $n-th$ sample
- $M_{n,k} \in \{0,1\}$ is a binary mask used to suppress gradients for missing or occluded joints

2) 3D Pose Estimation

the semantic graph convolutional network model is train using mean per joint position error (MPJPE)[15]. This loss compares the model predicted 3D keypoint position coordinates to ground truth keypoint coordinates. Formula (2) is to calculate the mean of the joint position error of the 3D keypoint position

$$(2)\ MPJPE = \frac{1}{N_F}\frac{1}{N_J}\sum_{f,j} \| P_{f,j} - P^2{}_{f,j} \| 2$$

Where:

- N_f is total frame
- N_j is total joints

- $\| P_{f,j} - P^2{}_{f,j} \| 2$ is the Euclidean distance between the predicted 3D coordinates P and ground-truth P^2coordinates or the J-th joint.

B. Evaluation Metrics

1) 2D Pose Estimation

The lightweight OpenPose model is evaluated by mean Average Precision with Object Keypoint Similarity where formula (3) is for calculate the mean of Average Precision :

$$(3)\ mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k$$

Where :

- n represent total of classes
- $\sum_{k=1}^{k=n} AP_k$ total of average precision
- AP represent total of P(Precision)

And formula (4) is for calculate the Object keypoint similarity for the similarity of 2D Keypoint:

$$(4)\ OKS = \frac{\sum_i \left[\exp\left(-\frac{d_i^2}{2s^2k_i^2}\right).\delta(v_i > 0)\right]}{\sum_i \delta\,(v_i > 0)}$$

Where:

- d_i represent the Euclidean distance between the predicted and ground truth keypoint
- $s_p\ is\ the\ scale\ of\ the\ person\ P.$
- k_i represent $i - th$ keypoint
- $\delta(v_i > 0)$ is the visibility flag, ensuring only visible keypoints contribute to the score

2) 3D Pose Estimation

The semantic graph convolutional network model is evaluated by Mean Per Joint Position Error (MPJPE) where the formula same as formula (2)

3) Cascade pipeline

For the cascade pipeline we use $F(.)$ which composite three distinct function where (5) formula is for calculate the total of speed of the pipeline inference:

$$(5)\ J_{3D} = \mathrm{F(I)} = f_{3D}(f_{2D}\left(f_{seq}(I)\right)$$

Where :

- f_{seq} denotes the human segmentation model (Yolov11-nano)
- f_{2D} denotes the 2D pose estimation model (OpenPose)
- f_{3D} denotes the 2D to 3D lifting model (SemGCN)

IV. Result and Discussion

Our proposed pipeline, YOLO11-OpenGCN, was evaluated on the CHI3D dataset. We measured the performance of the 2D and 3D pose estimation stages separately to analyze the impact of dataset size, data augmentation, and model hyper parameters

1) Video Segmentation: Yolov11nano – Segmentation

Since we only use pretrain model this pipeline only run the inference to produce the input for the lightweight OpenPose

The Yolov11-nano-Segmentation result are shown in Figure 1

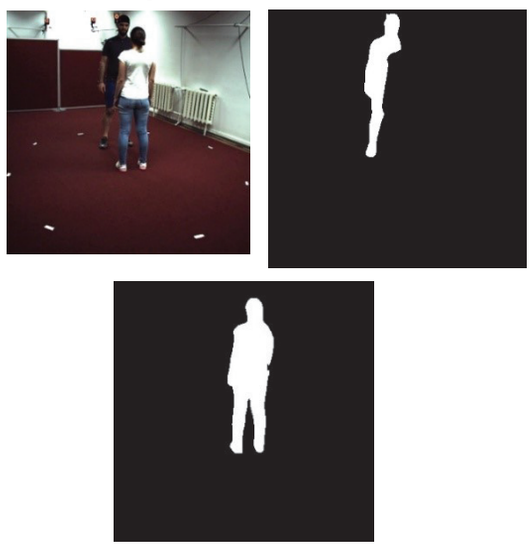


Figure 1: YOLOv11 nano segmentation mask for overlapping person without fine-tuning

2) 2D Pose Estimation: Lightweight OpenPose

We trained a lightweight OpenPose model based on the Daniil Osokin repository [16] to detect our custom 25-keypoint skeleton. We modified the model architecture to support 26 heatmaps and 48 Part Affinity Fields (PAFs) and introduced real-time augmentations such as motion blur and occlusion.

We evaluated performance using the standard Average Precision (AP) metric based on Object Keypoint Similarity (OKS). As shown in Table 1 for the model parameter and Table 2 for the model performance, training on the larger CHI3D dataset (100,000 images) with our custom configuration yielded superior results compared to the baseline. The best model, trained with a learning rate of 0.001 and a batch size of 256, achieved a final AP of 0.843.

Table 1 Parameter of lightweight OpenPose.

Dataset size	Augmentation	Learning Rate Head Only	Total Epoch Head Only	Learning Rate Fine Tune	Batch Size
50,000	N.Aug	0.008	10	0.0008	512
50,000	N.Aug (W.A CJ,BL,OC)	0.008	10	0.0008	512
100,000	N.Aug (W.A CJ,BL,OC)	0.004	10	0.0004	512
100,000	N.Aug (W.A CJ,BL,OC)	0.001	10	0.0001	512
100,000	N.Aug (W.A CJ,BL,OC)	0.001	10	0.0001	256

Table 2 Performance of lightweight OpenPose

Experiment	AP (IoU=0.50:0.95)	AP (IoU=0.50)	AP (IoU=0.75)
1	0.718	0.930	0.779
2	0.659	0.910	0.718
3	0.774	0.940	0.821
4	0.787	0.787	0.832
5	0.802	0.940	0.843

Where :

- (W.A CJ,BL,OC) means With Color Jitter , Blur and Occlusion
- N.Aug means normal augmentation

The lightweight OpenPose result are shown in Figure 2

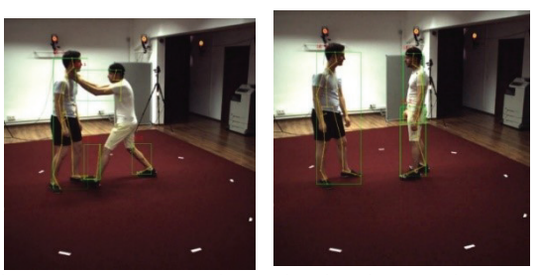


Figure 2: Openpose best checkpoint result on test data

3) 3D Pose Estimation: SemGCN

For the 3D lifting stage, we utilized the SemGCN model implementation by Zhao et al. [17] and modified the graph topology by redefining the adjacency matrix A. Unlike the standard 17-point Human3.6M skeleton, our graph consists of K=25 nodes. We defined the edge connections according to the kinematic chain of the CHI3D dataset, extending the spinal column to include additional spine joints and distinguishing between the clavicle and shoulder joints. The non-local edges (connecting geometrically distant but semantically related joints) were re-initialized to capture relationships specific to this denser topology ,was evaluated using two standard metrics:

MPJPE (Mean Per Joint Position Error): The average Euclidean distance between predicted and ground-truth 3D joints, measured in millimeters.

P-MPJPE (Procrustes-Aligned MPJPE): The MPJPE after aligning the predicted pose with the ground truth via a rigid transformation (rotation, translation, and scaling). This metric focuses purely on the reconstructed pose shape.

As shown in Table 3 for model parameter and Table 4 for the model performance, the best-performing model achieved a P-MPJPE of 73.86 mm. This was accomplished using a learning rate of 0.001, a dropout of 0.2, a batch size of 512, and horizontal flip augmentation. Interestingly and we achieve much more from the baseline model using the CHI3D dataset

Table 3 Parameter of SemGCN

	Total Pose Train	Augmentation	Epoch	Batch Size	Learning Rate
Others[9]	-	-	-	-	-
Ours	183,796	N.Aug	100	256	0.0001
Ours	183,796	Flip Horizontal	100	256	0.001
Ours	183,796	Flip Horizontal	100	256	0.001
Ours	183,796	Flip Horizontal	100	256	0.001
Ours	183,796	Flip Horizontal	100	256	0.001

Table 4 Performance of SemGCN

Experiment	MPJPE	P-MPJPE
1	149.4	-
2	236.5411	136.3220
3	159.3642	73.8689
4	511.9242	224.3092
5	208.4144	182.5542
6	158.5072	79.7532

Where :

- N.Aug means Normal Augmentation

The result of the SemGCN result are shown at Figure 3

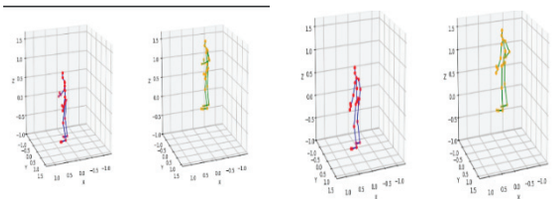


Figure 3 : Validation Result SemGCN

4) Cascade Pipeline

To evaluate the real-time feasibility of the YOLO11-OpenGCN pipeline, we measured the total inference latency by combining the processing times of the three cascaded models.

Table 5 details the individual performance of each module using the best-performing checkpoints. To demonstrate the efficiency of our approach, Table 6 compares our total pipeline speed against the baseline YOLOv5-HR-TCM framework [3]. While the baseline achieves only 3.146 FPS, our proposed YOLO11-OpenGCN pipeline achieves 15.74 FPS, representing a 5-fold improvement in processing speed. This significant speedup validates the suitability of our framework for real-time interactive applications.

Table 4 Performance of the YOLO11-OpenGCN Pipeline

Pipeline Name	Frame Per Second(FPS)
YOLOv11nano-Segmentation	51.12
OpenPose	49.08
SemGCN	42.37

Table 5 Comparison of overall inference speed (FPS)

Pipeline Name	Frame Per Second(FPS)
YOLOv5-HR-TCM[3]	3.146
YOLO11-OpenGCN	15.74

The result of the cascade pipeline are shown in Figure 4

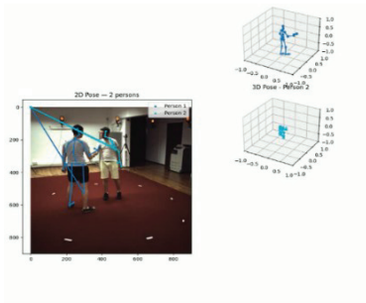


Figure 4 : Result of YOLO11-OpenGCN Module

V. CONCLUSIONS

The YOLO11-OpenGCN pipeline developed in this work was designed to balance accuracy with real-time performance for custom 3D human pose estimation. Our cascaded architecture—comprising YOLOv11-Nano Segmentation, lightweight OpenPose, and SemGCN proved highly efficient. Performance benchmarks confirm its suitability for live applications, with the segmentation, 2D detection, and 3D lifting stages achieving 51.12 FPS, 49.08 FPS, and 42.37 FPS, respectively.

While the model achieves a respectable P-MPJPE of 73.86 mm, the primary trade-off for this high speed is a reduction in robustness compared to heavier models. The system may struggle with severe occlusions or complex poses not well-represented in the training data. In summary, YOLO11-OpenGCN establishes a strong baseline for custom real-time HPE on consumer hardware. Future research will focus on enhancing robustness through temporal modeling and expanded training datasets to further stabilize 3D reconstruction.

ACKNOWLEDGMENT

We acknowledge the use of AI tools for linguistic refinement and grammar correction to enhance the clarity of this manuscript. The original research design, custom 25-keypoint skeleton methodology, and all mathematical evaluations including the 15.74 FPS performance benchmarks were developed well by the authors

References

[1] "Gaming Animation: Uses of Animation in the Gaming Industry." Accessed: Nov. 02, 2025. [Online]. Available: https://www.juegostudio.com/blog/gaming-animation

[2] "The Integration of Motion Capture in Game Animation." Accessed: Nov. 02, 2025. [Online]. Available: https://www.ixiegaming.com/blog/integration-of-motion-capture-in-game-animation/

[3] H. C. Nguyen, T. H. Nguyen, R. Scherer, and V. H. Le, "Unified End-to-End YOLOv5-HR-TCM Framework for Automatic 2D/3D Human Pose Estimation for Real-Time Applications," *Sensors*, vol. 22, no. 14, Jul. 2022, doi: 10.3390/s22145419.

[4] H. Sun, Y. Zhang, Y. Zheng, J. Luo, and Z. Pan, "G2O-Pose: Real-Time Monocular 3D Human Pose Estimation Based on General Graph Optimization," *Sensors*, vol. 22, no. 21, Nov. 2022, doi: 10.3390/s22218335.

[5] W. Ding and W. Li, "High Speed and Accuracy of Animation 3D Pose Recognition Based on an Improved Deep Convolution Neural Network," *Applied Sciences (Switzerland)*, vol. 13, no. 13, Jul. 2023, doi: 10.3390/app13137566.

[6] V. C. Lungu-Stan and I. G. Mocanu, "3D Character Animation and Asset Generation Using Deep Learning," *Applied Sciences (Switzerland)*, vol. 14, no. 16, Aug. 2024, doi: 10.3390/app14167234.

[7] J. Stenum, K. M. Cherry-Allen, C. O. Pyles, R. D. Reetzke, M. F. Vignos, and R. T. Roemmich, "Applications of pose estimation in human health and performance across the lifespan," Nov. 01, 2021, *MDPI*. doi: 10.3390/s21217315.

[8] S. Jia, "Frontiers in Computing and Intelligent Systems 3D Human Pose Estimation: A Survey".

[9] P. Hardy and H. Kim, "Unsupervised Multi-Person 3D Human Pose Estimation From 2D Poses Alone," Mar. 2024, [Online]. Available: http://arxiv.org/abs/2309.14865

[10] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep High-Resolution Representation Learning for Human Pose Estimation." [Online]. Available: https://github.com/leoxiaobin/

[11] T.-Y. Lin *et al.*, "Microsoft COCO: Common Objects in Context," Feb. 2015, [Online]. Available: http://arxiv.org/abs/1405.0312

[12] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN."

[13] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments," 2014. [Online]. Available: http://vision.imar.ro/human3.6m.

[14] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields *." [Online]. Available: https://youtu.be/pW6nZXeWlGM

[15] L. Zhao, X. Peng, Y. Tian, M. Kapadia, and D. N. Metaxas, "Semantic Graph Convolutional Networks for 3D Human Pose Regression," Mar. 2020, doi: 10.1109/CVPR.2019.00354.

[16] D. Osokin, "Real-time 2D Multi-Person Pose Estimation on CPU: Lightweight OpenPose," *International Conference on Pattern Recognition Applications and Methods*, vol. 1, pp. 744–748, 2019, doi: 10.5220/0007555407440748.

[17] "GitHub - garyzhao/SemGCN: The Pytorch implementation for 'Semantic Graph Convolutional Networks for 3D Human Pose Regression' (CVPR 2019)." Accessed: Nov. 06, 2025. [Online]. Available: https://github.com/garyzhao/SemGCN

[18] A. Howard *et al.*, "Searching for MobileNetV3".

[19] L. He, Y. Zhou, L. Liu, and J. Ma, "Research and Application of YOLOv11-Based Object Segmentation in Intelligent Recognition at Construction Sites," *Buildings*, vol. 14, no. 12, Dec. 2024, doi: 10.3390/buildings14123777.

[20] M. Fieraru, M. Zanfir, E. Oneata, A.-I. Popa, V. Olaru, and C. Sminchisescu, "Three-dimensional Reconstruction of Human Interactions." [Online]. Available: http://vision.imar.ro/ci3d.

Hybrid Scale Factor Adjustment for Spatial Resampling of VCM

1st Min-suk Kim *, 2nd Eun-bin An *, 3rd Sang-woon Kwak †, 4th Ayoung Kim †, 5th Soon-heung Jung †, 6th Kwang-deok Seo *
* *Division of Software, Yonsei University, Wonju, Korea*
† *ETRI(Electronics and Telecommunications Research Institute), Daejeon, Korea*
Emails: *minsuk98@yonsei.ac.kr, eunbin.an@yonsei.ac.kr, s.kwak@etri.re.kr, ayoung.kim@etri.re.kr, zeroone@etri.re.kr kdseo@yonsei.ac.kr*

***Abstract*— This paper presents a hybrid scale factor adjustment method for the Spatial Resampling module in the Video Coding for Machines (VCM) framework. Conventional resampling uses the Average Occupied Area Ratio (Average OAR) and input resolution score to determine the scale factor. While this approach maintains compression efficiency, it can fail to preserve small objects when the input resolution is high or when Region of Interest (ROI) retargeting is applied. The proposed hybrid approach adaptively switches between the conventional and object size-based methods depending on the input resolution, ensuring that every object remains above a minimum detectable size. Experimental results on SFU-HW and TVD datasets show improved trade-offs between compression efficiency and machine perception accuracy, achieving up to −3.99% BD-Rate gain and +4.51% BD-mAP improvement.**

Keywords*— *Video Coding for Machines, Spatial Resampling, scale factor adjustment, compression efficiency, machine perception

I. Introduction

The continuous increase in video data has made digital media the dominant form of information exchange. The growing use of mobile devices, high-speed networks, and smart sensors has significantly increased the share of video traffic in total data communication. Videos are now utilized not only for human viewing but also as essential inputs for automated systems that perform visual perception and decision-making[1]. Conventional video codecs have been optimized for human perception, emphasizing visual fidelity and subjective quality[2,3]. Compression methods in existing standards focus on removing spatial and temporal redundancy while maintaining visual similarity. Although this approach provides efficient compression for human viewing, it can degrade the performance of machine-based analysis tasks that rely on fine visual details. When a video is compressed for human consumption, structural information critical for object recognition or tracking is often lost.

To address these limitations, the Moving Picture Experts Group (MPEG), an international organization responsible for developing standards for video and audio coding, established the Video Coding for Machines (VCM) Ad-hoc Group (AHG) and initiated standardization discussions[4,5]. The goal of VCM is to design a new video coding framework optimized for machine perception while minimizing bitrate usage. Among its key modules, the Spatial Resampling process reduces spatial resolution while maintaining task-relevant information.

II. Limitations of Conventional Spatial Resampling

The Spatial Resampling module in VCM reduces image resolution to improve compression efficiency while maintaining inference performance for machine vision tasks[6]. Before resampling is applied, the system executes a pre-analysis stage that consists of three sequential steps as shown in Fig. 1. In valid object filtering, only detections that satisfy the confidence threshold are retained as valid objects. In initial scale decision, the system computes an initial scale decision using the frame-level object occupancy distribution (OOD), which reflects how detected objects occupy the image area. In scale adjustment, the system determines the final scale factor for each frame using the Average Occupied Area Ratio and the input resolution score. After this stage, the resampled frames are encoded by the Inner codec using the MI-RPR (Machine Inference-based Reference Picture Resampling) process to preserve coding efficiency. However, the scale factor adjustment still has several fundamental limitations.

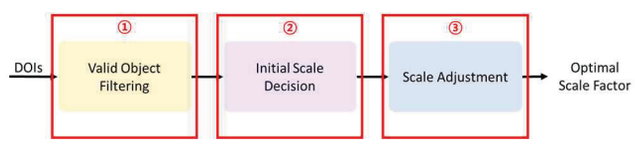


Fig. 1. Pre-analysis stages in the Spatial Resampling module

First, averaging object areas across the entire frame causes the influence of small objects to be diluted. When large and small objects coexist, the computed average primarily reflects larger regions, causing small objects to shrink excessively and fall below the detection threshold of inference models. This results in lower recognition accuracy in scenarios that include pedestrians, traffic signs, or other small but critical objects. Second, high-resolution content such as 2K or 4K inputs requires additional downscaling to match the input size constraints of typical object detection models. In this process, detailed features can disappear due to over-compression, leading to unstable performance and increased sensitivity to quantization noise. Third, the scale factor is determined before ROI Retargeting[7] is performed. Consequently, any subsequent resizing or movement of regions of interest is not reflected in the final spatial resampling scale factor decision. This causes inconsistencies between the scale factor applied during spatial resampling and the actual spatial characteristics after ROI Retargeting.

These limitations indicate that the conventional scale factor adjustment cannot simultaneously achieve optimal compression efficiency and inference accuracy. To address this, it is necessary to optimize the scale adjustment operation according to both resolution and object-level information. Fig. 2. shows the overall VCM encoding process, including proposed hybrid scale factor adjustment is applied.

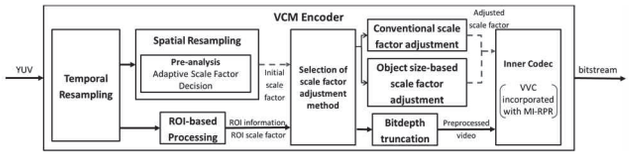


Fig. 2. VCM encoding process including the proposed hybrid scale factor adjustment

III. Proposed Hybrid Scale Factor Adjustment

The proposed method introduces a resolution-dependent mechanism that optimizes the selection between the conventional Average OAR-based adjustment and an object size-based correction method according to input characteristics. When the frame width is smaller than a defined resolution threshold $W_0 = 1280$, the conventional scale factor adjustment method is applied. In this range, the conventional approach already incorporates the input resolution score, which effectively constrains excessive down-scaling at low resolutions. As a result, the scale factor remains within a stable range without additional scale factor adjustment. Fig. 3. illustrates the resolution-based optimization mechanism that determines whether to apply conventional or object-based correction depending on the input resolution.

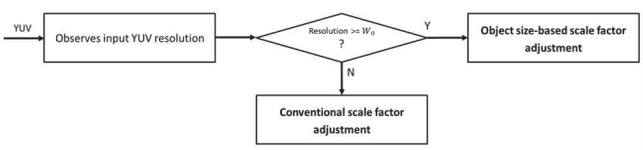


Fig. 3. Resolution-based selection mechanism for the hybrid scale factor adjustment

When the frame width is equal to or greater than W_0, the object size-based scale factor adjustment is activated. This mode guarantees that the size of every detected object remains larger than a minimum detectable threshold τ. For each object j, let a_j be the shorter side length of the bounding box, r_j the ROI scaling ratio, and s the spatial resampling scale factor. The size of the object after applying both ROI retargeting and Spatial Resampling is defined as (1):

$$a'_j = a_j \times r_j \times s \tag{1}$$

To ensure detectability, the condition $a'_j \geq \tau$ must be satisfied, leading to (2):

$$s \geq \frac{\tau}{a_j \times r_j} \tag{2}$$

For all detected objects, the adjusted scale factor is determined as (3):

$$s_adjusted = \max\left(\frac{\tau}{a_j \times r_j}\right) \tag{3}$$

The final applied spatial resampling scale factor is (4):

$$s_final = \max(s_initial, s_adjusted) \tag{4}$$

To maintain codec compatibility, s_final is quantized to the nearest upper value in a predefined discrete set {0.50, 0.75, 0.90}. The quantized scale factor is applied individually to each frame according to its computed value. This discrete mapping allows integration into the existing index-based scale factor structure of the VCM-RS (Video Coding for machine Reference Software) framework[8]. The quantized s_final value is then used by the subsequent MI-RPR process within the Inner Codec. Fig. 4. presents a flow chart of the object size-based scale factor adjustment.

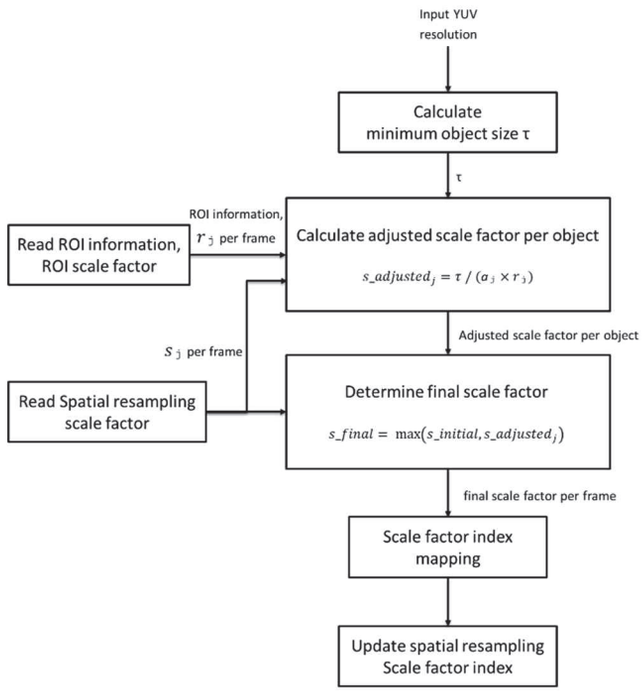


Fig. 4. Flow chart of the object size-based scale factor adjustment

IV. Experimental Results

The proposed method was implemented in the VCM-RS reference software and evaluated under the standard VCM common test conditions (VCM-CTC) using two datasets[9]: SFU-HW[10] for object detection and TVD[12] for object tracking. Both inner codec and end-to-end configurations were tested. In the SFU-HW experiments, the optimized scale adjustment method achieved consistent improvements. The BD-Rate decreased by −3.99%, and BD-mAP increased by +0.72 on average. For Class A (2560×1600) and Class B (1920×1080), the improvements were more pronounced. In the LD_e2e configuration, Class A achieved −16.93% BD-Rate and +4.51 BD-mAP improvement. Table 1. presents the detailed results for the SFU dataset, comparing the conventional and proposed methods under identical configurations. These results confirm that the hybrid adjustment effectively

TABLE I. Test results of the proposed method for the SFU dataset

	Object Detection	End-to-End BD-Rate [%]	End-to-End BD-mAP
RA_e2e	Class A	7.35%	-0.84
	Class B	-5.05%	0.63
	Class C	0.00%	0.00
	Class D	0.00%	0.00
	Average	**-0.65%**	**0.09**
LD_e2e	Class A	-16.93%	4.51
	Class B	-10.30%	1.39
	Class C	0.00%	0.00
	Class D	0.00%	0.00
	Average	**-3.99%**	**0.72**

AI_e2e	Class A	-5.91%	0.52
	Class B	-1.94%	0.06
	Class C	0.00%	0.00
	Class D	0.00%	0.00
	Average	**-0.98%**	**0.06**

For the TVD dataset, the proposed optimization also achieved better performance. The RA_e2e configuration achieved −6.08% BD-Rate and +0.89 BD-MOTA improvement, while the AI_e2e configuration achieved −7.37% BD-Rate and +34.05 BD-MOTA improvement. Through this experiment, it was confirmed that the proposed hybrid scale factor adjustment also works effectively in object tracking tasks.

TABLE II. TEST RESULTS OF THE PROPOSED METHOD FOR THE TVD DATASET

	Object Tracking	**End-to-End BD-Rate [%]**	**End-to-End BD-MOTA**
RA_e2e	TVD-01-1	-7.00%	1.04
	TVD-01-2	-11.65%	1.47
	TVD-01-3	-6.22%	0.71
	TVD-02-1	0.04%	0.00
	TVD-03-1	1.97%	-0.04
	TVD-03-2	-7.77%	1.82
	TVD-03-3	-11.94%	1.22
	Average	**-6.08%**	**0.89**
LD_e2e	TVD-01-1	17.05%	-2.76
	TVD-01-2	-3.42%	0.95
	TVD-01-3	2.65%	-0.37
	TVD-02-1	20.47%	-1.71
	TVD-03-1	-7.36%	1.61
	TVD-03-2	-4.59%	1.60
	TVD-03-3	-15.06%	2.43
	Average	**1.39%**	**0.25**
AI_e2e	TVD-01-1	-9.17%	0.24
	TVD-01-2	-32.20%	0.56
	TVD-01-3	-4.28%	0.26
	TVD-02-1	7.40%	-0.37
	TVD-03-1	-7.56%	0.64
	TVD-03-2	-11.71%	1.24
	TVD-03-3	5.96%	-0.19
	Average	**-7.37%**	**0.34**

V. CONCLUSION

This paper presented a scale factor optimization method for Spatial Resampling in Video Coding for Machines. The method determines the scale factor according to both resolution and object-level characteristics, preventing excessive down-scaling of small objects while maintaining overall compression efficiency. Through experiments conducted on the SFU-HW and TVD datasets, the proposed method consistently improved both coding efficiency and perception accuracy compared to the conventional Average OAR-based approach. In particular, it achieved noticeable gains in high-resolution sequences where small objects are more susceptible to loss after Spatial Resampling.

The proposed hybrid structure is fully compatible with the existing VCM framework and requires no modification to the syntax or bitstream structure of the reference software. Its modular design enables straightforward integration with other functional components, including ROI retargeting and MI-RPR-based inner coding. The proposed method has been adopted in the VCM standard and is currently integrated into the official VCM pipeline, operating after the Spatial Resampling and ROI retargeting stages.

ACKNOWLEDGMENT

This work was supported by Institute of Information & communications. Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT)(No. 2020-0-00011, Video Coding for Machine).

REFERENCES

[1] J. Zhang and D. Tao, “Empowering things with intelligence: A survey of the progress, challenges, and opportunities in artificial intelligence of things,” IEEE Internet of Things Journal, vol. 8, no. 10, pp. 7789–7807, May 2021.

[2] Information technology — High efficiency coding and media delivery in heterogeneous environments — Part 2: High Efficiency Video Coding, ISO/IEC 23008-2:2023, 2023.

[3] Information technology — Coded representation of immersive media — Part 3: Versatile Video Coding, ISO/IEC 23090-3:2023, 2023.

[4] ISO/IEC JTC 1/SC 29/WG 2, ″Call for Proposals on Video Coding for Machines,″ISO/IEC, Output Document N00246, October 2022.

[5] ISO/IEC JTC 1/SC 29/WG 2, ″Call for Evidence on Video Coding for Machines,″ISO/IEC, Output Document N00215, July 2022.

[6] E. B. An, A. Y. Kim, S. H. Jung, S. W. Kwak, J. Lee, W. S. Cheong, H. G. Choo, and K. D. Seo, (2024). “Adaptive spatial down-sampling method based on object occupancy distribution for video coding for machines,” EURASIP Journal on Image and Video Processing, 36, 2024. Available: https://doi.org/10.1186/s13640-024-00647-y

[7] O. Stankiewicz, T. Grajek, S. Mackowiak, J. Stankowski, S. Rozek, M. Lorkiewicz, M. Wawrzyniak, and M. Domanski, “Region-of-Interest-Based Video Coding for Machines,” in Proceedings of the IEEE International Conference on Multimedia and Expo Workshops (ICMEW), Los Alamitos, CA, USA: IEEE Computer Society, July 2024, pp. 1–6, doi: 10.1109/ICMEW63481.2024.10645441.

[8] ISO/IEC JTC 1/SC 29/WG 4, “Algorithm description of tools in vcm reference software,” ISO/IEC, Output Document N00700, July 2024.

[9] ISO/IEC JTC 1/SC 29/WG 4, “Common test conditions for video coding for machines,”ISO/IEC, Output Document N00699, July 2025.

[10] H. Choi, E. Hosseini, S. Ranjbar Alvar, R. A. Cohen, I. V. Bajić, A. Karabutov, Z. Yin, and E. Alshina, “A dataset of labelled objects on raw video sequences,” Data in Brief, vol. 34, Art. 106701, 2021. doi: 10.1016/j.dib.2020.106701. [Online]. Available: https://www.researchgate.net/publication/348000819

[11] X. Xu, S. Liu, and Z. Li, “Tencent Video Dataset (TVD): A video dataset for learning-based visual data compression and analysis,” arXiv preprint arXiv:2105.05961, May 2021. [Online]. Available: https://arxiv.org/abs/2105.05961

Daily Municipal Solid Waste Forecasting: A Comparison of SARIMA and LSTM

1st Syalom Veninda Runtuwene
Department of Electrical and Information Engineering
Universitas Gadjah Mada
Yogyakarta, Indonesia
syalomvenindaruntuwene@mail.ugm.ac.id

2nd Rudy Hartanto
Department of Electrical and Information Engineering
Universitas Gadjah Mada
Yogyakarta, Indonesia
rudy@ugm.ac.id

3rd Adhistya Erna Permanasari
Department of Electrical and Information Engineering
Universitas Gadjah Mada
Yogyakarta, Indonesia
adhistya@ugm.ac.id

***Abstract*— Daily fluctuations in municipal solid waste (MSW) generation complicate operational planning for waste management, particularly in developing cities with growing economic activity and increasing population density such as Moratuwa. This study compares the forecasting performance of a seasonal statistical model (SARIMA) and a deep learning model (LSTM) for daily total MSW in Moratuwa using data from 2015–2018. Evaluation is conducted using a rolling/walk-forward 1-step ahead scheme on the 2018 test period to reflect daily operational use. Beyond point forecasts, the study assesses predictive uncertainty through 95% prediction intervals (SARIMA based on model output and LSTM via Monte Carlo Dropout) and examines LSTM robustness through repeated runs with multiple random seeds and rolling-origin evaluation across several cutoff points. The results show that both approaches achieve comparable performance, with LSTM slightly outperforming SARIMA in MAE, RMSE, and MAPE on the test period (MAPE of 22.88% for LSTM and 23.58% for SARIMA). The robustness evaluation indicates that LSTM exhibits smaller performance variability, suggesting better consistency against training randomness. However, both models still struggle to capture extreme spikes on certain days, highlighting the high volatility of the daily MSW series. Further improvements are expected by incorporating explanatory variables to better model extreme events.**

Keywords—Municipal Solid Waste, SARIMA, LSTM, Robustness

I. INTRODUCTION

Daily human activities, both directly and indirectly, always produce waste. In this study, municipal solid waste (MSW) refers to the overall municipal waste stream, rather than a specific waste category. Globally, MSW generation is estimated to reach around 2.01 billion tons per year, and at least 33% of it is still not managed in an environmentally safe manner [1]. This condition makes poor waste management and unsafe disposal practices a real threat to the environment and human health. The challenges are not isolated: rapid urbanization, population growth, and the impacts of climate change are further increasing the complexity and dynamics of MSW management [2]. Furthermore, this increasing trend in generation is projected to continue until 2050 and shows no signs of slowing, so that pressure on waste management systems and environmental-health risks will increase [1].

Poor waste management is often not just about littering, but also about failing to predict waste generation, resulting in poor planning: disposal facilities/capacity may be unnecessary or even insufficient, collection services may be suboptimal, and landfill and recycling management may become skewed. Therefore, accurate forecasting is needed so authorities can take early action, improve existing systems, and minimize environmental impacts and health risks when waste is not handled properly [3]. On the operational side, this challenge is even more severe because logistics activities have many constraints (fluctuating demand, limited vehicle capacity/number, time constraints, and numerous and dispersed service points), while transportation costs are often the most dominant component in the logistics cost structure; this means that without proper planning (including route optimization), costs and service quality will easily leak everywhere [4].

SARIMA (Seasonal Autoregressive Integrated Moving Average) is a statistical method that excels at capturing seasonal patterns and trends in time series data. SARIMA's advantage lies in its ability to simultaneously accommodate seasonal and non-seasonal components, making it highly relevant for waste data, which typically exhibit seasonal fluctuations, and is suitable for operational applications and medium-term planning [5], [6].

LSTM (Long Short-Term Memory) is relevant for univariate waste generation forecasting because its gating mechanism allows the model to learn long-term temporal dependencies and mitigates the vanishing gradient problem of traditional RNNs [7]. Studies in the waste management domain also report that LSTM outperforms ARIMA when the time series exhibits strong temporal dependencies, non-linear patterns, and seasonal variations [8].

Although seasonal statistical models such as SARIMA and deep learning models like LSTM have both been used for waste generation forecasting, the need for focused empirical evaluation remains strong, especially in the context of developing cities. MSW recording and data systems often face capacity limitations, incompleteness, and inconsistent definitions and measurement methods, resulting in only approximate data.

On the method side, LSTM is designed to model sequential data and address the vanishing gradient problem through a gating mechanism (input–forget–output), potentially capturing long-term dependencies in time series [7]. In waste management applications, LSTM has been reported to be capable of capturing complex temporal patterns, but its performance is sensitive to configuration and risks

overfitting. Therefore, regularization practices such as dropout and early stopping are often recommended; in fact, LSTM may be less superior on very small datasets or when the data does not exhibit sufficient temporal dependencies [8].

Based on this gap, this study aims to compare the performance of SARIMA and LSTM on univariate MSW time series of Moratuwa City to empirically assess the balance between prediction accuracy and model stability on fluctuating city operational data.

II. METHODS

A. Study Area and Data

This study uses the Daily Solid Waste Dataset published on Kaggle [9]. The dataset contains daily waste volume data. Specifically, the Moratuwa Urban Council dataset contains 1376 observations of daily MSW volume (in kg) with varying values reported in reference studies, and has missing values associated with irregularities in collection and reporting by the transport authorities [10]. In this study, only two attributes are used, namely the date and the daily MSW volume, while handling missing/outliers is explained in the preprocessing subsection.

B. Data Preprocessing

Table I summarizes the raw Moratuwa waste dataset collected from 12 January 2014 to 11 December 2018, showing the number of records available for each waste type. The dataset is dominated by major categories such as Municipal Solid Waste (1,307 records), Sorted Organic Waste (1,208 records), Slaughter House Waste (1,105 records), and Bulky Waste (1,022 records), while several minor categories contain very limited observations (e.g., C&D Waste: 102, Industrial Sludge Waste: 41, and multiple categories with fewer than 20 records).

TABLE I. SUMMARY OF MORATUWA DATASET BY WASTE TYPE (2014–2018)

Waste Type	Row Count
Municipal Solid Waste	1307
Sorted Organic Waste	1208
Slaghter House Waste	1105
Bulky Waste	1022
Industrial Waste	842
Soil With Waste	777
Sanitary Waste	236
C&D Waste	102
Indutrial Sludge Waste	41
Soil	16
Saw Dust	14
Polythyne & Regiform	11
Mesuring	5
Wood Debris	3
Special Waste	1
Wood Trank	1

After the initial screening of the raw Moratuwa dataset (Table I), the time series was reindexed to a daily frequency to enforce a continuous calendar day index. Because the original records are irregular and some dates are not reported, daily reindexing introduced additional missing entries corresponding to those unobserved days. A subsequent data quality check indicated that 2014 contained an excessively high proportion of missing observations, which could compromise model estimation and bias the evaluation. Therefore, the study period was restricted to 2015–2018, resulting in a complete daily index of 1441 days.

To improve the usability of the dataset for forecasting, waste categories with very limited observations were aggregated into a single “Others” group. Specifically, Soil With Waste, Sanitary Waste, C&D Waste, Industrial Sludge Waste, Soil, Saw Dust, Polythyne & Regiform, Measuring, Wood Debris, Special Waste, and Wood Trunk were merged due to their sparse counts in the raw data. Table II summarizes the post-processing dataset where each waste type shares the same daily row count (1441) after reindexing, and reports the total number of missing values per category. Importantly, the missing-value counts in Table II reflect the accumulated missingness from three sources: (1) missing entries already present in the original dataset, (2) missing dates introduced by enforcing daily continuity through reindexing, and (3) additional missing values created during outlier treatment, where anomalous observations were flagged and set to missing.

TABLE II. MORATUWA DAILY REINDEXED DATASET AND MISSING-VALUE SUMMARY BY WASTE TYPE (2015–2018)

Waste Type	Row Count	Missing Value
Municipal Solid Waste	1441	76
Sorted Organic Waste	1441	178
Slaghter House Waste	1441	295
Bulky Waste	1441	397
Industrial Waste	1441	561
Others	1441	510

All missing values are imputed before the modeling stage. This imputation procedure is applied consistently to all waste types: Municipal Solid Waste, Sorted Organic Waste, Slaughterhouse Waste, Bulky Waste, Industrial Waste, and Others. Imputation is performed using a supervised learning approach based on the XGBoost Regressor. For each waste type, the data is first formed into predictor features consisting of: (1) calendar features, namely day-of-week (dow) and month, and (2) lag features of the waste value itself, namely lag-1, lag-7, lag-14, and lag-28. These lags are selected to capture short-term dependencies as well as weekly patterns that commonly appear in daily data.

A mask is then created to ensure that only rows with complete features are included in the training and prediction processes. The rows that meet the complete set of features are then divided into two groups: (i) rows with non-NaN garbage values as training data (known target), and (ii) rows with NaN garbage values as data to be predicted (missing target). The XGBoost model is trained using the training data and then used to predict values in the missing rows.

In implementation, the imputation model uses the configuration as shown in Table III. The imputation model uses an XGBoost Regressor with n_estimators=400, learning_rate=0.05, max_depth=4, subsample=0.8, colsample_bytree=0.8, random_state=42, and n_jobs=4.

TABLE III. XGBOOST HYPERPARAMETERS FOR IMPUTATION

Waste Type	Row Count
n_estimators	400
learning_rate	0.05
max_depth	4
subsample	0.8
colsample_bytree	0.8
random_state	42
n_jobs	4

C. Seasonal Autoregressive Integrated Moving Average (SARIMA) Model

The SARIMA model extends ARIMA to explicitly handle seasonal patterns and the possibility of seasonal unit roots [11], [12]. For a time series with seasonal period s, SARIMA is commonly written in multiplicative form as ARIMA$(p, D, q) \times (P, D_s, Q)_s$, and can be expressed as:

$$\Phi(L^s)\phi(L)(1-L)^d(1-L^s)^{D_s}y_t = c + \Theta(L^s)\theta(L)\varepsilon_t \quad (1)$$

where $D_s$denotes the seasonal integration order, indicating the extent of seasonal differencing applied. The operator $(1-L^s)^{D_s}$ performs seasonal differencing and accounts for non-stationarity in observations that recur at the same seasonal positions across successive cycles. Here, $\Phi(L)$ is a stable seasonal autoregressive (AR) polynomial of order P, analogous in form to $\phi(L)$for the non-seasonal component. Likewise, $\Theta(L)$ is an invertible seasonal moving-average (MA) polynomial of order Q, analogous to $\theta(L)$in the non-seasonal part.

D. Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) is an artificial neural network architecture that improves upon the Recurrent Neural Network (RNN). LSTM is equipped with memory units and gating mechanisms, such as the input gate, forget gate, and output gate, which regulate the flow of information so that important information can be retained over long periods while irrelevant information can be forgotten. This approach enables LSTM to capture temporal dependencies and time patterns effectively, making it widely used in applications such as time-series forecasting, language translation, sentiment analysis, and the prediction of economic and health conditions. The following are the stages used in this research [13], [14].

1. **Data split (Train/Validation/Test):** The test set was defined from 1 January 2018 to the end of the observation period, while data prior to this date were used for training. To monitor generalization during training, the last 20% of the training data were separated as the validation set.
2. **Normalization:** Prior to training, the data were normalized to the 0–1 range using MinMaxScaler.
3. **Supervised data construction (Lookback Window):** The normalized time series was converted into a supervised learning format using a sliding-window approach.
4. **Model architecture and training configuration:** The LSTM architecture and training settings were specified.
5. **Rolling 1-step ahead evaluation strategy:** Evaluation was conducted using a rolling (walk-forward) 1-step ahead scheme on the test period.
6. **Evaluation metrics and predictive uncertainty:** Forecast accuracy on the test set was measured using MAE, RMSE, and MAPE.
7. **Robustness testing**.

III. RESULT

A. SARIMA Results

1) Experimental Setup and Data Split: SARIMA MSW generation series in Moratuwa, which was reindexed to a daily frequency and had missing values handled to form a continuous series. The observation period spans 1 January 2015 to 11 December 2018, totaling 1,441 observations. Out-of-sample evaluation used a time-based split with a cutoff on 1 January 2018, yielding a training set covering 2015–2017 (n = 1,096) and a test set for 2018 (n = 345).

2) Stationarity Assessment and Model Identification: A weekly seasonal pattern was assumed with a period of 7 days. Stationarity tests were conducted under several minimal differencing candidates, namely the combinations of d and D: (0,0), (1,0), (0,1), and (1,1). The ADF and KPSS results indicated that several differencing candidates could produce a stationary series, whereas the original data showed inconsistent indications between ADF and KPSS (ADF suggested the series tended to be stationary, while KPSS suggested non-stationarity). Given the strong evidence of a weekly pattern, seasonal differencing with D = 1 and s = 7 was retained to support identification of the model's seasonal components. In Figure 1, the ACF of the seasonally differenced series (D = 1, s = 7) shows a significant spike at lag 7, indicating remaining weekly dependence and supporting the inclusion of seasonal terms in the model.

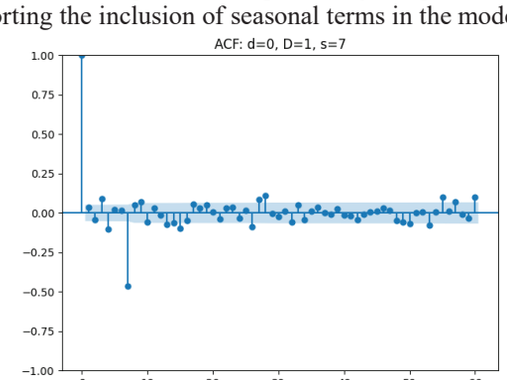


Fig. 1. Autocorrelation function (ACF) of the seasonally differenced MSW series (d = 0, D = 1, s = 7)

3) AIC-based Selection of SARIMA Structure: Next, a restricted grid search was conducted for the parameters (p, q, P, Q), with p and q in {0, 1, 2} and P and Q in {0, 1}, under the seasonal setting s = 7 and seasonal differencing D = 1.

Model selection was based on the minimum AIC, and the best configuration was obtained for SARIMA(0,0,0)(0,1,1,7). The grid search results are presented in Table IV.

TABLE IV. AIC RANKING OF CANDIDATE SARIMA MODELS

p	q	P	Q	AIC
0	0	0	1	33177.902004
0	0	1	1	33178.017547
2	2	0	1	33189.517395
2	2	1	1	33189.945405
1	2	0	1	33259.012759
1	2	1	1	33259.441094
0	2	0	1	33265.028856
0	2	1	1	33265.654437
2	1	1	1	33285.831634
1	1	0	1	33290.935375

4) Prediction Performance in the Test Period with 1-Step Ahead Rolling: Forecast performance was evaluated using a rolling (walk-forward) 1-step ahead scheme on the test period, where each day the model predicts one step ahead and then updates the model state using the actual value observed for that day [15], [16]. This scheme reflects daily operational use and avoids the overly optimistic bias that can arise from multi-step forecasting without information updates. On the test set, the model achieved an MAE of 18,350.42, an RMSE of 24,464.65, and a MAPE of 23.58%. The actual versus predicted plot in Fig. 3 shows that the model tracks the average level and gradual changes, but tends to be less responsive to extreme daily spikes, which contributes to the relatively high MAPE.

5) Prediction Uncertainty via 95% Prediction Interval: Beyond point forecasts, predictive uncertainty was analyzed using 95% prediction intervals derived from the model's forecast output at each step of the rolling 1-step evaluation. The visualization in Fig. 3 shows that the interval bands capture the general variation, but some extreme observations can fall outside the intervals, consistent with the high day-to-day volatility in the test data.

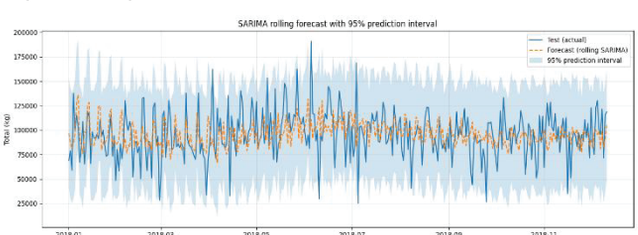


Fig. 2. Observed vs Predicted using SARIMA During the Test Period with 95% prediction interval

6) Robustness across Different Cutoff Dates (Rolling-Origin Evaluation: Robustness was assessed using a rolling-origin evaluation with multiple cutoff points and a fixed test horizon (e.g., 90 days). The results show that SARIMA's performance varies across cutoffs, but remains within a relatively consistent range. The robustness summary is reported in Table V.

TABLE V. ROBUSTNESS SUMMARY OF SARIMA

	MAE	RMSE	MAPE
mean	19692.830652	26615.402874	24.619068
std	1292.739025	2633.969913	3.072034

B. LSTM Results

1) Data Splitting and Preprocessing for LSTM: The data were organized at a daily frequency for the period from 1 Jan 2015 to 11 Dec 2018, totaling 1,441 observations, with no missing values after time interpolation and forward/backward filling (NaN = 0). The dataset was then split chronologically into train–validation–test to keep the evaluation fair for future data. The test period was defined as 1 Jan 2018 to 11 Dec 2018 (345 days). Data before 2018 were used as the full training set (train_full), and the last 20% of this portion was used as the validation set: train = 877 days (1 Jan 2015–26 May 2017) and validation = 219 days (27 May 2017–31 Dec 2017). To stabilize neural network training, the target values were normalized to the [0, 1] range using MinMaxScaler.

2) Supervised Data Construction and Lookback Window: LSTM requires sequence-shaped inputs. Therefore, the normalized time series was transformed into supervised pairs (window to next value) using a lookback window of 56 days. This means the model uses the previous 56 days to predict the next day's value (t+1). With this setup, the resulting data shapes were X_train (821, 56, 1) and X_val (163, 56, 1), with corresponding targets y_train (821,) and y_val (163,) respectively.

3) LSTM Model Architecture: The model uses a compact univariate architecture (single feature), consisting of an input layer with 56 timesteps and 1 feature, one LSTM layer with 64 units, Dropout with a rate of 0.2 for regularization, a Dense layer with 32 units and ReLU activation, and a final Dense layer with 1 unit as the prediction output. The total number of trainable parameters is 19,009. The model was compiled using the Adam optimizer and mean squared error (MSE) as the loss function.

4) Model Training and Early Stopping: The model was trained for up to 200 epochs with a batch size of 32, using EarlyStopping (monitor = val_loss, patience = 10, restore_best_weights = True). In this run, training stopped at epoch 18, indicating that validation performance no longer improved meaningfully after several epochs (suggesting convergence and helping prevent overfitting).

5) Forecast Performance on the Test Period with Rolling 1-Step Ahead: Evaluation was conducted using a rolling (walk-forward) 1-step ahead scheme on the test period. For each day in the test set, the model predicted the next-day value (t+1) based on the previous 56 days, and the

input “history” was then updated using the actual observed value for that day to generate the next prediction (Fig. 3). The test-set metrics were MAE 18,040.89, RMSE 23,921.27, and MAPE 22.88%.

6) Predictive Uncertainty: 95% Prediction Interval via Monte Carlo Dropout: In addition to point forecasts, predictive uncertainty was estimated using Monte Carlo (MC) Dropout. At each rolling prediction step, dropout was kept active during inference (training=True), and the model generated 200 predictive samples for the same day. The 95% prediction interval was constructed from the 2.5th and 97.5th percentiles, while the prediction line was reported as the mean of the sampled forecasts. The visualization shows interval bands that track the general variation, although extreme observations may still fall outside the intervals at certain times.

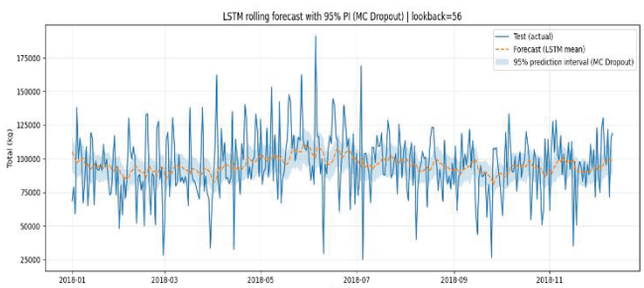


Fig. 3. Observed vs Predicted using LSTM During the Test Period with 95% prediction interval

7) LSTM Robustness: The model was retrained 10 times using different random seeds and then evaluated using the same rolling scheme. The summary (mean ± standard deviation) is reported in Table VI. The relatively small standard deviation indicates that LSTM performance is fairly consistent with respect to seed variation under this configuration.

TABLE VI. ROBUSTNESS SUMMARY OF LSTM

	MAE	RMSE	MAPE
mean	18174.699121	24006.957502	23.026361
std	300.892121	318.327871	0.625960

C. Comparison Between SARIMA and LSTM

Table VII summarizes the comparison of the average (mean) evaluation metrics from the robustness procedure for both models, where LSTM achieves lower mean values across all metrics compared to SARIMA.

TABLE VII. MEAN PERFORMANCE COMPARISON (ROBUSTNESS EVALUATION)

Metric	SARIMA (mean)	LSTM (mean)
MAE	19692.83065	18174.69912
RMSE	26615.40287	24006.95750
MAPE	24.61906	23.02636

Table VIII reports the standard deviation (std) of the evaluation metrics from the robustness evaluation. LSTM shows smaller performance variability than SARIMA across all metrics.

TABLE VIII. STANDARD DEVIATION OF PERFORMANCE METRICS (ROBUSTNESS EVALUATION)

Metric	SARIMA (std)	LSTM (std)
MAE	1292.73902	300.89212
RMSE	2633.96991	318.32787
MAPE	3.07203	0.62596

As a point-forecast benchmark on the test period under the rolling 1-step ahead scheme, Table IX reports error values that are also broadly consistent with the robustness summary.

TABLE IX. POINT-FORECAST PERFORMANCE COMPARISON ON THE TEST SET (ROLLING 1-STEP AHEAD)

Metric	SARIMA	LSTM
MAE	18350.42	18,040.89
RMSE	24464.65	23,921.27
MAPE	23.58	22.88

IV. DISCUSSION

Based on the rolling/walk-forward one-step-ahead evaluation on the test period (Table IX), the performance of LSTM and SARIMA appears comparable, with a slight advantage for LSTM. LSTM yields lower errors in terms of MAE, RMSE, and MAPE than SARIMA, but the differences are not large. This pattern suggests that for Moratuwa’s daily total MSW time series, the short-term historical information used by both models (via state updating in SARIMA and the input window/lookback in LSTM) is sufficient to predict the average daily level, but not strong enough to consistently capture extreme spikes.

A more noticeable difference emerges when considering stability/robustness. In Table VII (mean) and especially Table VIII (standard deviation), LSTM shows smaller standard deviations than SARIMA. This indicates that LSTM’s performance is more consistent under the tested sources of variation (e.g., different initializations/seeds in repeated runs), whereas SARIMA tends to exhibit noisier performance across experiments. From a practical standpoint, this matters: two models may have similar average error, but the model with lower variability is typically more reliable for operational use because its performance is less dependent on specific conditions.

From a methodological perspective, the close performance between SARIMA and LSTM indicates that higher model complexity does not automatically translate into substantially better accuracy when the input information is still limited (univariate) and the series is dominated by irregular day-to-day fluctuations. The simpler SARIMA model can remain competitive because weekly seasonality and the state-updating mechanism capture the most predictable components of the data. LSTM provides a marginal improvement by learning mild nonlinearities and pattern interactions within the input window, but it is still

constrained when facing extreme spikes that are inherently difficult to predict without additional explanatory variables.

V. Conclusion

This study compares the performance of SARIMA and LSTM for forecasting daily total MSW in Moratuwa using a rolling/walk-forward one-step-ahead evaluation scheme that reflects day-to-day operational use. Overall, both approaches achieve comparable performance, with LSTM showing a slight advantage over SARIMA in terms of MAE, RMSE, and MAPE. However, the accuracy gap is relatively small, indicating that the improvement gained from a deep learning model does not constitute a substantial leap when the input remains univariate.

In terms of stability, LSTM exhibits lower variability in performance (smaller standard deviations of the metrics) than SARIMA, suggesting more consistent results with respect to randomness in training. Nevertheless, both SARIMA and LSTM remain limited in capturing extreme spikes on certain days. This highlights that Moratuwa's daily MSW series is highly volatile and includes anomalous events that are difficult to predict using only the historical total MSW values.

These findings are relevant for cities with characteristics similar to Moratuwa, such as developing or peri-urban areas experiencing increasing economic activity and population density, where waste management operations and record-keeping can show sharp day-to-day fluctuations. In such contexts, seasonal statistical models like SARIMA can serve as an efficient and strong baseline, while LSTM may be preferable when the primary objective is performance consistency and when sufficient computational resources and data are available.

As a direction for future work, performance improvements are expected when the models incorporate exogenous variables, so that extreme spikes have learnable drivers that can be leveraged during forecasting.

Acknowledgment

The author would like to thank the Moratuwa municipal waste management authority for making the daily MSW dataset openly available, as well as the supervisors and reviewers for their constructive feedback.

References

[1] H. S. Salem, Y. Yihdego, and M. Y. Pudza, ‘‘Scientific and technical per spectives of waste-to-energy conversion,’’ J. Nature Sci. Sustain. Technol., vol. 15, no. 3, pp. 193–217, 2021.

[2] K. Raab, G. Tolotti, and R. Wagner, ‘‘Challenges in solid waste man agement: Insights into the disposal behavior of suburban consumers in Guatemala city,’’ Frontiers Sustain. Cities, vol. 3, 2021.

[3] Rashmi G and S. Sathish Kumar K, “Cloud based Forecast of Municipal Solid Waste Growth using AutoRegressive Integrated Moving Average Model: A Case Study for Bengaluru,” *(IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 13, no. 9, pp. 909–913, 2022.

[4] T. R. Damayanti, A. L. Kusumaningrum, Y. D. Susanty, and S. Susilawati Islam, “Route Optimization Using Saving Matrix Method – A Case Study at Public Logistics Company in Indonesia,” in Proc. 5th NA Int. Conf. Industrial Engineering and Operations Management (IEOM), Detroit, MI, USA, Aug. 10–14, 2020.

[5] J. Huang, C.-W. Chu, and Y.-C. Tsai, “Container throughput forecasting for international ports in Taiwan,” Journal of Marine Science and Technology (Taiwan), vol. 28, no. 5, pp. 456–469, 2020, doi: 10.6119/JMST.202010_28(5).0015.

[6] H. N. Durmus Senyapar and A. Aksoz, “Empowering sustainability: A consumer-centric analysis based on advanced electricity consumption predictions,” Sustainability, vol. 16, no. 7, art. no. 2958, 2024, doi: 10.3390/su16072958.

[7] A. Yadav, C. K. Jha, and A. Sharan, “Optimizing LSTM for time series prediction in Indian stock market,” *Procedia Computer Science*, vol. 167, pp. 2091–2100, 2020, doi: 10.1016/j.procs.2020.03.257.

[8] M. A. Olawumi, B. I. Oladapo, and R. A. Olawale, “Revolutionising waste management with the impact of Long Short-Term Memory networks on recycling rate predictions,” *Waste Management Bulletin*, vol. 2, pp. 266–274, 2024, doi: 10.1016/j.wmb.2024.08.006.

[9] Ivantha, “Daily Solid Waste Dataset,” Kaggle, dataset. Accessed: Nov. 30, 2025. [Online]. Available: https://www.kaggle.com/datasets/ivantha/daily-solid-waste-dataset.

[10] O. Mudannayake, D. Rathnayake, J. D. Herath, D. K. Fernando, and M. Fernando, “Exploring Machine Learning and Deep Learning Approaches for Multi-Step Forecasting in Municipal Solid Waste Generation,” IEEE Access, vol. 10, pp. 122570–122585, 2022, doi: 10.1109/ACCESS.2022.3221941.

[11] M. Bilgili, “Time series forecasting on cooling degree-days (CDD) using SARIMA model,” Natural Hazards, vol. 118, pp. 2569–2592, 2023, doi: 10.1007/s11069-023-06109-4.

[12] V. Shah, T. Gandhi, V. Gupta, and R. K. Gupta, “Efficient Medical Supply Chain Forecasting using Time Series Analysis,” in *2023 IEEE International Conference on Computer Vision and Machine Intelligence (CVMI)*, 2023, doi: 10.1109/CVMI59935.2023.10465213.

[13] S S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[14] A. Riyadi, N. Rokhman, and L. Heryawan, “A review on long short-term memory combination development,” *IAES International Journal of Artificial Intelligence*, vol. 14, no. 6, pp. 4427–4441, 2025, doi: 10.11591/ijai.v14.i6.pp4427-4441.

[15] S. Suradhaniwar, S. Kar, S. S. Durbha, and A. Jagarlapudi, “Time series forecasting of univariate agrometeorological data: A comparative performance evaluation via one-step and multi-step ahead forecasting strategies,” *Sensors*, vol. 21, no. 7, Art. no. 2430, 2021, doi: 10.3390/s21072430

[16] H. Hewamalage, K. Ackermann, and C. Bergmeir, “Forecast evaluation for data scientists: Common pitfalls and best practices,” *Data Mining and Knowledge Discovery*, vol. 37, pp. 788–832, 2023, doi: 10.1007/s10618-022-00894-5.

Joint Resource Management and Trajectory Optimization through Quantum-Inspired Reinforcement Learning in UAVs

Ihsan Ullah[1], Ghulam Musa Raza[2], Muhammad Nadeem Ali [3], Muhammad Asim [4],
Bareera Anam [5], Byung-Seo Kim[6]
[1,2,3,4,5]Hongik University, Sejong City 30016, Republic of Korea
danish1852@gmail.com, ghulammusaraza96@gmail.com, nadeemali499@hotmail.com,
araja3356@gmail.com, anambareera123@gmail.com, jsnbs@hongik.ac.kr

***Abstract*—This paper presents an Information-Centric Networking (ICN)-enabled Unmanned Aerial Vehicle (UAV) framework, where UAVs act as flying ICN nodes to deliver and cache content for ground users with heterogeneous Quality-of-Service (QoS) demands in a disaster environment. The QoS information embeds priority and deadline metadata, enabling urgency-aware scheduling for efficient content dissemination. To jointly optimise the UAV trajectory and resource allocation according to ground user demand, we adopt a quantum-inspired reinforcement learning (QiRL) approach. The proposed approach employs an amplitude-memory-based probabilistic action selection and a Grover-like amplitude amplification mechanism to balance exploration and exploitation efficiently by evaluating and ranking the collapse probabilities of potential actions, compared with traditional reinforcement learning techniques. Simulation results demonstrate that the proposed QiRL-ICN scheme significantly improves energy efficiency, deadline satisfaction, and convergence speed compared with conventional reinforcement learning methods.**

***Index Terms*—UAVs, ICN, Quantum-inspired Reinforcement learning, Resource Management, Trajectory Optimization.**

I. Introduction

Unmanned Aerial Vehicles (UAVs) have gained significant attention as an emerging technology for next-generation wireless communication systems due to their flexibility, rapid deployment, and cost-effectiveness. Compared with terrestrial communication infrastructures, UAV-assisted networks offer controllable mobility, enabling aerial platforms to dynamically adjust their positions to improve link quality and coverage [1]. One of the most important research challenges in such systems is the joint optimization of trajectory and resource management, since the UAV's flight path and transmission strategy directly affect the network's connectivity, throughput, and energy consumption.

Machine Learning (ML) techniques like Reinforcement Learning (RL) have been widely employed for UAV control and communication optimization. RL allows UAVs to autonomously learn optimal movement and resource allocation policies through trial-and-error interaction with the environment, without requiring explicit channel or user models. This learning capability enables adaptive decision-making under dynamic network conditions and supports intelligent resource management, including power control, bandwidth allocation, and user scheduling. Despite these advantages, conventional RL methods still face limitations when applied to high-dimensional joint trajectory resource problems, where continuous state and action spaces lead to slow convergence and unstable policy learning.

A fundamental difficulty in RL-based UAV systems lies in maintaining a proper balance between exploration and exploitation during learning. Classical exploration strategies such as ϵ-greedy or Boltzmann selection rely on stochastic sampling that often treats all unexplored actions uniformly, neglecting the relative quality between near-optimal and poor actions. This uniform exploration can result in redundant search and delayed convergence, particularly in large action spaces involving both spatial motion and communication resource variables. Therefore, more efficient and adaptive learning mechanisms are required to accelerate policy convergence and enhance the overall performance of UAV-assisted networks through intelligent trajectory control and dynamic resource management.

To address these challenges, this paper presents a Quantum-Inspired Reinforcement Learning (QiRL) framework for joint UAV trajectory and urgency-aware resource optimization in Information-Centric Networking (ICN). The QiRL model applies amplitude-memory-based probabilistic action selection and Grover-style amplitude amplification to accelerate learning while balancing exploration and exploitation. By embedding priority and deadline metadata within ICN Interest packets, the UAV performs demand-driven trajectory control and urgency-based resource allocation. Simulation results confirm that the proposed scheme improves energy efficiency, deadline satisfaction, and convergence stability over conventional RL approaches.

II. System Model

We consider a UAV-assisted Information-Centric Network (ICN), where UAVs act as aerial base stations and flying nodes to deliver requested content to K ground users in a disaster scenario. Each ground user generates Interest packets containing priority and deadline metadata to indicate QoS urgency.

The UAV determines its 3D position $\mathbf{c}(t) = [x(t), y(t), z(t)]$ and allocates transmit power $P_{\mathrm{Tx}}^{(k,n)}$ and bandwidth B_n across ground users based on urgency and channel quality. The objective is to maximize urgency-weighted throughput while minimizing energy consumption, subject to UAV flight, power, and QoS constraints.

III. Proposed Methodology

The proposed framework jointly optimizes the UAV trajectory and urgency-aware resource allocation in a disaster environment. The UAV acts as a flying base-station node capable of caching, forwarding, and delivering requested content to ground users. Ground users generate Interest packets containing priority and deadline metadata, which enable the UAV to evaluate the urgency and spatial distribution of user content requests. The UAV gathers information on user demands and resource competition, which is then evaluated in parallel by the QiRL agent to determine the optimal decision.

A. Learning Framework

In the proposed QiRL framework, each joint action comprising a UAV trajectory movement and a resource allocation strategy is represented by a quantum state vector given in Equation (7).

$$|\psi_t\rangle = \sum_{i=1}^{N} A_i(t)|a_i\rangle, \tag{1}$$

where $|a_i\rangle$ denotes the i^{th} composite action (trajectory + resource allocation), and $A_i(t)$ is its complex probability amplitude satisfying $\sum_i |A_i(t)|^2 = 1$. Initially, the amplitudes are uniformly distributed, forming a quantum superposition that allows the UAV to explore multiple possible actions in parallel.

A Grover-style amplitude amplification operator $\hat{G}$ is applied to update amplitudes toward optimal actions identified by a reward-based oracle $\hat{O}$. Mathematically, the evolution of the action state is given by Equation (7).

$$|\psi_{t+1}\rangle = \hat{G}\hat{O}|\psi_t\rangle, \tag{2}$$

where, $\hat{O}$ inverts the phase of the amplitudes corresponding to high-reward (feasible) actions, and $\hat{G}$ performs an inversion about the mean to amplify them according to Equation (7).

$$A_i(t+1) = A_i(t) + \mu\big(A_i(t) - \bar{A}(t)\big), \quad \forall i \in \mathcal{O}, \tag{3}$$

with $\bar{A}(t)$ being the mean amplitude and μ the amplification coefficient. Low-performing actions, particularly those leading to high energy consumption or deadline violations, are penalized through phase inversion, thereby reducing their future probability given by Equation (7).

$$A_i(t+1) = -A_i(t), \tag{4}$$

This quantum-inspired update rule enhances exploration ability while maintaining convergence stability, allowing the UAV to evaluate multiple trajectory–allocation combinations simultaneously and converge toward optimal policies faster than conventional RL.

B. Decision and Update Process

At each decision epoch, the UAV observes the state is given by Equation (7).

$$s_t = \{\mathbf{c}(t), E(t), \{U_k(t), h_k(t)\}_{k=1}^{K}\}, \tag{5}$$

where, $\mathbf{c}(t)$ is the UAV position, $E(t)$ the residual energy, $U_k(t)$ the urgency score, and $h_k(t)$ the channel condition. The QiRL agent samples an action $|a_i\rangle$ from the amplitude distribution $|A_i(t)|^2$. Upon execution, the UAV receives a reward defined in Equation (7), which balances urgency-weighted throughput and energy consumption.

$$r_t = \alpha \sum_{k \in \mathcal{S}_t} U_k(t) \log_2\big(1 + \mathrm{SINR}_k(t)\big) - \beta E_{\mathrm{cons}}(t), \tag{6}$$

The temporal-difference (TD) error serves as a feedback signal that simultaneously refines the amplitude-memory vector and updates the actor–critic network, as formulated in Equation (7) and Equation (8), respectively.

$$\delta_t = r_t + \xi V(s_{t+1}) - V(s_t) \tag{7}$$

$$A_i(t+1) = \mathrm{norm}\big(A_i(t) + \eta_A \,\mathrm{sign}(\delta_t)\, \mathbf{1}\{i = i_t\}\big), \tag{8}$$

where, η_A is the amplitude learning rate and $\mathrm{sign}(\delta t)$ shows whether to increase or decrease probability (amplify or suppress).

Through iterative application of these quantum-inspired operators, the UAV moves toward spatial regions with high aggregated urgency and throughput, while simultaneously allocating transmission resources to users with higher priority and tighter deadlines—achieving *energy-efficient, urgency-aware optimization* across both trajectory and communication layers.

IV. Conclusion and Future work

This paper presented a Quantum-Inspired Reinforcement Learning (QiRL) framework for joint UAV trajectory optimization and urgency-aware resource management in ICN. By leveraging quantum principles such as amplitude superposition and Grover-style amplification, the UAV efficiently explores and converges toward optimal trajectory–allocation policies. The proposed scheme effectively balances urgency-weighted throughput and energy consumption, offering a promising direction for energy-efficient and adaptive UAV communication systems.

Future research will extend the QiRL framework to multi-UAV cooperation, resource–computation integration, and realistic channel modeling to enhance scalability, robustness, and learning stability.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2022R1A2C1003549).

REFERENCES

[1] X. Yan, X. Fang, C. Deng, and X. Wang, "Joint optimization of resource allocation and trajectory control for mobile group users in fixed-wing uav-enabled wireless network," *IEEE Transactions on Wireless Communications*, vol. 23, no. 2, pp. 1608–1621, 2024.

A Cross Layer Design for Joint Caching and Routing in Information-Centric Internet of Vehicles

Muhammad Nadeem Ali[1], Ihsan Ullah[2], Ghulam Musa Raza[3], Muhammad Asim[4], Bareera Anam[5], and Byung-Seo Kim[6]
[1,2,3,4,5,6] Hongik University, Sejong City 30016, Republic of Korea
nadee@mail.hongik.ac.kr, danish1852@gmail.com, ghulammusaraza96@gmail.com,
anambareera123@gmail.com, araja3356@gmail.com, jsnbs@hongik.ac.kr

***Abstract*—Named Data Network (NDN) is an emerging Internet architecture that address the several limitations of conventional Transmission Control Protocol/Internet Protocol (TCP/IP) by offering unique features of content caching, request aggregation, and stateful forwarding strategies. Content caching is a critical feature of NDN that directly lower the latency and backhaul network traffic load by serving the request from its content store (CS). The existing literature has seen a large contribution in designing various caching schemes; however, these caching schemes mainly perform caching decisions without the prior knowledge of subsequent network layers, such as physical layer. In this paper, we are motivated to design a cross-layer caching approach to that consider the physical connectivity strength with the nearby network node in terms of signal strength and then formulate a caching decision. A central entity between the cache decision and the physical layer is proposed to decide the cache decision based on the current connectivity capability with surrounding nodes.**

I. Introduction

The rapid growth of smart applications has transformed traditional networks into highly dense infrastructures comprising billions of end-user devices, access points, relay nodes, and remote servers, including edge, fog, and cloud servers. This dense ecosystem demands high connectivity and seamless user experiences across diverse applications ranging from critical services to entertainment. The current Internet architecture, commonly known as the Transmission Control Protocol/Internet Protocol (TCP/IP), despite significant advancements, faces several limitations due to its location-based communication paradigm. To overcome TCP/IP limitations, Information-Centric Networking (ICN) [1] was proposed, and Named Data Networking (NDN) group is most prominent among all.

NDN brings the networking features based on the content name, rather than the host location. The content name-based network brings several key features, such as request aggregation, custom forwarding strategies, and introduces in-network caching. Additionally, NDN supports in-network caching, where content can be temporarily stored at intermediate nodes, reducing latency and enhancing the user experience. In NDN, user requests are represented by Interest packets, while the corresponding content is delivered through Data packets [2]. The Figure 1 illustrates the key difference between the TCP/IP and NDN network architectures.

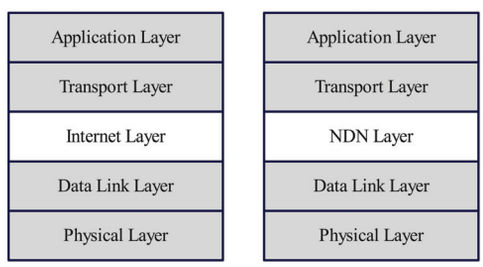


Fig. 1. TCP/IP network model vs NDN network model.

Named Data Networking (NDN) has demonstrated its capability to enhance various performance metrics in the Internet of Things (IoT), edge computing, satellite communication, unmanned aerial vehicle (UAV), and Internet of Vehicles (IoV) environments. Among these networks, the IoV is highly dynamic due to frequent mobility, which often leads to intermittent disconnections. Moreover, the IoV requires frequent content processing and handling at higher speeds compared to UAV networks, where the network topology remains relatively stable.

NDN provides in-network caching, allowing content to be stored temporarily. This reduces Interest packet traversal in the network and enables end-users to retrieve content from any node that has cached it. This directly lowers content retrieval time and improves server performance. Cached content is selected based on several factors, including content freshness, popularity, and the location of the caching node. Similarly, NDN forwarding strategies offer customizable schemes to address dynamic network environments. In wireless networks, NDN employs a broadcast approach, where an Interest packet is forwarded to all neighboring nodes. Any node that has the requested content responds. If none of the neighbors can serve the data, the Interest is forwarded to higher-level nodes.

Despite these features, NDN architecture faces challenges similar to those in TCP/IP, where each layer operates independently and communicates only through well-defined interfaces. At the NDN layer, the caching module stores content but lacks

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2022R1A2C1003549).

knowledge of physical layer conditions, such as signal strength and channel state. To address this limitation, a potential solution is cross-layer design. A centralized entity can inform the caching module about current connectivity with neighboring nodes, including signal strength and channel state, enabling the caching scheme to optimize decisions in accordance with the underlying layers.

II. Related Work

Shu-Ming T. et al. [3] proposed a heterogeneous interference graph neural network (HIGNN) to address cross-layer design issues in video streaming for device-to-device (D2D) communication. The proposed scheme jointly handles source encoding rate and transmission power control. Similarly, Zhaoming Hu et al. [4] introduced a joint scheme to predict user content preferences and optimize network resource allocation, directly affecting prediction accuracy, latency, and backhaul traffic. The authors employ a DRL-based reward function to optimize the multi-objective function, thereby improving routing decisions through continuous observation of network conditions.

III. Proposed Methodology

A vehicular environment is illustrated in Figure 2, comprising several vehicles, a roadside unit (RSU), and a controller mounted on the RSU. The RSU is connected to higher-level servers, such as the cloud. In this environment, vehicles can cache content and provide it to any requesting vehicle within the network. Existing caching schemes typically store data based on characteristics such as content popularity, freshness, and similar parameters. However, these schemes often overlook other critical factors, including the routing information of neighboring vehicles and the current signal strength between vehicles. A vehicle may cache large content, but due to connectivity conditions, including its own signal strength and the routing information of neighboring vehicles, it may become an ineffective node for content delivery.

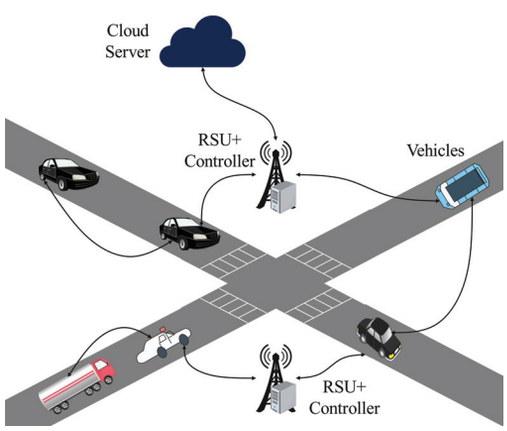


Fig. 2. IoV environment for joint content caching and routing in ICN.

In this study, we propose a scheme based on a central entity positioned between vehicles and infrastructure, such as Roadside Units (RSUs). This entity collects relevant information about each vehicle, including caching capacity, content status (in terms of popularity and freshness), and signal strength. Furthermore, the central entity gathers routing information for each vehicle to estimate connectivity with specific vehicles. We refer to this entity as the Central Caching Routing Unit (CCRU). Within the CCRU, a Q-learning-based system is deployed to exploit joint caching and routing information, aiming to enhance the cache hit ratio. This improvement directly reduces content retrieval latency and increases the interest satisfaction rate and cache hit ratio. However, this approach requires frequent information exchanges, which consume additional bandwidth. This proposed scheme requires extending the NDN functionality to acquire cache, channel, and mobility states, a mobility prediction module, neighbor cache information, a cost function for joint caching and routing discovery protocols, and RSU integration for cache placement and replacement..

IV. Conclusion

This study aims to jointly optimize caching and routing in information-centric IoV networks by utilizing a cross-layer caching and routing control unit (CCRU). The CCRU integrates the vehicle caching scheme, the connectivity state of vehicles, and the routing information of nearby vehicles, enabling coordinated and efficient data management across the network.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2022R1A2C1003549).

References

[1] B. Ahlgren, C. Dannewitz, C. Imbrenda, D. Kutscher, and B. Ohlman, "A survey of information-centric networking," *IEEE Communications Magazine*, vol. 50, no. 7, pp. 26–36, 2012.

[2] M. Imran, M. N. Ali, M. S. U. Din, M. A. U. Rehman, and B.-S. Kim, "An efficient communication and computation resources sharing in information-centric 6g networks," *IEEE Internet of Things Journal*, vol. 11, no. 16, pp. 27 275–27 294, 2024.

[3] S.-M. Tseng, S.-T. Wen, C. Fang, and M. Norouzi, "Cross layer power allocation by graph neural networks in heterogeneous d2d video communications," *IEEE Access*, vol. 13, pp. 44 484–44 496, 2025.

[4] Z. Hu, C. Fang, Z. Wang, J. Chen, S.-M. Tseng, and M. Dong, "Joint content caching and request routing for user-centric many-objective metaverse services," *IEEE Transactions on Network Science and Engineering*, vol. 12, no. 3, pp. 1911–1925, 2025.

A Cross-Layer ResDNN-Based Multicast Routing Protocol in Flying Ad-Hoc Networks

Amalia Amalia*, Yushintia Pramitarini†, Ridho Hendra Yoga Perdana‡, Kyusung Shim§, and Beongku An¶
*Dept. of Software and Communications Engineering in Graduate School, Hongik University, Republic of Korea
†Institute of Industrial Technology, Korea University, Republic of Korea
‡ School of Information and Communication Engineering, Chungbuk National University, Republic of Korea
§School of Computer Engineering and Applied Mathematics, Hankyong National University, Republic of Korea
¶Dept. of Software and Communications Engineering, Hongik University, Republic of Korea
Emails: amalia@mail.hongik.ac.kr, shintia88@korea.ac.kr, perdanarhy@ieee.org, kyusung.shim@hknu.ac.kr, beongku@hongik.ac.kr

***Abstract*—This paper proposes a novel cross-layer residual deep neural network (ResDNN)-based multicast (CRDM) routing protocol in flying ad-hoc networks(FANETs). The main contributions of this paper can be summarized as follows. Firstly, we consider the integration of cross-layer design and ResDNN to improve network performance. Next, we employ a cross-layer design that jointly incorporates physical layer information such as mobility, position, channel capacity, and residual energy, with the network layer information (route information), enabling optimal route selection through cost-function maximization. Last, we design a deep learning (DL) framework based on a ResDNN model to infer the optimal weight values of the cost function. The simulation results show that the proposed CRDM routing protocol achieves strong scalability, maintains a high packet delivery ratio (PDR), low routing delay, and low control overhead.**

***Index Terms*—Cross-layer, FANETs, Machine learning, Multicast Routing, ResDNN**

I. Introduction

The rapid evolution of 5G and beyond 5G communication has enabled new paradigms for autonomous and aerial networking, giving rise to flying ad-hoc networks (FANETs) [1]. FANETs, comprising highly mobile autonomous aerial vehicles (AAVs), provide flexible and adaptive communication infrastructures for various applications, including surveillance, disaster management, environmental monitoring, and intelligent transportation systems [2], [3]. However, the dynamic topology and frequent link disconnections caused by the high mobility of AAVs make routing a challenging task, particularly for multicast communication, which requires simultaneous and reliable data delivery to multiple destinations.

Ensuring quality of service (QoS) in multicast routing under such dynamic conditions demands adaptive and intelligent routing strategies. Several studies have explored the use of metaheuristic-based optimization techniques for efficient route selection. For instance, particle swarm optimization (PSO) algorithms [4] and hybrid genetic–firefly algorithms [5] have been employed to identify optimal routing paths. However, these approaches rely on iterative optimization methods, which struggle to adapt to rapid topology changes in FANETs [6], [7]. Moreover, the increasing scale of AAV deployments intensifies the need for routing protocols that can intelligently process cross-layer information while minimizing overhead. To overcome the limitations, we propose a deep learning (DL)-based multicast routing protocol using a residual deep neural network (ResDNN) to enhance overall network performance. The main contributions of this paper can be summarized as follows:

- We propose a novel cross-layer ResDNN-based multicast (CRDM) routing protocol in FANETs to establish an optimal multicast route from a source node to multiple destination nodes.
- We formulate a cost function maximization to determine the optimal multicast route based on a cross-layer design that integrates physical layer information such as mobility (speed and direction), position, residual energy, and channel capacity with network layer information (hop count).
- We design a DL framework based on the ResDNN model to optimize the weights of the cost function, leading to more accurate and robust multicast route selection in dynamic network environments.
- The simulation results show that the proposed CRDM routing protocol effectively establishes multicast routes while achieving high packet delivery ratio (PDR), low routing delay, and reduced control overhead.

II. System Model

A. System Model

In this paper, we propose a CRDM routing protocol in FANETs, which is designed to select the optimal route as illustrated in Fig. 1. This network consists of a set of flying access points (F-AP) nodes $\mathcal{N} = \{\mathtt{F}_n\}_{n=1}^{N}$ and a set of user equipment nodes $\mathcal{L} = \{\mathtt{UE}_l\}_{l=1}^{L}$, among which a node is designated as the source node $(\mathtt{S})$ and multiple destination nodes $(\mathtt{D})$. We assume each node obtains its position through a global positioning system (GPS), providing accurate location information for route discovery. However, due to highly dynamic node mobility, routing in FANETs remains challenging. To overcome this, the proposed CRDM routing protocol uses a cross-layer design that integrates the combined

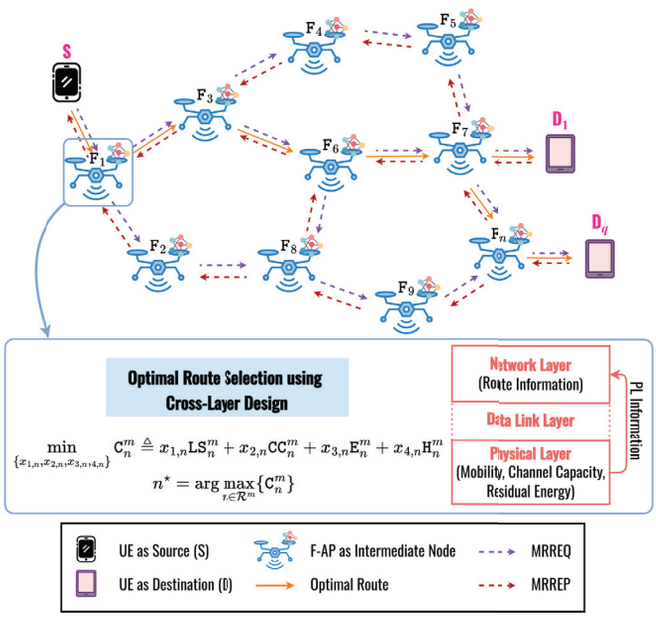


Fig. 1. The basic concept of the proposed CRDM routing protocol.

weight problem of physical layer information (i.e., mobility, residual energy, channel capacity) and network layer information (i.e., hop count) into a cost function. Furthermore, we use ResDNN to optimize this cost function for optimal route selection. Unlike traditional routing protocols that focus solely on the shortest route, the proposed CRDM routing protocol evaluates the cost values of multiple candidate routes while considering link stability, residual energy, and channel capacity. Consequently, the optimal route is determined as follows:

- $\mathtt{S} \rightarrow \mathtt{F}_1 \rightarrow \mathtt{F}_3 \rightarrow \mathtt{F}_6 \rightarrow \mathtt{F}_7 \rightarrow \mathtt{D}_1$
- $\mathtt{S} \rightarrow \mathtt{F}_1 \rightarrow \mathtt{F}_3 \rightarrow \mathtt{F}_6 \rightarrow \mathtt{F}_7 \rightarrow \mathtt{F}_n \rightarrow \mathtt{D}_q$.

B. The Proposed CRDR Protocol

The proposed protocol considers optimizing metrics such as link stability, channel capacity, residual energy, and hop count to select the optimal route. The algorithm of the proposed CRDM routing protocol can be summarized as follows:

- **Step 0: Initialization**
  - Each node obtains mobility (speed and direction), position, channel capacity, and residual energy from the physical layer by using the cross-layer design to establish a route in the network layer.
  - When multiple destination $\mathtt{D}_q$ nodes need to receive certain data from S node, they join a multicast membership by broadcasting a join request (JR) packet to the nearest F-AP. The (JR) packet contains the following information:

$$\langle \mathtt{Type}, \mathtt{S_{ID}}, \mathtt{N_{ID}}, \mathtt{Stat} \rangle$$

where Type indicates packet type, $\mathtt{S_{ID}}$ indicates the source node ID, $\mathtt{N_{ID}}$ indicates the node ID, and Stat indicates the node status (join or leave), respectively. Furthermore, the F-AP records the multicast ID ($\mathtt{M_{ID}}$) associated with this request in its table and proceeds to **Step 1**.

Multicast Tree Route Discovery Process:

- **Step 1: Source Generates and Sends MRREQ**
  The S node generates a multicast route request (MRREQ) packet and sends it to its F-AP, which contains the following fields:

$$\langle \mathtt{Type}, \mathtt{S_{ID}}, \mathtt{MG_{ID}}, \mathtt{M_{ID}}, \mathtt{S_{seq}}, \mathtt{H}, \mathtt{TTL} \rangle$$

where $\mathtt{MG_{ID}}$ indicates the ID of the multicast group, $\mathtt{S_{seq}}$ indicates the source sequence number, H indicates the hop count, and TTL indicates the time to live of the packet in the network, respectively. Furthermore, it records the sender's ID, updates the routing table, and broadcasts MRREQ, then goes to **Step 2**.
- **Step 2: Intermediate Nodes Operation for MRREQ**
  We consider only F-APs that can forward the packet. Thus, when UEs receive a MRREQ packet, it will be dropped. Furthermore, when the intermediate node receives a MRREQ packet, it evaluates the freshness of the packet by the following steps:
  - **Step 2.1:** If the $\mathtt{S_{seq}}$ at the received MRREQ is larger than $\mathtt{S_{seq}}$ at the routing table, then go to **Step 2.2**. Conversely, if the $\mathtt{S_{seq}}$ at the received MRREQ equals to $\mathtt{S_{seq}}$ at the routing table and the $\mathtt{M_{ID}}$ at the received MRREQ is equal to $\mathtt{M_{ID}}$ at the routing table and $\{\mathtt{H}+1\}$ at the received MRREQ is less than H at the routing table, then go to **Step 2.3**. Otherwise, the packet will be dropped.
  - **Step 2.2:** If the TTL at the received MRREQ is less than or equal to the TTL at the routing table, then go to **Step 2.3**. Otherwise, the packet will be dropped.
  - **Step 2.3:** If $\mathtt{N_{ID}}$ is the same as the $\mathtt{MG_{ID}}$, it records the sender's ID, updates the routing table, and broadcasts MRREQ to the neighbor nodes, then proceeds to **Step 3**. Otherwise, it records the sender's ID, updates the routing table, and broadcasts MRREQ to the neighbor nodes, then goes to **Step 2**.

Multicast Route Reply Process:

- **Step 3: F-AP Operation for MRREP**
  When $\mathtt{F}_n$ receives the MRREQ packet, it records the sender's ID, updates the routing table, and replies using a multicast route reply (MRREP) packet to the previous node. The MRREP packet contains the following information:

$$\langle \mathtt{Type}, \mathtt{S_{ID}}, \mathtt{D_{ID}}, \mathtt{M_{ID}}, \mathtt{D_{seq}}, \mathtt{pos}, \mathtt{sp}, \mathtt{dir}, \mathtt{E}, \mathtt{H} \rangle$$

where $\mathtt{D_{ID}}$ indicates the ID of the destination, $\mathtt{D_{seq}}$ indicates the destination sequence number, pos indicates the node position, sp indicates the node speed, dir indicates the node direction, and E indicates the node residual energy, respectively. Then, proceeds to **Step 4**.
- **Step 4: Intermediate Nodes Operation for MRREP**
  Upon receiving MRREP packet, intermediate nodes $\mathtt{F}_n$ process the following steps:

– **Step 4.1:** If $\mathtt{F}_n$ receives multiple MRREP packets for the same $\mathtt{D}_{\mathtt{ID}}$, it proceeds to **Step 4.2**. Otherwise, it directly proceeds to **Step 5**.
– **Step 4.2:** $\mathtt{F}_n$ calculates a cost function for each candidate next-hop node n at m-th hop by optimizing the weight values of link stability (LS), channel capacity (CC), residual energy (E), and hop count (H). The proposed cost function can be expressed as

$$\min_{\substack{\{x_{1,n},x_{2,n},\\ x_{3,n},x_{4,n}\}}} \mathtt{C}_n^m \triangleq x_{1,n}\mathtt{LS}_n^m + x_{2,n}\mathtt{CC}_n^m + x_{3,n}\mathtt{E}_n^m + x_{4,n}\mathtt{H}_n^m, \tag{1a}$$

$$\text{s.t} \quad x_{1,n} + x_{2,n} + x_{3,n} + x_{4,n} = 1, \tag{1b}$$

$$x_{1,n} > 0, x_{2,n} > 0, x_{3,n} > 0, x_{4,n} > 0, \tag{1c}$$

$$\mathtt{LS}_n^m \geq \mathtt{LS}^{\mathtt{th}}, \tag{1d}$$

$$\mathtt{CC}_n^m \geq \mathtt{CC}^{\mathtt{th}}, \tag{1e}$$

$$\mathtt{E}_n^m \geq \mathtt{E}^{\mathtt{th}}, \tag{1f}$$

$$\mathtt{H}_n^m \leq \mathtt{H}^{\mathtt{th}}, \tag{1g}$$

where (1b) and (1c) establish the requirement that the total weight must equal one and be a positive number. Constraints (1d)-(1g) specify link stability, channel capacity, residual energy, and hop count requirements. The objective function (1a) is nonconvex subject to the nonconvex constraints, making the problem highly challenging to solve optimally. In addition, finding the optimal weights in (1b) typically requires multiple iterations and is computationally demanding. To address this, we utilize the ResDNN model explained in Section III. By using this approach, we can deliver faster and more accurate results compared to conventional iterative methods when determining optimal values. The formulations for LS, CC, E, and H are given in [7], [8]. To make fairness all parameters (LS, CC, E, H) are normalized into the range [0,1] using the min–max normalization rule **k** = (**k**-min(**k**))/(max(**k**)-min(**k**)), where $\mathbf{k} \in \{\mathbf{ls}, \mathbf{cc}, \mathbf{e}, \mathbf{h}\}$, with **ls** = $[\mathtt{LS}_1^1, \ldots \mathtt{LS}_N^M]$, **cc** = $[\mathtt{CC}_1^1, \ldots \mathtt{CC}_N^M]$, **e** = $[\mathtt{E}_1^1, \ldots \mathtt{E}_N^M]$, and **h** = $[\mathtt{H}_1^1, \ldots \mathtt{H}_N^M]$. The optimal next-hop node is then selected based on the MRREP with the highest cost function value, which is determined as

$$n^\star = \arg\max_{n \in \mathcal{R}^m} \{\mathtt{C}_n^m\}, \tag{2}$$

where $\mathcal{R}$ denotes the set of neighboring candidate nodes. Then, proceeds to **Step 4.3**.
– **Step 4.3:** $\mathtt{F}_n$ records the sender's ID of the MRREP packet, updates the routing table, and forwards it to the previous node, then proceeds to **Step 5**.

- **Step 5: Source F-AP Operation for MRREP Packet** Upon receiving the MRREP packet, the F-AP of the source process performs the following steps:
  – **Step 5.1:** If the F-AP of the source receives all the MRREP packets from the multicast member, it informs S node, then goes to **Step 6**. Otherwise, go to **Step 5.2**.
  – **Step 5.2:** If the timer $\leq 2 \times$ TTL threshold, all the MRREP packets from the multicast member wait until the timer $> 2 \times$ TTL threshold. Otherwise, it informs S node and returns to **Step 1**.

Data Transmission Process:

- **Step 6: Data Transmission at The Source Node** Upon receiving all the MRREP packets, the S node multicasts the data packet to its destination nodes.

Furthermore, Table I presents the multicast tree routing table obtained. In this table, $\mathtt{S}_{\mathtt{ID}}$ denotes the source ID, $\mathtt{MG}_{\mathtt{ID}}$ represents the multicast group ID, $\mathtt{M}_{\mathtt{ID}}$ corresponds to the multicast ID, $\mathtt{S}_{\mathtt{seq}}$ indicate the source sequence number, $\mathtt{D}_{\mathtt{seq}}$ indicate the destination sequence number, $\mathtt{PN}_{\mathtt{ID}}$ represent the previous node ID, $\mathtt{NN}_{\mathtt{ID}}$ represent the next node ID, and C denotes the associated cost value, respectively.

TABLE I
ROUTING TABLE OF THE PROPOSED CRDM ROUTING PROTOCOL

$\mathtt{S}_{\mathtt{ID}}$	$\mathtt{MG}_{\mathtt{ID}}$	$\mathtt{M}_{\mathtt{ID}}$	$\mathtt{S}_{\mathtt{seq}}$	$\mathtt{D}_{\mathtt{seq}}$	$\mathtt{PN}_{\mathtt{ID}}$	$\mathtt{NN}_{\mathtt{ID}}$	C

III. THE PROPOSED DEEP LEARNING FRAMEWORK

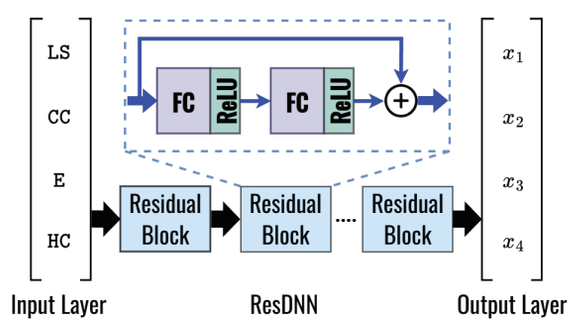


Fig. 2. The architecture of the proposed ResDNN model.

In the proposed CRDM routing protocol, a DL framework based on a ResDNN is integrated into the routing process to enhance decision-making and optimize network path selection. ResDNN is employed as it effectively captures complex nonlinear relationships and mitigates the vanishing gradient issue through residual connections. This architecture supports deeper network construction and improves the accuracy of the learning process [9]. The architecture of the proposed ResDNN can be seen in Fig. 2. It takes key network parameters such as LS, CC, E, and H as input features. These inputs are processed through multiple residual blocks to predict the optimal to predict the optimal weight of cost value x_1, x_2, x_3, and x_4. By learning the complex nonlinear relationships among these parameters, the ResDNN enables data-driven route selection, thereby improving overall performance in highly dynamic network environments. The proposed ResDNN model is constructed using two residual blocks, where each

block contains two fully connected hidden layers, and each layer comprises 64 neurons, as presented in Table II.

TABLE II
THE LAYER STRUCTURE OF THE PROPOSED RESDNN MODEL

Layer	Size	Activation / Operation
Input	4	-
Residual Block 1	FC1: 64	ReLU
	FC2: 64	ReLU
Skip Connection 1	-	Element-wise Addition
Residual Block 2	FC3: 64	ReLU
	FC4: 64	ReLU
Skip Connection 2	-	Element-wise Addition
Output	4	-

IV. PERFORMANCE EVALUATION

To evaluate the proposed CRDM routing protocol, we use the NS3 simulator under different maximum speeds (15 km/h, 30 km/h, 45 km/h, 60 km/h, and 75 km/h). We deploy 50 and 100 mobile nodes consisting of UE and F-AP in urban area measuring 1200 m $\times$ 1200 m. The 3D Gauss–Markov mobility model is adopted to realistically characterize the highly dynamic and three-dimensional movement patterns of UAVs in a FANET [6]. Table III shows the simulation environment and parameters. In addition, we measured the accuracy between the predicted optimal value and output data from the test set by calculating the root mean square error (RMSE) of the proposed ResDNN model [10].

TABLE III
SIMULATION ENVIRONMENTS AND PARAMETERS

Parameters	Value
Simulator	NS3 version 3.35
Simulation area	1,200 $\times$ 1,200 m^2
Mobility model	3D Gauss-Markov Mobility Model [6]
Simulation time	100 seconds
Session length	5 seconds
Number of nodes	50, 100
Node's speed range	[15:15:75 (km/h)]
RSSI threshold	-80 dbm
MAC protocol	802.11a

A. Performance Metrics

To comprehensively evaluate the effectiveness of the proposed CRDM routing protocol, the following metrics are considered [11]–[13]:

- PDR: defined as the ratio of the number of data packets successfully received at the multicast destination nodes to the total number of data packets transmitted by the source node.
- Routing delay: defined as the average delay to establish a multicast route from source to multicast destinations per one session.
- Control overhead: defined as the average number of control packets involved in the route creation process received by each node per multicast data delivered.

B. Simulation Results

Firstly, we evaluate the accuracy of the proposed ResDNN model by calculating the RMSE. RMSE evaluates the error between the model's predictions and the true values in the test dataset, where a lower RMSE indicates more accurate predictions. Fig. 3 illustrates that the prediction performance

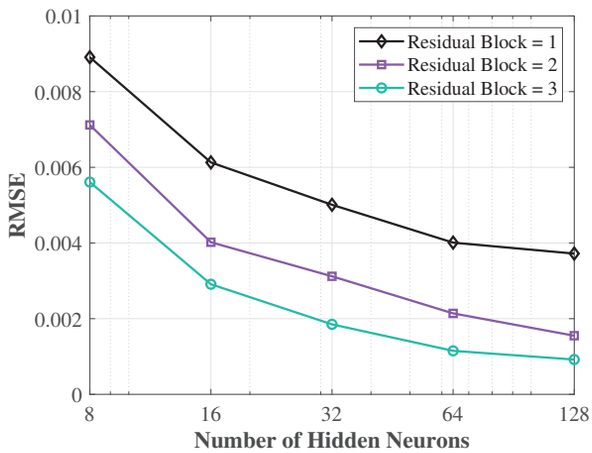


Fig. 3. RMSE versus number of hidden neurons.

of the ResDNN model is influenced by both the number of hidden neurons and the number of residual blocks. The RMSE decreases as the number of hidden neurons increases, indicating that more hidden layers enable the model to better learn complex nonlinear patterns. Moreover, adding more residual blocks yields additional performance gains. For instance, the configuration with three residual blocks consistently achieves lower RMSE than those with one or two blocks under the same hidden-neuron configurations. This demonstrates the effectiveness of residual learning, where skip connections support stable gradient flow and help deeper networks achieve higher accuracy without performance degradation.

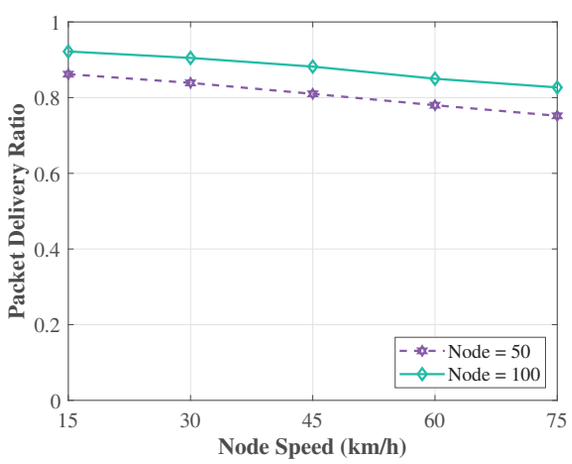


Fig. 4. PDR versus node speed.

Next, we evaluate the network performance metrics of the proposed CRDM routing protocol. Fig. 4 shows the PDR as a function of node speed with a different number of nodes. As the node speed increases, the PDR gradually decreases. This degradation is mainly due to the increased mobility of nodes, which causes frequent link breakages and route disruptions, making it difficult for routes to remain stable. Meanwhile, increasing the number of nodes results in a higher PDR. With greater node density, the network benefits from more available neighbors and alternative routes, which improve connectivity.

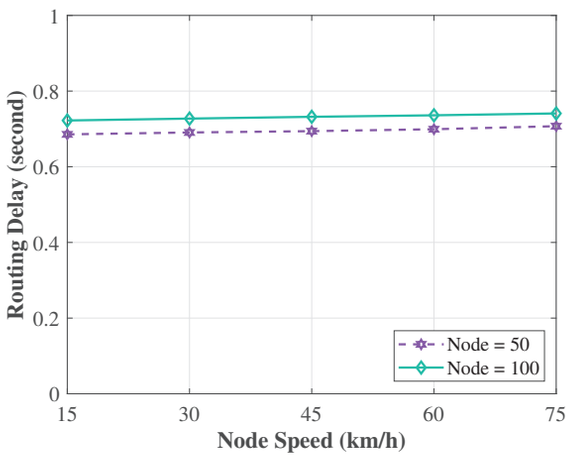


Fig. 5. Routing delay versus node speed.

Fig. 5 shows the routing delay as a function of node speed with a different number of nodes. The routing delay increases as the node speed increases due to frequent network topology changes, which leads to frequent interruptions in data transmission. Moreover, the routing delay increases as the number of nodes increases. In denser networks, a larger number of nodes participate in the routing process, resulting in higher control overhead and longer route establishment times. Consequently, the combined effects of higher mobility and increased network size contribute to a longer overall routing delay.

Fig. 6 shows control overhead as a function of node speed with a different number of nodes. The control overhead slightly increases as the node speed increases, since a higher speed leads to more frequent link breakages. Furthermore, the control overhead increases with the number of nodes in the network. This is mainly attributed to the higher node density, which results in more frequent routing updates, neighbor discovery operations, and control packet exchanges. Consequently, larger networks incur higher control overhead to maintain reliable multicast routing under dynamic mobility conditions.

Finally, Table IV compares the average execution time required to obtain the optimal cost-function weights using the conventional approach and the ResDNN approach. As the number of nodes increases, the average execution time of the conventional approach gradually increases from 0.554 s

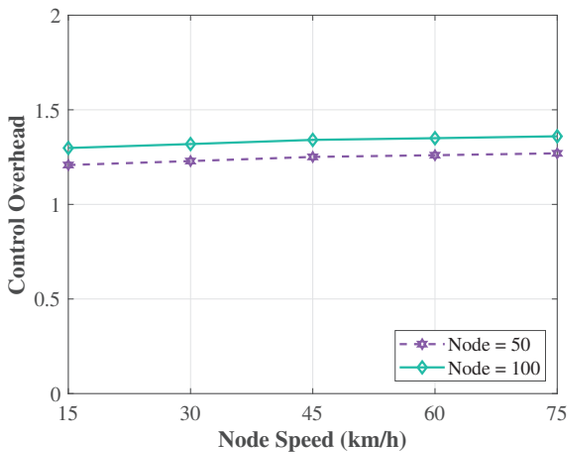


Fig. 6. Control overhead versus node speed.

TABLE IV
THE AVERAGE EXECUTION TIME FOR CONVENTIONAL VS. RESDNN

Number of Nodes {F, UE}	Conventional [s]	ResDNN [s]
{25, 5}	0.554	0.124
{40, 10}	0.611	0.161
{55, 15}	0.688	0.208
{80, 20}	0.785	0.265

to 0.785 s due to its reliance on iterative and computation-intensive optimization procedures. In contrast, the ResDNN approach exhibits a much lower execution time, increasing only slightly from 0.124 s to 0.265 s as the network size increases. This demonstrates that the ResDNN approach scales more efficiently with network size. The performance gain is attributed to the fact that most of the computational complexity is shifted to the offline training phase, enabling the trained ResDNN model to infer the optimal cost-function weights almost instantly during the routing process.

V. CONCLUSION

In this paper, we proposed a novel CRDM routing protocol in FANETs. The proposed CRDM routing protocol integrated a cross-layer design and ResDNN to improve network performance. Specifically, we employed a cross-layer design that combined physical layer information such as mobility, position, channel capacity, and residual energy, with network layer information (route information), enabling optimal route selection through cost-function maximization. In addition, we designed a DL framework based on the ResDNN model to infer the optimal weight values of the cost function. The simulation results demonstrated that the proposed CRDM routing protocol achieved strong scalability, maintaining high PDR, low routing delay, and low control overhead.

ACKNOWLEDGMENT

This work was supported by National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2022R1A2B5B01001190). Prof. Beongku An is the corresponding author.

REFERENCES

[1] M. K. Banafaa, O. Pepeoglu, I. Shayea, A. Alhammadi, Z. A. Shamsan, M. A. Razaz, M. Alsagabi, and S. Al-Sowayan, "A Comprehensive Survey on 5G-and-Beyond Networks With UAVs: Applications, Emerging Technologies, Regulatory Aspects, Research Trends and Challenges," *IEEE Access*, vol. 12, pp. 7786–7826, 2024.

[2] S. A. H. Mohsan, M. A. Khan, F. Noor, I. Ullah, and M. H. Alsharif, "Towards the Unmanned Aerial Vehicles (UAVs): A Comprehensive Review," *Drones*, vol. 6, no. 6, 2022.

[3] Y. Pramitarini, R. H. Y. Perdana, K. Shim, and B. An, "DLSMR: Deep Learning-Based Secure Multicast Routing Protocol against Wormhole Attack in Flying Ad Hoc Networks with Cell-Free Massive Multiple-Input Multiple-Output," *Sensors*, vol. 23, no. 18, 2023.

[4] X. Bao, H. Li, G. Zhao, L. Chang, J. Zhou, and Y. Li, "Efficient clustering V2V routing based on PSO in VANETs," *Measurement*, vol. 152, p. 107306, 2020.

[5] G. D. Singh, M. Prateek, S. Kumar, M. Verma, D. Singh, and H.-N. Lee, "Hybrid Genetic Firefly Algorithm-Based Routing Protocol for VANETs," *IEEE Access*, vol. 10, pp. 9142–9151, 2022.

[6] Y. Pramitarini, R. Hendra Yoga Perdana, K. Shim, and B. An, "Secure Multicast Routing Against Collaborative Attacks in FANETs With CF-mMIMO and STAR-RIS: Blockchain and Federated Learning Design," *IEEE Internet Things J.*, vol. 12, no. 12, pp. 22 404–22 426, 2025.

[7] A. Amalia, Y. Pramitarini, R. H. Y. Perdana, K. Shim, and B. An, "Energy-Efficient Multicast Routing Protocol Using FL-Based Optimal Route Selection in IoT-Enabled MANETs With RIS and CF-mMIMO," *IEEE Internet Things J.*, vol. 12, no. 22, pp. 47 588–47 606, 2025.

[8] A. Amalia, Y. Pramitarini, R. Hendra Yoga Perdana, K. Shim, and B. An, "A Cross-Layer FL-Based Clustering Protocol to Support Multicast Routing in IoT-Enabled MANETs With CF-mMIMO," *IEEE Access*, vol. 13, pp. 3881–3899, 2025.

[9] X. Wang, Y. Han, W. Xue, G. Yang, and G. Zhang, "Stable climate simulations using a realistic general circulation model with neural network parameterizations for atmospheric moist physics and radiation processes," *Geoscientific Model Development*, vol. 15, pp. 3923–3940, 2022.

[10] A. Amalia, Y. Pramitarini, R. H. Y. Perdana, K. Shim, and B. An, "A Deep-Learning-Based Secure Routing Protocol to Avoid Blackhole Attacks in VANETs," *Sensors*, vol. 23, no. 19, 2023.

[11] Y. Pramitarini, R. H. Y. Perdana, T.-N. Tran, K. Shim, and B. An, "A Hybrid Price Auction-Based Secure Routing Protocol Using Advanced Speed and Cosine Similarity-Based Clustering against Sinkhole Attack in VANETs," *Sensors*, vol. 22, no. 15, 2022.

[12] T.-N. Tran, T.-V. Nguyen, K. Shim, D. B. da Costa, and B. An, "A Deep Reinforcement Learning-Based QoS Routing Protocol Exploiting Cross-Layer Design in Cognitive Radio Mobile Ad Hoc Networks," *IEEE Trans. Veh. Technol.*, vol. 71, no. 12, pp. 13 165–13 181, 2022.

[13] T.-N. Tran and B. An, "QoS Multicast Routing Utilizing Cross-Layer Design for IoT-Enabled MANET in RIS-Aided Cell-Free Massive MIMO," *IEEE Internet Things J.*, vol. 11, no. 7, pp. 11 876–11 893, 2024.

Transfer Learning for Joint Precoding and Phase Shift Optimization in Mobile RIS-Aided Networks

Yosefine Triwidyastuti*, Tri Nhu Do†, Kyusung Shim‡, and Beongku An§
*Dept. of Software and Communications Engineering in Graduate School, Hongik University, Sejong, Republic of Korea.
†Department of Electrical Engineering, Polytechnique Montréal, Montréal, Québec, Canada.
‡School of Computer Engineering and Applied Mathematics, Hankyong National University, Anseong, Republic of Korea.
§Department of Software and Communications Engineering, Hongik University, Sejong, Republic of Korea.
Emails: *yosefine@mail.hongik.ac.kr, †tri-nhu.do@polymtl.ca, ‡kyusung.shim@hknu.ac.kr, §beongku@hongik.ac.kr

***Abstract*—Mobile networks have significant challenges in optimizing their networks, especially short coherence time and training data scarcity. To maximize the sum rate performance in mobile RIS-aided networks, we propose unsupervised multitask CNN with transfer learning in optimizing discrete phase shift at RIS, beamforming at multi-antenna BS, and power allocation for mobile users. The unsupervised method is utilized to eliminate the need for labeled data and reduce the computational cost, while transfer learning is beneficial for training in mobile networks that have limited data due to environment variation. Numerical results show that target model has higher sum rate although the CNN model is trained with limited target domain.**

***Index Terms*—beamforming, CNN, mobile nodes, RIS, transfer learning, unsupervised learning.**

I. Introduction

Node mobility introduces significant challenges for reconfigurable intelligent surface (RIS) optimization in wireless networks. Rapid time variations in channel coefficients that are primarily caused by Doppler spread, lead to channel aging and time-selective fading, which degrade the reliability of beamforming and resource allocation [1]. Consequently, mobile RIS-assisted systems must perform joint beamforming and phase shift optimization under short coherence times, demanding algorithms with extremely low latency and computational complexity to maintain performance within sub-millisecond decision intervals [2], [3]. Mobility fundamentally limits the controllability of propagation environments in next-generation wireless systems by reducing the temporal stability of RIS configurations, thereby constraining how precisely and how long the wireless channel can be shaped. At the same time, mobility-driven dynamics generate highly non-stationary data, which challenges AI models by requiring continual adaptation, rapid retraining, and robust generalization from limited, fast-evolving datasets in such environments [4].

Transfer learning is one of the potential candidates to make rapid decision-making with fast sampling efficiency [5]. Transfer learning is an advanced machine learning (ML) technique that involves transferring knowledge from one domain (source domain) to another (target domain). This approach is beneficial in scenarios where target domain has limited data but shares similarities with source domain [4]. However, high mobility causes frequent domain shifts due to rapidly changing channel statistics, which can significantly reduce the relevance of knowledge transferred from the source domain. As a result, negative transfer may occur when previously learned representations fail to generalize to fast-evolving target environments, requiring continual domain adaptation and lightweight online fine-tuning. In addition, from an optimization and radio resource allocation perspective, mobility shortens channel coherence time and introduces uncertainty into channel state information, making classical optimization solutions quickly outdated [6]. This necessitates robust, predictive, and low-complexity optimization frameworks that can operate with partial or imperfect information while maintaining real-time responsiveness.

In this paper, we address joint precoding and phase shift (JPPS) optimization problem in mobile MIMO networks assisted by RIS with discrete configuration constraint. We establish unsupervised learning to train CNN model without label and cost low computation. Consecutively, we employ transfer learning to tackle domain shift under mobility scenarios. The main contributions of this paper can be summarized as follows:

- We formulate JPPS optimization problem for sum rate (SR) maximization with discrete phase shift (PS) constraint and diverse node mobility.
- We establish unsupervised multitask CNN model to optimize PS at RIS, multi-antenna beamforming (BF) at base station (BS), and power allocation (PA) for mobile users simultaneously without the need of labeled data.
- We employ transfer learning to retrain the unsupervised multitask CNN model and handle the domain shift problem in mobile networks that have limited training data.
- Numerical results show that target model obtains higher SR performance with lower training time and maintained execution time from source model, validating transfer learning efficacy under mobility scenarios.

II. System Model

Let us consider a downlink transmission from a fixed BS (S) with M antennas located in y-z plane to U single-antenna mobile users (D_u) with velocity $\vec{v}_u$ (different speed v_u and direction η_u for each user). We assume that signal transmissions from S to all users are blocked by obstacles. Hence, there is no direct link transmission and a fixed RIS (R) with N passive reflecting elements located in x-z plane assists

the signal transmission. Fig. 1 illustrates the system model of mobile RIS-aided network, where the signal transmission in main channel comes from S to R, then is reflected from R to all mobile users.

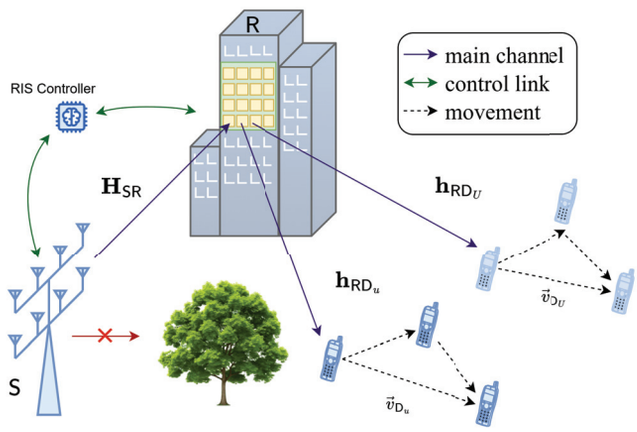


Fig. 1: System Model of Mobile RIS-MIMO

The transmitted signal with data $\mathbf{s} = [s_1, \ldots, s_U]^{\mathsf{T}} \in \mathbb{C}^{U\times 1}$ can be written as

$$\mathbf{x} = \mathbf{V}\mathbf{s} = \mathbf{W}\mathbf{P}\mathbf{s} = \sum_{u=1}^{U} \sqrt{p_u}\mathbf{w}_u s_u \in \mathbb{C}^{M\times 1}, \tag{1}$$

where $\mathbf{V} = \mathbf{W}\mathbf{P} \in \mathbb{C}^{M\times U}$ is the precoding matrix with BF matrix $\mathbf{W}$ and PA matrix $\mathbf{P}$ for all users. $\mathbf{W} \triangleq \{\mathbf{w}_u\}_{u=1}^{U} = [\mathbf{w}_1, \ldots, \mathbf{w}_U] \in \mathbb{C}^{M\times U}$ is the collection of BF vectors for all users that represent the spatial diversity for multi-antennas at BS towards all users. $\mathbf{P} = \mathrm{diag}\{\sqrt{p_1}, \ldots, \sqrt{p_U}\} \in \mathbb{R}^{U\times U}$ is the transmit PA across all users.

The received signal at each user with PS matrix $\mathbf{\Theta} = \mathrm{diag}\{e^{\mathrm{j}\theta_1}, \ldots, e^{\mathrm{j}\theta_N}\}$ can be expressed as

$$r_u = \sum_{d=1}^{U} \sqrt{p_d}\ \mathbf{h}_{\mathsf{RD}_u}^{\mathsf{H}} \mathbf{\Theta}\mathbf{H}_{\mathsf{SR}}\mathbf{w}_d s_d + n_{\mathsf{D}_u}, \tag{2}$$

where $n_{\mathsf{D}_u} \sim \mathcal{CN}(0, \sigma_n^2)$ is additive white Gaussian noise (AWGN) at user u. For the sake of simplicity, we assume that the amplitude reflection coefficient is $\kappa_n = \kappa = 1\ \forall\ n$. $\mathbf{H}_{\mathsf{SR}} \in \mathbb{C}^{N\times M}$ and $\mathbf{h}_{\mathsf{RD}_u} \in \mathbb{C}^{N\times 1}$ are the channel coefficients for $\mathsf{S}-\mathsf{R}$ and $\mathsf{R}-\mathsf{D}_u$ links, respectively.

The received signal to interference plus noise ratio (SINR) at user u can be written as

$$\gamma_u(\mathbf{\Theta}, \mathbf{P}, \mathbf{W}) = \frac{p_u |\mathbf{h}_{\mathsf{RD}_u}^{\mathsf{H}} \mathbf{\Theta}\mathbf{H}_{\mathsf{SR}}\mathbf{w}_u|^2}{\sum_{d\neq u} p_d |\mathbf{h}_{\mathsf{RD}_u}^{\mathsf{H}} \mathbf{\Theta}\mathbf{H}_{\mathsf{SR}}\mathbf{w}_d|^2 + \sigma_n^2}, \tag{3}$$

where $|\cdot|$ is absolute function. Hereafter, the achievable rate of user u can be expressed as

$$R_u(\mathbf{\Theta}, \mathbf{P}, \mathbf{W}) = B_W \log_2\big(1 + \gamma_u(\mathbf{\Theta}, \mathbf{P}, \mathbf{W})\big), \tag{4}$$

where B_W is the system bandwidth.

The objective of this paper is to maximize the SR by optimizing PS, PA, and BF matrices simultaneously that can be written as

$$\underset{\mathbf{\Theta},\mathbf{P},\mathbf{W}}{\text{maximize}} \quad \mathrm{SR} = \sum_{u=1}^{U} R_u(\mathbf{\Theta}, \mathbf{P}, \mathbf{W}) \tag{5a}$$

$$\text{subject to} \quad \theta_n \in \left\{0, \frac{2\pi}{2^B}, \cdots, \frac{2\pi(2^B-1)}{2^B}\right\}\ \forall\ n, \tag{5b}$$

$$p_u > 0 \quad \forall\ u, \tag{5c}$$

$$\sum_{u=1}^{U} p_u \leq P_{\mathsf{S}}, \tag{5d}$$

$$\|\mathbf{w}_u\|^2 \leq 1 \quad \forall\ u, \tag{5e}$$

where B is the number of phase quantization bits at the RIS, P_{S} is the total transmit power available at S, and $\|\cdot\|$ is Euclidean norm. Constraint (5b) represents discrete phase shift at R with 2^B-level uniform quantization. Constraint (5c) and (5d) indicate PA constraints for all user. Constraint (5e) represents unit-norm BF targeted each user.

Optimization problem in (5) is non-convex due to the coupling of $\mathbf{\Theta}$, $\mathbf{P}$, and $\mathbf{W}$ at the SINR functions along with the complicated sum of log of the fractional quadratic expressions and discrete constraint in (5b) that is more practical than implementing continuous PS [7]. As a result, the optimization problem is NP-hard and there is no standard method to obtain its optimal solution efficiently [8].

In this study, the optimal solution of (5) is collected into an output vector $\mathbf{y}$ that can be expressed as

$$\begin{aligned}\mathbf{y} &= [\mathbf{\Theta}, \mathbf{P}, \mathbf{W}] \in \mathbb{R}^{(N+U+2MU)\times 1} \\ &= \big[\theta_1,..,\theta_N, p_1,..,p_U, |w_{11}|,..,|w_{MU}|, \angle w_{11},..,\angle w_{MU}\big]^{\mathsf{T}},\end{aligned} \tag{6}$$

where $\mathbf{y}$ is the stack of all optimization variables. The optimization task in (5) can be rewritten as

$$\mathbf{y}^* = \arg\max_{\mathbf{y}\in\mathcal{C}} \sum_{u=1}^{U} R_u(\mathbf{\Theta}, \mathbf{P}, \mathbf{W}), \tag{7}$$

where $\mathcal{C}$ is the set collection of optimization parameter space that can be written as

$$\mathcal{C} = \left\{\theta_n \in \mathcal{Q},\ p_u > 0,\ \textstyle\sum_u p_u \leq P_{\mathsf{S}},\ \|\mathbf{w}_u\|^2 \leq 1\ \forall u\right\}, \tag{8}$$

where $\mathcal{Q} = \left\{0, \frac{2\pi}{2^B}, \ldots, \frac{2\pi(2^B-1)}{2^B}\right\}$.

III. Unsupervised Multitask CNN With Transfer Learning

To solve the optimization problem in (5), we propose unsupervised multitask CNN. Unsupervised learning with customized loss function eliminates the need for labeled data, which is greatly beneficial in mobile networks. Meanwhile, the multi-task architecture is capable of handling several number of optimization variables simultaneously within a short execution time [9]. In addition, CNN model is utilized to capture the spatial features of channel coefficients between RIS elements and mobile users.

TABLE I: Input Matrix Components

Row Indices	Column Indices	Values
1 to U	1 to N	$\lvert h_{\mathsf{R}_n\mathsf{D}_u}\rvert$
1 to U	$N+1$ to $2N$	$\angle h_{\mathsf{R}_n\mathsf{D}_u}$

TABLE II: Output Vector Partition

Indices	Values	Component
1 to N	θ_n	only phase
N+1 to N+U	p_u	only magnitude
(N+U+1) to (N+U+MU)	$\lvert\mathbf{w}_u\rvert$	magnitude
(N+U+MU+1) to (N+U+2MU)	$\angle\mathbf{w}_u$	phase

The input matrix for the CNN model is collected from complex channel coefficients in RIS-MIMO networks that can be written as

$$\mathbf{X}=\begin{bmatrix} \lvert h_{\mathsf{R}_1\mathsf{D}_1}\rvert & \dots & \lvert h_{\mathsf{R}_N\mathsf{D}_1}\rvert & \angle h_{\mathsf{R}_1\mathsf{D}_1} & \dots & \angle h_{\mathsf{R}_N\mathsf{D}_1} \\ \vdots & & \vdots & \vdots & & \vdots \\ \lvert h_{\mathsf{R}_1\mathsf{D}_U}\rvert & \dots & \lvert h_{\mathsf{R}_N\mathsf{D}_U}\rvert & \angle h_{\mathsf{R}_1\mathsf{D}_U} & \dots & \angle h_{\mathsf{R}_N\mathsf{D}_U} \end{bmatrix}, \quad (9)$$

where $h_{\mathsf{R}_n\mathsf{D}_u}$ denotes the channel coefficient from the n-th element at R to the u-th user and $\mathbf{X}\in\mathbb{R}^{U\times 2N}$. The detailed partition of the input matrix can be seen in Table I.

The input matrix in (9) is presented by the magnitude and phase components of the channel coefficients since the output vector for the optimization methods in (7) is composed of the PS matrix that is the phase values, PA matrix that represents gain or magnitude, and BF matrix with complex values. Thus, mapping function can be conducted directly from the magnitude and phase components of the input matrix into the magnitudes and phases of the output vector. The detailed partition for the output vector in (7) is shown in Table II.

CNN model can be considered as a parameterized model that maps the input $\mathbf{X}$ into the output $\mathbf{y}$. The mapping in the CNN model can be expressed as

$$\hat{\mathbf{y}} = f_{\mathbf{\Gamma}}(\mathbf{X}), \quad (10)$$

where $\mathbf{\Gamma}$ denotes all weights and biases in CNN model.

In this paper, the source model is built from convolution layers and fully-connected (FC) layers that can be written as

$$f_{\mathbf{\Gamma}_\mathrm{S}}(\mathbf{X})=\left[f^{\mathtt{act}}\circ f^{\mathtt{fc}}_{\mathbf{\Gamma}^\mathrm{S}_4}\circ f^{\mathtt{fc}}_{\mathbf{\Gamma}^\mathrm{S}_3}\circ f^{\mathtt{fl}}\circ f^{\mathtt{p}}\circ f^{\mathtt{c}}_{\mathbf{\Gamma}^\mathrm{S}_2}\circ f^{\mathtt{p}}\circ f^{\mathtt{c}}_{\mathbf{\Gamma}^\mathrm{S}_1}\right](\mathbf{X}) \quad (11)$$

where $f^{\mathtt{lay}}_{\mathbf{\Gamma}^\mathrm{S}_l}$ is the operation at the l-th layer with $\mathtt{lay}\in\{\mathtt{c},\mathtt{p},\mathtt{fl},\mathtt{fc},\mathtt{act}\}$ represents convolution, pooling, flatten, FC, and activation layer. $\mathbf{\Gamma}^\mathrm{S}_l$ is the weights and biases at the l-th layer of the source model. Hence, $\mathbf{\Gamma}_\mathrm{S}\triangleq\{\mathbf{\Gamma}^\mathrm{S}_1,\mathbf{\Gamma}^\mathrm{S}_2,\mathbf{\Gamma}^\mathrm{S}_3,\mathbf{\Gamma}^\mathrm{S}_4\}$ denotes the parameters in the source model that should be optimized during training to correctly map the input and output relation.

On the other side, the target model is constructed from the source model with additional FC layer that can be expressed as

$$f_{\mathbf{\Gamma}_\mathrm{T}}(\mathbf{X}) = \left[f^{\mathtt{act}}\circ f^{\mathtt{fc}}_{\mathbf{\Gamma}^\mathrm{T}_5}\circ f^{\mathtt{fc}}_{\mathbf{\Gamma}^\mathrm{T}_4}\circ f^{\mathtt{fc}}_{\mathbf{\Gamma}^\mathrm{T}_3}\circ f^{\mathtt{fl}}\circ f^{\mathtt{p}}\circ f^{\mathtt{c}}_{\mathbf{\Gamma}^\mathrm{T}_2}\circ f^{\mathtt{p}}\circ f^{\mathtt{c}}_{\mathbf{\Gamma}^\mathrm{T}_1}\right](\mathbf{X}), \quad (12)$$

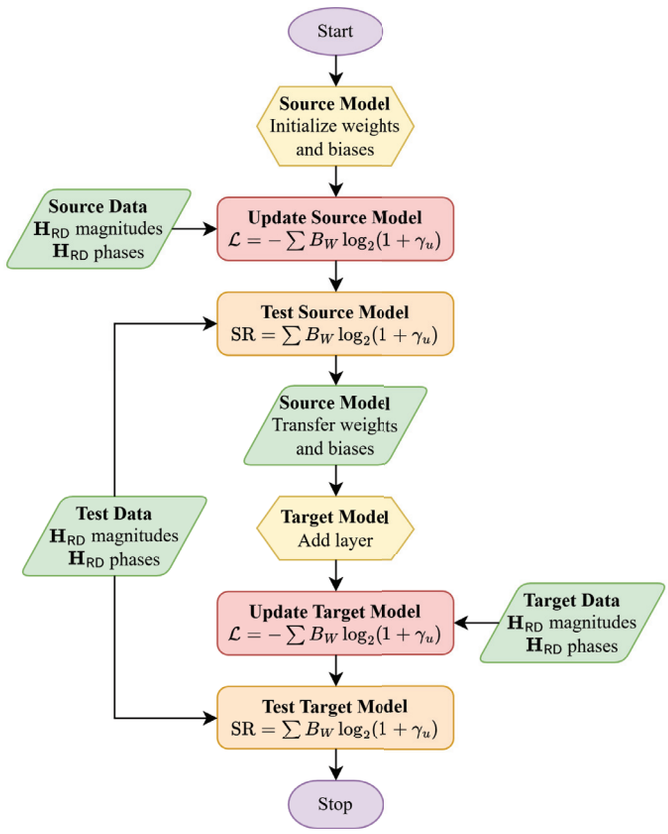


Fig. 2: Training pipeline with transfer learning

where $\mathbf{\Gamma}_\mathrm{T}\triangleq\{\mathbf{\Gamma}^\mathrm{T}_1,\mathbf{\Gamma}^\mathrm{T}_2,\mathbf{\Gamma}^\mathrm{T}_3,\mathbf{\Gamma}^\mathrm{T}_4,\mathbf{\Gamma}^\mathrm{T}_5\}$ are the weights and biases of the target model that should be optimized during training in the limited target domain.

Adding new layers is intuitive method in transfer learning [10]. Since the target dataset is different from source dataset, we need to add new layers located directly after the transferred layers to learn the specific characteristics of features in target dataset. The transferred layers in the front layers correspond to general characteristics of features that can be applicable to source and target datasets [11]. In this study, fine-tuning is applied to the target model, where the transferred parameters from the source model is retrained with target domain data to learn more data characteristics even though target model is only trained with insufficient amount of data [4].

Using assumption for the same number of channels or kernels at every convolution layer ($C^{(l)}=C\ \forall\ l$) and same number of neurons at FC layers ($n^{(l)}=C$), we can obtain the computational complexity of the source and target model in Table III with input matrix $\mathbf{X}$ as in (9). $N^\mathrm{S}_{\mathtt{batch}}$ is the number of batches in source domain data, whereas $N^\mathrm{T}_{\mathtt{batch}}$ is the number of batches in target domain data. It is noted that $N^\mathrm{T}_{\mathtt{batch}}\ll N^\mathrm{S}_{\mathtt{batch}}$ since the target domain is very limited. Therefore, the training complexity of the target model is less complex than the source model. However, their inference complexity is nearly same because target model only adds one FC layer after the source model.

TABLE III: Computational Complexity Comparison

Unsupervised CNN Model	Computational Complexity
Source Model Training	$\approx \mathcal{O}\left(N_{\texttt{epoch}}\ S_{\texttt{batch}}\ N^{\mathsf{S}}_{\texttt{batch}}\left(MNC^2+UNC^2+MUC\right)\right)$
Target Model Training	$\approx \mathcal{O}\left(N_{\texttt{epoch}}\ S_{\texttt{batch}}\ N^{\mathsf{T}}_{\texttt{batch}}\left(MNC^2+UNC^2+MUC\right)\right)$
Source Model Inference	$\approx \mathcal{O}\left(MNC^2+UNC^2+MUC\right)$
Target Model Inference	$\approx \mathcal{O}\left(MNC^2+UNC^2+MUC\right)$

For the backward propagation to train the source and target model without label, we propose a customized loss function that can be written as

$$\mathcal{L}(\mathbf{X},\hat{\mathbf{y}}) = -\mathrm{SR}\Big(\mathbf{X}, f_{\mathbf{\Gamma}}(\mathbf{X})\Big) = -\sum_{u=1}^{U} R_u\Big(\mathbf{X},\hat{\mathbf{\Theta}},\hat{\mathbf{P}},\hat{\mathbf{W}}\Big). \tag{13}$$

To satisfy all constraints in (5), we modify the loss function with the projected constraint set. The modified loss function can be expressed as

$$\mathcal{L}(\mathbf{X}) = -\sum_{u=1}^{U} R_u\big(\mathbf{X},\mathbf{\Theta}^{\mathsf{q}},\mathbf{P}^{\mathsf{n}},\mathbf{W}^{\mathsf{n}}\big). \tag{14}$$

The predicted PS $\hat{\mathbf{\Theta}}$ is quantized to the closest discrete value using the following calculation.

$$\mathbf{\Theta}^{\mathsf{q}} = \frac{2\pi}{2^B}\,\mathrm{round}\left(\frac{2^B}{2\pi}\hat{\mathbf{\Theta}}\right). \tag{15}$$

Concurrently, the predicted PA $\hat{\mathbf{P}}$ is projected by the sum of the PA for all users that can be expressed as

$$p_u^{\mathsf{n}} = \min\left\{1, \frac{P_{\mathsf{S}}}{\sum_{u=1}^{U}\hat{p}_u}\right\}\hat{p}_u. \tag{16}$$

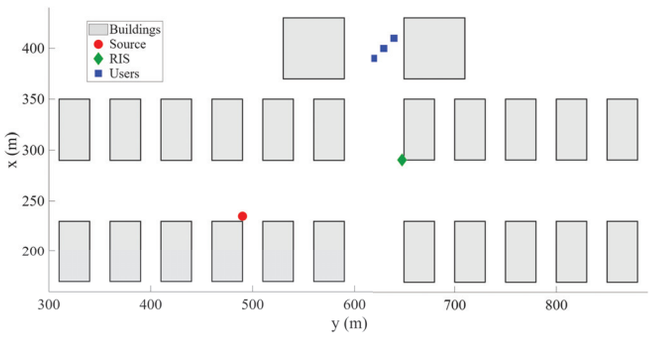
Lastly, the predicted BF $\hat{\mathbf{W}}$ is normalized by its user norm that can be written as

$$\mathbf{w}_u^{\mathsf{n}} = \frac{\hat{\mathbf{w}}_u}{\max\{1,\|\hat{\mathbf{w}}_u\|\}}. \tag{17}$$

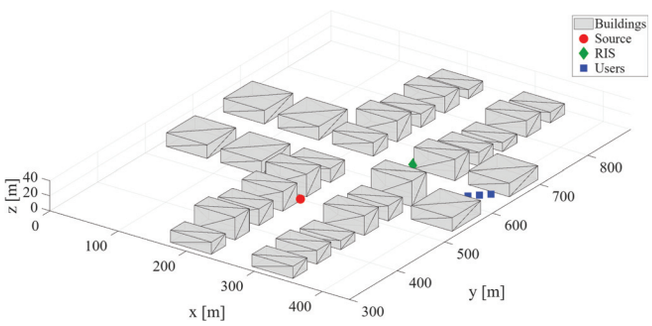
The training pipeline of the source and target model using transfer learning is shown in Fig. 2. First, source model is built using (11) with random initialization for its weights and biases. Then, source model parameters are updated using (14) with training data from source domain. Next, the trained source model conducts inference phase to calculate the resulted SR with test data. Then, source model's parameter values are copied and added one new FC layer to built the target model in (12). Target model parameters are then updated using (14) with training data from target domain. Lastly, target model's SR is evaluated using test data.

IV. Channel Data Generation Using Ray Tracing with Integrated Mobility Models

To generate realistic narrowband channel matrices for the JPPS problem, we compute the channel matrices ($\mathbf{H}_{\mathsf{SR}}$ and $\mathbf{H}_{\mathsf{RD}} \in \mathbb{C}^{N\times U}$) via a physics-based ray-tracing model. The constructed scene consists of multiple rectangular buildings whose 2-D and 3-D plots are shown in Fig. 3. In this scene, S and R both have height of 6m, while all mobile users' height is 2m.



(a) 2-D plot



(b) 3-D plot

Fig. 3: Illustration of the considered network simulation with (a) a top view and (b) a 3D view.

All mobile users are set under two different mobility scenarios, which are random walk mobility (RWM) and Gauss-Markov mobility (GMM). RWM captures the unpredictable movements of users, representing high randomness in mobility. On the other hand, GMM model has smooth movement that is more similar with real mobile node movement for the simulation of a personal communication system (PCS) [12].

A high-frequency ray-tracing model based on shooting-and-bouncing rays (SBR) is employed to obtain channel coefficients for all links, allowing up to four reflections. For every transmitter–receiver pair, the ray tracer returns a set of propagation paths, where each path (k) is characterized by a propagation distance (d_k), a total path loss (P_k), and all reflection points. The complex baseband gain of each path is computed as

$$g_k = a_k\ e^{-\mathrm{j}\ 2\pi f_c \tau_k}, \tag{18}$$

where f_c is the carrier frequency, $\tau_k = d_k/c$ is the propagation delay, c is the speed of light (3×10^8m/s), and $a_k = 10^{-P_k/20}$

is the path attenuation derived from the ray-tracer's path-loss output.

To enhance robustness, we retain only the K strongest propagation paths ranked by their linear power ($|a_k|^2$). The narrowband channel coefficient for the m-th antenna of S and n-th element of RIS is obtained by

$$(\mathbf{H}_{\mathsf{SR}})_{n,m} = \sum_{k=1}^{K_{nm}} g_k = \sum_{k=1}^{K_{nm}} a_k \; e^{-\mathrm{j}\; 2\pi f_c \tau_k}. \tag{19}$$

Similarly, the channel coefficient for the u-th user and n-th element of RIS can be expressed as

$$(\mathbf{H}_{\mathsf{RD}})_{n,u} = \sum_{k=1}^{K_{nu}} g_k = \sum_{k=1}^{K_{nu}} a_k \; e^{-\mathrm{j}\; 2\pi f_c \tau_k}. \tag{20}$$

This procedure is applied to every array element of S and R, producing fully populated channel matrices that incorporate the 3-D geometry of the environment, material reflections, and multipath effects with parameter settings in Table IV. Hence, the resulting CSI faithfully represents the physical propagation conditions and can be directly used for beamforming and discrete RIS phase-shift optimization.

TABLE IV: Simulation Parameter Settings

Parameter	Value
Carrier frequency (f_c)	28 GHz
Antenna spacing or wavelength ratio	0.5
System bandwidth (B_W)	50 MHz
Maximum number of paths (K)	25
Maximum number of reflections	4
Buildings material	concrete
Number of BS antennas (M)	4
Number of users (U)	3
Transmit power at S, (P_{S})	5 W
AWGN variance (σ^2)	-80 dBm
Number of RIS quantization bits (B)	3 bits
Learning rate (η)	0.001
Kernel size at every convolution layer	3×3
Number of convolution layers	2 layers
Number of stride (s)	1
Number of channels or kernels (C)	64
Number of source domain data ($N^{\mathsf{S}}_{\mathtt{patch}} \times S_{\mathtt{batch}}$)	100,000
Number of target domain data ($N^{\mathsf{T}}_{\mathtt{batch}} \times S_{\mathtt{batch}}$)	10,000
Batch size ($S_{\mathtt{batch}}$)	50
Number of epochs ($N_{\mathtt{epoch}}$)	30
Number of test data	1,000

V. Results and Discussion

Fig. 4 displays the loss minimization from the source and target model during their training. Target model training has lower loss than the source model that indicates the successful transfer learning from the source model to the target model. Source model parameters successfully learn the general features from the source domain and transfer the trained knowledge to the target model.

Fig. 5 shows the SR performance of unsupervised multitask CNN with transfer learning in RWM scenario. Target model performance is high because it received knowledge from source model. Thus, it learns data characteristics from all

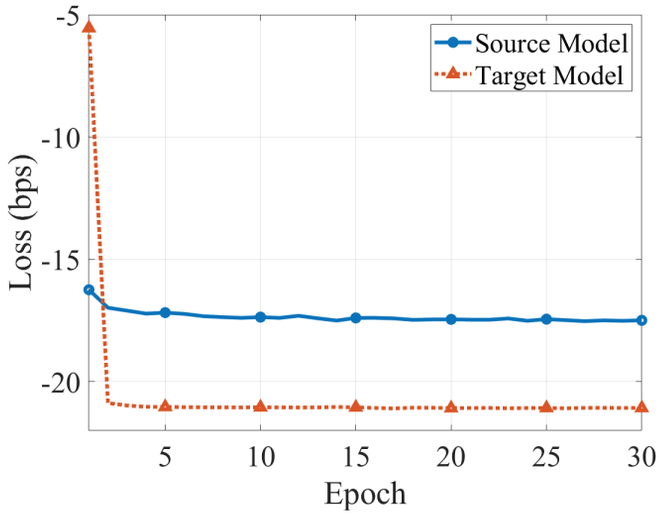


Fig. 4: Training loss of the source and target model.

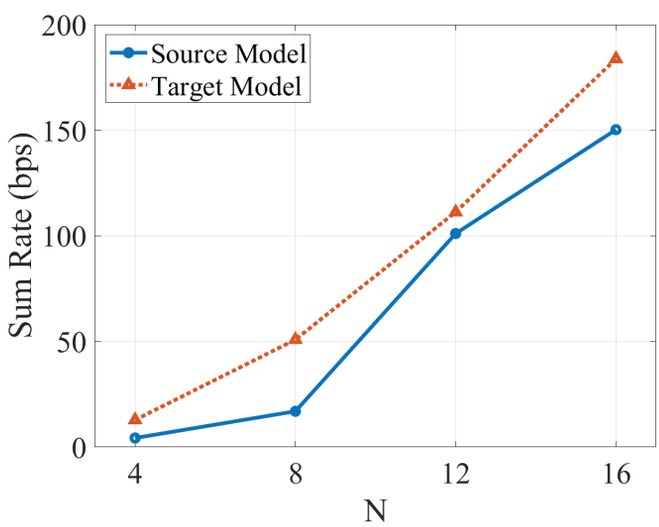


Fig. 5: SR performance of the source and target model.

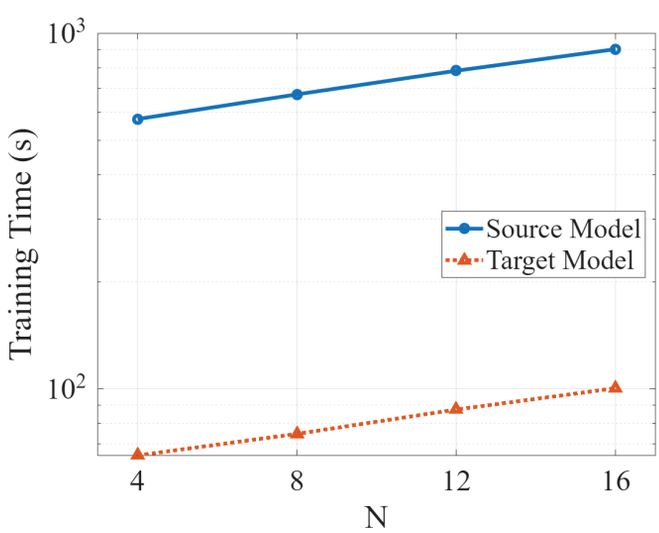


Fig. 6: Training time of the source and target model.

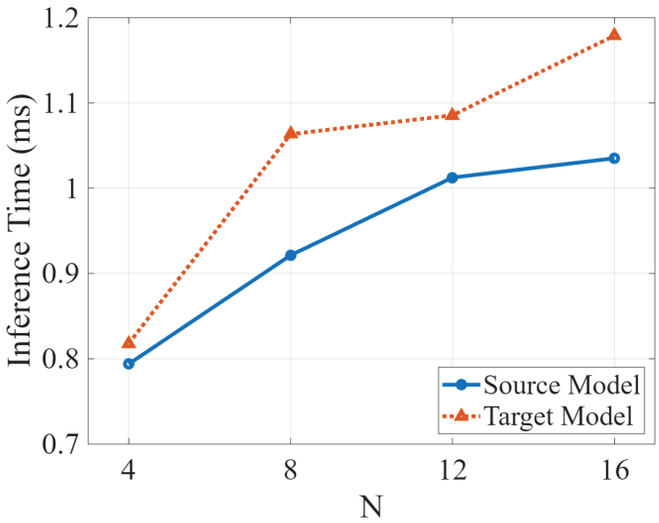


Fig. 7: Execution time of the source and target model.

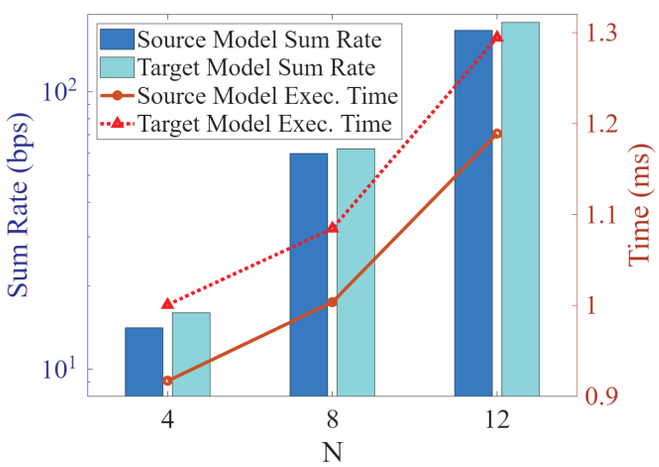


Fig. 8: CNN performance under GMM scenario.

source domain and target domain, while source model only learns data from source domain.

Fig. 6 depicts the consumed time for CNN training in source model and target model. Since target domain only has limited data, its training time is much shorter than source model's training time. This validates that less training time and higher performance can be achieved by the implementation of transfer learning, making it suitable for mobile networks.

Fig. 7 presents the required execution time of the source model and target model under RWM scenario. Target model requires more execution time because it has more FC layer than the source model. However, their time difference is insignificant since only one FC layer that is added to target model. Hence, target model execution is still feasible for mobile networks.

Fig. 8 portrays the performance comparison between source model and target model under GMM scenario. The unsupervised multitask CNN performance under GMM scenario has similar trend with RWM scenario. Target model has higher SR and execution time than source model because of the transferred knowledge from the source model and the addition of new FC layer in target model.

VI. Conclusion

This paper studied JPPS optimization problem in RIS-assisted mobile networks for SR maximization. We proposed unsupervised multitask CNN with transfer learning to obtain discrete PS, multi-antenna BF, and mobile users PA simultaneously. We trained the multitask CNN model with unsupervised learning to relieve the need of labeled data in wireless networks and consecutively with transfer learning to ease the limited data available in mobile networks. Numerical results show that transfer learning effectively trained the multitask CNN model with maintained execution time in the limited target domain data, suitable for mobile networks with rapid environment change.

Acknowledgment

This work was supported by National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2022R1A2B5B01001190). Prof. Beongku An is the corresponding author.

References

[1] C. Pan, G. Zhou, K. Zhi, S. Hong, T. Wu, Y. Pan, H. Ren, M. D. Renzo, A. L. Swindlehurst, R. Zhang, and A. Y. Zhang, "An Overview of Signal Processing Techniques for RIS/IRS-Aided Wireless Systems," *IEEE J. Sel. Top. Signal Process.*, vol. 16, no. 5, pp. 883–917, 2022.

[2] R. Liu et al., "6G Enabled Advanced Transportation Systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 9, pp. 10 564–10 580, Sept. 2024.

[3] T. L. Nguyen, G. Kaddoum, T. N. Do, and Z. J. Haas, "Ground-to-AAV and RIS-Assisted AAV-to-Ground Communication Under Channel Aging: Statistical Characterization and Outage Performance," *IEEE Trans. Commun.*, vol. 73, no. 9, pp. 8062–8079, 2025.

[4] Y. Triwidyastuti, T. N. Do, R. H. Y. Perdana, K. Shim, and B. An, "Transfer Learning-Empowered Physical Layer Security in Aerial Reconfigurable Intelligent Surfaces-Based Mobile Networks," *IEEE Access*, vol. 13, pp. 5471–5490, 2025.

[5] K. Shafique and M. Alhassoun, "Going beyond a Simple RIS: Trends and Techniques Paving the Path of Future RIS," *IEEE Open J. Antennas Propag.*, vol. 5, no. 2, pp. 256–276, April 2024.

[6] O. Wang, H. He, S. Zhou, Z. Ding, S. Jin, K. B. Letaief, and G. Y. Li, "Fast Adaptation for Deep Learning-Based Wireless Communications," *IEEE Commun. Mag.*, vol. 63, no. 10, pp. 158–164, Oct. 2025.

[7] B. Zheng, C. You, W. Mei, and R. Zhang, "A Survey on Channel Estimation and Practical Passive Beamforming Design for Intelligent Reflecting Surface Aided Wireless Communications," *IEEE Commun. Surveys Tuts.*, vol. 24, pp. 1035–1071, Secondquarter 2022.

[8] H. Zhou, M. Erol-Kantarci, Y. Liu, and H. V. Poor, "A Survey on Model-Based, Heuristic, and Machine Learning Optimization Approaches in RIS-Aided Wireless Networks," *IEEE Commun. Surveys Tuts.*, vol. 26, no. 2, pp. 781–823, 2024.

[9] Y. Triwidyastuti, T. N. Do, R. H. Y. Perdana, K. Shim, and B. An, "Unsupervised Learning-Based Joint Beamforming and Phase-Shift Optimization for RIS-Assisted DeepMIMO With Large-Scale Arrays," *IEEE Access*, vol. 13, pp. 150 852–150 871, 2025.

[10] H. Venkateswara, S. Chakraborty, and S. Panchanathan, "Deep-Learning Systems for Domain Adaptation in Computer Vision: Learning transferable feature representations," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 117–129, 2017.

[11] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. 27th NIPS - Volume 2*. Cambridge, MA, USA: MIT Press, 2014, p. 3320–3328.

[12] B. Liang and Z. J. Haas, "Predictive Distance-Based Mobility Management for Multidimensional PCS Networks," *IEEE/ACM Trans. Netw.*, vol. 11, no. 5, pp. 718–732, 2003.

Large-Scale Benchmarking of AI Models for Predictive Resource Allocation in Open RAN

1st Md Shakirul Islam *, 2nd Eun-Sung Jung *
* *Dept. of Software and Communications Engineering, Hongik University, Sejong, Korea*
Emails: shakirul.islam@mail.hongik.ac.kr, ejung@hongik.ac.kr

***Abstract*—The increasing complexity of Open Radio Access Network (O-RAN) environments requires predictive intelligence to manage heterogeneous slice resources efficiently. Traditional static allocation approaches fail to adapt to diverse service ty, such as eMBB, URLLC, and mMTC, leading to resource underutilizationion and degraded Quality of Service (QoS). While earlier studies explored AI-driven xApps/rApps frameworks for intelligent resource allocation, they were mostly limited to small-scale datasets, restricting generalization. In this work, we considered a large-scale, multi-base station dataset with over 600,000 time-series samples across 18 experimental configurations, enabling realistic benchmarking under varied traffic and scheduler conditions. Multiple classical and deep learning models are evaluated under different experimental settings. With over 600,000 multi-BS samples, CNN1D achieved the lowest NRMSE (0.085) and highest accuracy (0.860), outperforming the best classical baseline by up to a 15% error reduction, while lightweight models such as DT and KNN retained superior training efficiency for real-time deployment.**

***Index Terms*—Open RAN, Network Slicing, Resource Allocation, Machine Learning, Deep Learning, 5G, RIC, xAPP/rAPP**

I. Introduction

The rapid growth of connected devices and the demand for high-speed, low-latency and reliable services have placed increasing pressure on modern wireless networks. To address these challenges, next-generation systems such as 5G and 6G emphasize scalable and efficient architectures. Among them, the Open Radio Access Network (Open RAN) has emerged as a key enabler, offering openness, interoperability and virtualization through standardized interfaces and intelligent control functions such as the RAN Intelligent Controller (RIC) and its xApps/rApps. By decoupling hardware and software, Open RAN enables flexible deployment and more adaptive resource orchestration across heterogeneous environments [1].

Network slicing further enhances this flexibility by partitioning the physical infrastructure into virtual slices tailored to diverse QoS requirements, supporting services such as enhanced Mobile Broadband (eMBB), Ultra-Reliable Low-Latency Communications (URLLC), and massive Machine Type Communications (mMTC). However, maintaining slice performance under dynamic traffic, user mobility and heterogeneous scheduler behavior requires predictive, fine-grained resource allocation mechanisms rather than static or rule-based schemes.

Although Deep Reinforcement Learning (DRL) and other adaptive optimization techniques have been explored for slice-level control, their computational complexity and slow convergence limit practical deployment in real O-RAN systems. This highlights the need for lighter, predictive AI models that can anticipate short-term load variations and guide resource decisions without incurring the overhead of online DRL training.

Artificial Intelligence (AI) and Machine Learning (ML) have therefore emerged as promising tools for forecasting traffic patterns, detecting anomalies and supporting dynamic PRB allocation. By predicting near-future network conditions, ML-driven control can improve efficiency and stability while maintaining QoS across slices [2], [3].

Despite these advances, selecting appropriate AI models for real-time resource allocation in O-RAN remains a non-trivial challenge. Practical deployment requires not only high predictive accuracy, but also robustness across heterogeneous base stations, scheduler policies, and traffic regimes, together with low training and inference overhead suitable for near-real-time xApp operation. However, existing studies often rely on limited-scale or single-BS datasets, providing little guidance on model selection trade-offs under realistic O-RAN conditions.

To address this gap, we conduct a large-scale benchmarking study using over 600k time-series samples across multi-BS and multi-scheduler configurations. Both classical machine learning and deep learning models are evaluated under three settings—Single-BS, Multi-BS generalization, and Full-Scheduler scenarios—while also analyzing the impact of causal feature design (with and without grant/throughput inputs) on predictive performance.

The contributions of this work are threefold:

(i) a unified benchmarking protocol for classical and deep models in O-RAN-based network slicing;
(ii) a cross-BS generalization analysis across heterogeneous cells and schedulers; and
(iii) a causal variant study demonstrating how feature design influences next-step PRB prediction accuracy.

The rest of the paper is organized as follows. Section II reviews related studies on AI-driven resource allocation. Section III describes the system architecture, dataset, and experimental methodology. Section IV presents the evaluation results, and Section V concludes the paper.

II. Related Work

Network slicing has emerged as a fundamental technology that empowers 5G networks to accommodate a wide range

of services and application requirements within a shared infrastructure. However, traditional approaches that rely solely on static or rule-based physical resource management often struggle with scalability, adaptability and real-time decision-making in complex and dynamic network conditions. These limitations have motivated the adoption of AI-driven resource management frameworks, particularly within Open Radio Access Network (O-RAN) architectures, where intelligent control, virtualization and data-driven optimization can enhance flexibility, efficiency and end-to-end service quality.

Lotfi et al. [4] introduced an Attention-based Deep Reinforcement Learning (ADRL) approach to manage network slicing in O-RAN environments, addressing the diverse demands of 5G services. This approach deployed distributed DRL agents within the O-RAN framework that prioritize critical information using a value-attention mechanism, optimizing resource allocation and reducing SLA violations. The ADRL technique achieved a notable 32.8% improvement in QoS compliance compared to traditional DRL approaches. However, the findings were limited to simulation environments, lacking real-world validation in complex and dynamic networks.

In a subsequent work, Lotfi and Afghah [5] proposed an integrated framework that combines an LSTM-based traffic prediction model (rApp) running in the non-real-time RIC with a DDRL agent (xAPP) operating in the near-real-time RIC. Their LSTM-based model predicted traffic load patterns, allowing the DDRL agent to optimize dynamic resource allocation. Simulation results demonstrated a 7.7% improvement in cumulative rewards, higher throughput and lower SLA violations compared to standard DRL methods.

Yeh et al. [6] proposed a deep learning-based xApp-enabled slicing framework integrated into the near-real-time RIC to guarantee SLA compliance for various IoT services. Their approach significantly improved delay-sensitive performance while maintaining good throughput efficiency. Shen et al. [7] further examined the role of AI in optimizing network slicing for next-generation wireless networks (NGWNs), proposing an AI-assisted slicing architecture utilizing ML and RL for traffic prediction and dynamic resource allocation. Despite these advances, scalability and real-time adaptation remain open challenges.

Bega et al. [8] proposed an AI-based framework for networking slice management integrating DRL and CNNs into admission control resource orchestration and radio scheduling, demonstrating up to 80% performance gains. However, the authors highlighted challenges in achieving trustworthy and interpretable AI for real-world deployment.

Most of these studies, however, remain limited in scope. They primarily rely on small-scale or simulation-based datasets, lacking generalization across multiple base stations, and scheduler configurations. Moreover, previous research seldom evaluated feature explainability or cross-slice performance, which are crucial for understanding model scalability and reliability in real-world O-RAN systems. To bridge these gaps, Alshagri et al. [11] conducted benchmarking of multiple AI models within an O-RAN setup; however, their analysis was restricted to small-scale datasets and limited slice configurations. In contrast, our study extends this work to a large-scale, multi-base station environment, introducing two experimental variants-one utilizing previously granted resource information and another excluding it- to enable the model to learn dynamic allocation behavior based on requested PRBs, buffer load and downlink bit rate. This approach allows a more comprehensive evaluation of how different AI models adapt to varying resource availability and traffic conditions across base stations.

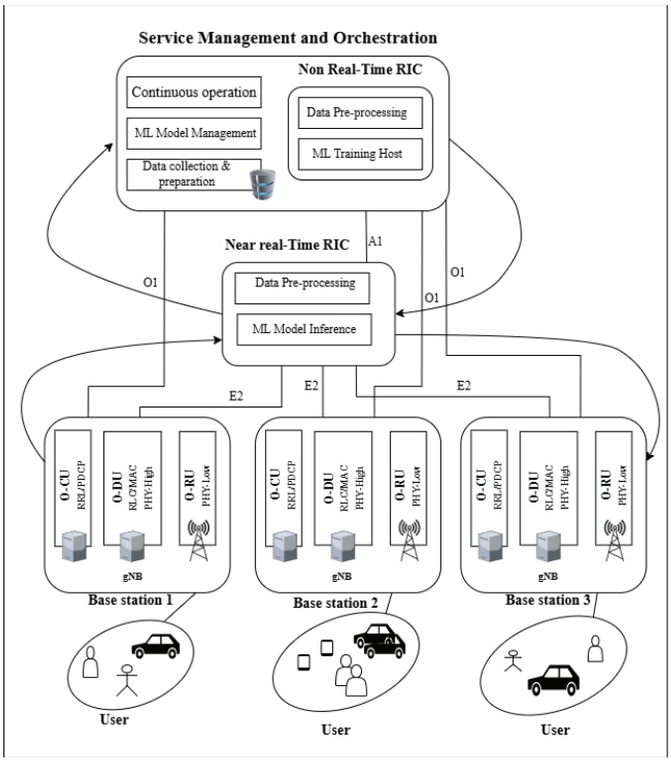


Fig. 1. Overall architecture of the Open RAN system.

III. System Model

This section outlines the proposed system model used for benchmarking intelligent resource allocation within the Open Radio Access Network (O-RAN) framework. It introduces the overall architecture, explains the dataset composition and scaling methodology and presents the evaluation metrics used to analyze model performance, fairness and deployability.

A. Overall Architecture

In O-RAN, resource management is coordinated by the RAN Intelligent Controller (RIC), which consists of a non-real-time (non-RT) RIC and a near-real-time (near-RT) RIC. The non-RT RIC is responsible for long-term policy generation, optimization, and offline model training, while the near-RT RIC hosts xApps that execute AI-driven control decisions under strict latency constraints.

Figure 1 illustrates the end-to-end procedural workflow of the proposed O-RAN-based resource allocation framework,

explicitly showing how telemetry and control decisions flow across the SMO layer, non-RT RIC, near-RT RIC, and distributed RAN nodes.

Step 1: KPI Collection at the RAN Nodes. At the bottom of Fig. 1, user traffic is served by base stations composed of Radio Units (RUs), Distributed Units (DUs), and Centralized Units (CUs). These RAN nodes continuously collect slice-level and cell-level KPIs, including requested PRBs, granted PRBs, downlink buffer occupancy, and achieved throughput.

Step 2: Near-Real-Time Inference via the E2 Interface. The collected KPIs are transmitted through the E2 interface to the near-RT RIC (middle of Fig. 1). Here, xApps utilize pre-trained machine learning models to perform low-latency inference and predict near-future PRB allocation in a slice-aware manner. The inferred allocation decisions are fed back to the RAN nodes via E2, enabling scheduler-aware resource adaptation in near real time.

Step 3: Long-Term Learning and Optimization via the A1 Interface. In parallel, operational data is periodically forwarded through the A1 interface to the non-RT RIC within the SMO layer (top of Fig. 1). At this stage, rApps aggregate historical multi-base-station data and support offline training and policy optimization.

Step 4: Model Update and Continuous Adaptation. Updated models and control policies generated at the non-RT RIC are redistributed to the near-RT RIC via A1. This enables continuous refinement of xApp intelligence across heterogeneous base stations, schedulers, and traffic conditions.

While previous work by Alshagri *et al.* [11] introduced xApp/rApp frameworks for resource orchestration, our study focuses on the learning and inference engine underpinning these control functions. Concretely, we train predictors that map KPI sequences to the next-step PRB allocation target ($grant_{t+1}$), and benchmark classical and deep models under Single-BS, Multi-BS, and Full-Scheduler evaluation settings.

B. Dataset Design and Scaling

The experiments utilize an extended version of the open-access Colo-RAN dataset [9], which contains over 7.4 GB of performance data originally collected from the Colosseum network emulator. The dataset includes seven base stations, 42 users and more than 73 hours of operational logs. It captures the temporal evolution of key metrics such as throughput, buffer occupancy and PRB utilization under slice-specific scheduling policies. Each record contains 31 KPIs and is provided as part of a structured directory reflecting schedulers, traffic scenarios, experiment repetitions and base stations.

The dataset is organized hierarchically as shown in Fig. 2, where each scheduler contains 28 traffic scenarios, each scenario includes 5 experimental repetitions and each experiment consists of measurements from 7 base stations, with separate BS-level, UE-level and slice context metric files.

From the full corpus of approximately 35 million rows, we adopt a balanced sampling strategy selecting 6 representative traffic scenarios per scheduler and 3 repetitions per scenario,

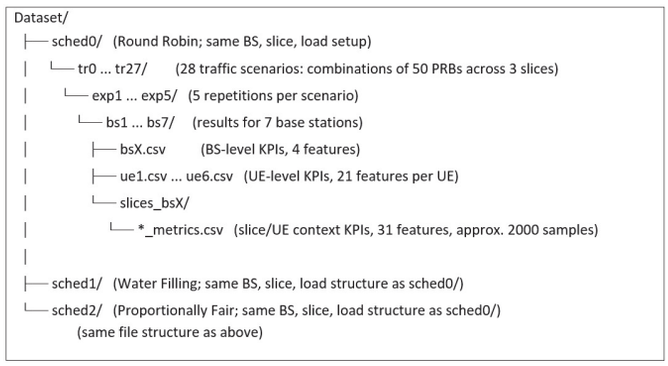


Fig. 2. Dataset hierarchy

producing a stratified 600k row subset. This sampling preserves the distribution across Base Station IDs, schedulers and traffic scenarios, enabling fair evaluation of cross-BS generalization and scheduler-specific behaviors without temporal leakage.

From the full corpus (≈35M rows), we use a **stratified 600k row subset** preserving the distribution across schedulers, traffic scenarios, Base Station IDs and slices to study Cross-BS generalization and scheduler effects without temporal leakage.

Each record contains 31 KPIs. We focus on the following core features and target used across our benchmarks:

```
[Timestamp, slice_id, sum_requested_prbs,
  sum_granted_prbs, dl_buffer [bytes],
       tx_brate downlink [Mbps]]
```

These capture demand (`sum_requested_prbs`), allocation (`sum_granted_prbs`), backlog pressure (`dl_buffer`), and achieved throughput (`tx_brate downlink`), which are the principal drivers for slice-level resource control in O-RAN.

For reproducibility, data were normalized per base station and stored in separate files. The framework supports modular experimentation by allowing:

- **Cross-BS Station Evaluation:** Training on BS1-BS5 and testing on BS6-BS7 to examine spatial generalization.
- **Temporal Forecasting:** sequential train/test splits to simulate real-time future prediction without temporal overlap.
- **Scheduler-Aware Benchmarking:** Each scheduler (PF, RR, TD) is analyzed individually to observe scheduling bias on slice fairness and throughput.

This design facilitates dataset scalability and encourages reproducibility for future researchers aiming to benchmark new AI or transformer-based architectures (e.g., Time-LLM) on realistic O-RAN traffic dynamics.

The data used in this study are derived from the publicly available Colo-RAN dataset[1]. This ensures full reproducibility and aligns with open research practices in the O-RAN community.

[1]Dataset Link: https://github.com/winelab/colosseum-oran-coloran-dataset

C. Evaluation Metrics

A comprehensive evaluation strategy was adopted to measure accuracy, fairness and deployability of machine learning and deep learning models. The adopted metrics are categorized into four major groups. The following are the detailed calculations for these metrics [10]:

a) Regression Metrics:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{1}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{2}$$

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}, \tag{3}$$

$$\text{NRMSE}_{\text{range}} = \frac{\text{RMSE}}{y_{\max} - y_{\min}}, \qquad \text{NRMSE}_{std} = \frac{\text{RMSE}}{\sigma_y}. \tag{4}$$

b) Classification Metrics: (High/low demand states via a train-set threshold.)

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{5}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{6}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{7}$$

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{8}$$

c) Fairness (SLA Alignment): We quantify equity of predicted PRB allocations across slices using Jain's Fairness Index:

$$J = \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n \sum_{i=1}^{n} x_i^2},$$

where x_i is the predicted PRB for slice i; higher J indicates more equitable allocation.

d) Efficiency: We track wall-clock training time, per sample inference latency and model size to assess deployability in near-real-time xApp settings (reported in the results section).

IV. Results and Discussion

This section presents an evaluation of multiple AI models for dynamic resource allocation in an Open RAN framework across three settings: (i) Single-BS, (ii) Multi-BS (cross-cell generalization) and (iii) Full-Scheduler large-scale configuration. Classical machine learning models: Decision Tree (DT), Random Forest (RF), Support Vector Regression (SVR), and K-nearest neighbors (KNN) along with deep learning models such as ANN, LSTM and CNN1D were benchmarked under diverse traffic and scheduler conditions. Table I summarizes the key hyperparameters and evaluation metrics.

Each model is evaluated using RMSE, MAE, and R^2 for regression accuracy, as well as Accuracy, Precision, Recall and F1-score for binary allocation. Training time is also reported

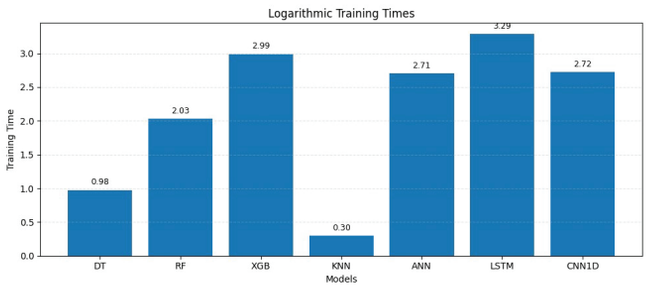


Fig. 3. The Training Time (logarithmic) of ANN, CNN, LSTM, XGB, DT, RF and KNN models.

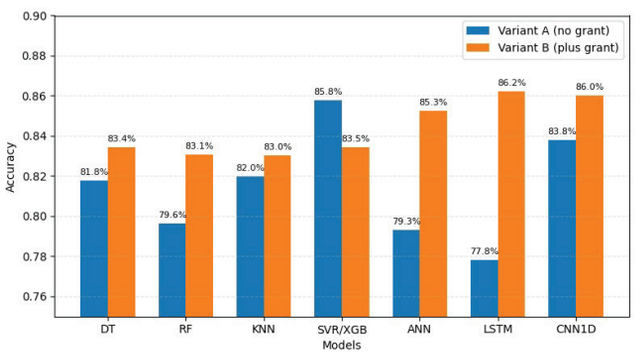


Fig. 4. Accuracy comparison across models for Variant A (no_grant) vs. Variant B (plus_grant).

to capture computational efficiency. Classical models such as DT and KNN train the fastest, making them attractive for lightweight or edge-side deployment, while ANN and CNN1D require moderate computation and LSTM incurs the highest cost due to recurrent dependencies (Fig. 3).

Two feature configurations are analyzed: **Variant A (no grant)**, which removes contemporaneous grant/throughput values to avoid causal leakage and **Variant B (plus grant)**, which incorporates these features to provide short-term scheduler context. As shown in Fig. 4, Variant B consistently improves accuracy and stability across all models.

Prediction stability is assessed using the $\text{NRMSE}_{\text{std}}$ metric, which reflects variance-normalized error across the test set and functions as an analog to confidence-interval reporting. Across Single-BS, Multi-BS and Full-Scheduler settings, $\text{NRMSE}_{\text{std}}$ remains low (typically < 0.21), confirming that results are not influenced by outliers or atypical traffic spikes. Jain's Fairness Index is also computed where applicable. Fairness is moderate in the Single-BS and Multiple-BS settings (≈ 0.57 – 0.65) and highest in the Full-Scheduler scenario (≈ 0.89 – 0.90), indicating more balanced PRB distribution under heterogeneous scheduling. These results collectively highlight the trade-offs between model complexity, accuracy and computational efficiency for real-time O-RAN resource allocation.

A. Single-BS Results

The Single-BS setup (BS1) includes three slices with a strict temporal split (train: 4,533; test: 1,121). As shown in Table

TABLE I
Simulation parameters of AI models

Parameters	Values
Learning Rate	ANN = 1e-3, LSTM = 3e-4, CNN1D = 1e-3
Optimizer	Adam (weight_decay = 1e-5)
Loss Function	Mean Squared Error (MSE)
Training / Testing Data	600,000 Records (Total)
Dataset Source	ColO-RAN (7 BS, 42 UEs)
CPU	Intel i7-14700 @2.1GHz
RAM	16.0 GB
GPU (DL Models)	NVIDIA Tesla T4 (16 GB)
Decision Tree Criterion	Gini Index (default sklearn)
Random State (DT/RF)	42
Max Depth (DT)	8
Min Samples Leaf	2
Min Samples Split	5
No. of Estimators (RF)	200
C of SVM / SVR	10.0
Kernel Type (SVM / SVR)	Radial Basis Function (RBF)
Max Iterations (SVM)	10,000
Dual (SVM)	False
Scaler Used	StandardScaler
No. of Neighbors (KNN)	9
Batch Size (DL Models)	512
Epochs (DL Models)	ANN = 10, LSTM = 30, CNN1D = 10
Activation Function	ReLU (hidden layers), Linear (output)
LSTM Units	128
Dropout (LSTM)	0.2
Validation Split	0.2
Performance Metrics	RMSE, MAE, R^2, Accuracy, Precision, F1

TABLE II
SINGLE-BS (BS1) RESULTS USING GRANT-BASED FEATURE INPUTS. METRICS: STANDARDIZED NRMSE ↓, R^2 ↑, ACCURACY ↑, F1 ↑, AND TRAINING TIME (S) ↓.

Model	RMSE	R^2	Acc.	F1	Train(s)
Decision Tree	0.196	0.962	0.851	0.855	0.012
Random Forest	0.198	0.961	0.841	0.845	2.900
SVR (RBF)	0.206	0.958	0.834	0.842	0.810
KNN (k=9, dist)	0.196	0.962	0.839	0.841	0.003
ANN	0.191	0.964	0.838	0.841	0.604
LSTM	0.192	0.963	0.829	0.813	1.646
CNN1D	0.189	0.964	0.850	0.860	0.623

II, deep models outperform classical baselines, with CNN1D achieving the lowest standardized error (NRMSE = 0.189) and highest F1 (0.860). ANN offers similar performance with substantially lower cost. Among classical models, DT achieves strong accuracy (0.851) with negligible training time (0.012s), while KNN shows similar accuracy (0.839). RF and SVR incur higher training overhead but provide only marginal improvements.

B. Multi-BS (Cross-cell Generalization)

To evaluate generalization, models are trained on BS1-BS5 and tested on unseen BS6-BS7 (train: 28,440; test: 11,385). Table III shows all models benefiting from the inclusion of grant-based features. Among classical models, SVR achieves the best performance (RMSE 0.100, R^2 0.971). RF also performs consistently with RMSE $\approx$ 0.113. DT remains lightweight but shows reduced generalization capacity.

TABLE III
MULTI-BS (CROSS-CELL) RESULTS USING GRANT-BASED FEATURE INPUTS. METRICS: STANDARDIZED NRMSE↓, R^2↑, ACCURACY↑, F1↑, AND TRAINING TIME (S)↓.

Model	RMSE	R^2	Acc.	F1	Train(s)
Decision Tree	0.107	0.965	0.782	0.796	0.081
Random Forest	0.113	0.962	0.790	0.800	18.450
SVR (RBF)	0.100	0.970	0.846	0.836	30.871
KNN (k=9, dist)	0.113	0.961	0.785	0.795	0.016
ANN	0.090	0.975	0.873	0.858	3.944
LSTM	0.093	0.974	0.862	0.855	11.815
CNN1D	0.096	0.972	0.876	0.856	3.363

Deep-learning models ANN, LSTM, and CNN1D-outperform the classical baselines. ANN achieves the lowest RMSE (0.090) and highest R^2 (0.975), while LSTM follows closely. CNN1D also performs robustly (RMSE 0.096), reflecting strong adaptability to heterogeneous base stations and scheduler variations.

C. Full-Scheduler, Large-Scale Results

The Full-Scheduler experiment uses all traffic patterns, Schedulers, and experiment configurations (train: 433,264; test: 175,591). Table IV shows that CNN1D achieves the best overall performance with RMSE 0.085, R^2 0.992, and F1 = 0.849. ANN performs comparably (RMSE 0.086) with lower computational overhead, while LSTM maintains high accuracy but incurs the highest training time.

TABLE IV
FULL-SCHEDULER, LARGE-SCALE RESULTS (ACROSS ALL SCHEDULERS, TRAFFIC TYPES, AND EXPERIMENTS). METRICS: STANDARDIZED NRMSE_{std}↓, R^2↑, ACCURACY↑, F1↑, AND TRAINING TIME (S)↓.

Model	RMSE_{std}	R^2	Acc.	F1	Train(s)
Decision Tree	0.107	0.988	0.834	0.829	0.945
Random Forest (GPU)	0.100	0.989	0.831	0.827	10.789
XGBoost (CUDA)	0.109	0.987	0.835	0.831	97.739
KNN (GPU)	0.110	0.987	0.830	0.825	0.004
ANN (GPU)	0.086	0.992	0.853	0.847	50.733
LSTM (GPU)	0.092	0.991	0.862	0.847	193.269
CNN1D (GPU)	**0.085**	**0.992**	**0.860**	**0.849**	52.666

Among classical models, Random Forest provides the strongest performance (RMSE 0.100, R^2 0.989). DT remains the fastest model, suitable for edge deployment, though with moderate accuracy. KNN and XGBoost show higher prediction variability under large-scale diversity.

Overall, deep learning especially CNN1D generalizes best under large-scale and heterogeneous scheduling conditions, while classical models remain strong candidates when interpretability and minimal compute cost are required.

D. Summary Across Settings

Table V consolidates the best performing models across all three settings. In *Single-BS*, CNN1D provides the lowest error (NRMSE ≈ 0.189), while DT provides the best classical performance. In *Multi-BS*, ANN achieves the lowest deep-model error (NRMSE ≈ 0.090) and SVR leads among classical models. In the Full-Scheduler scenario, CNN1D provides the lowest error (NRMSE $\approx$ 0.085), while RF offers the

TABLE V
SUMMARY OF MODEL PERFORMANCE ACROSS SCENARIOS. EACH MODEL SHOWS NRMSE↓ AND ACCURACY↑.

Scenarios	Classical Machine Learning Models								Deep Learning Models					
	DT		RF		SVR/XGB		KNN		ANN		LSTM		CNN1D	
	RMSE	Acc.	RMSE	Acc.	RMSE	Acc.	RMSE	Acc.	RMSE	Acc.	RMSE	Acc.	RMSE	Acc.
Single-BS	**0.196**	**0.851**	0.198	0.841	0.206	0.834	0.196	0.839	0.191	0.838	0.192	0.829	**0.189**	**0.850**
Multi-BS	0.107	0.782	0.113	0.790	**0.100**	**0.846**	0.113	0.785	**0.090**	0.873	0.093	0.862	0.096	**0.876**
Full-Scheduler	0.107	0.834	**0.100**	0.831	0.109	**0.835**	0.110	0.830	0.086	0.853	0.092	**0.862**	**0.085**	0.860

most stable classical baseline. Overall, deep models provide the strongest accuracy-efficiency trade-off, whereas RF and XGBoost remain reliable classical baselines for real-time scheduler-aware allocation in O-RAN.

V. CONCLUSION

This work conducted a large-scale benchmarking study of classical and deep learning models for PRB allocation in O-RAN, evaluating seven algorithms across Single-BS, Multi-BS, and Full-Scheduler configurations. The results show that deep models, particularly CNN1D, achieve superior accuracy and stability under heterogeneous base stations and schedulers, while lightweight classical models such as Decision Trees and Random Forests remain attractive for fast, low-overhead inference at the edge. Incorporating grant-based contextual features further improves predictive performance without degrading slice-level fairness. Overall, the study highlights practical trade-offs between accuracy, computational cost, and deployability, providing concrete guidance for selecting AI models for near-real-time xApps in disaggregated O-RAN systems. Future work will explore Time-LLM-based temporal models to capture long-range dependencies and further enhance cross-cell generalization.

VI. ACKNOWLEDGMENT

This work was supported in part by the BK21 FOUR (Fostering Outstanding Universities for Research) program funded by the Ministry of Education (MOE) and the National Research Foundation (NRF) of Korea, and in part by the NRF grant funded by the Ministry of Science and ICT (MSIT) (No. 2022R1F1A1073976).

REFERENCES

[1] M. Dhandapani and V. Vetriselvi, "DRL-based customised resource allocation for sub-slices in 6G network slicing," *IET Networks*, 2024.

[2] Y. Cai, P. Cheng, Z. Chen, M. Ding, B. Vucetic, and Y. Li, "Deep reinforcement learning for online resource allocation in network slicing," *IEEE Transactions on Mobile Computing*, 2024.

[3] P. Valcarce, P. Kela, S. Mandelli, and H. Viswanathan, "The role of AI in 6G MAC," in *Proc. 2024 European Conference on Networks and Communications (EuCNC) 6G Summit*, 2024.

[4] F. Lotfi, F. Afghah, and J. Ashdown, "Attention-based Open RAN slice management using deep reinforcement learning," *arXiv preprint arXiv:2306.09490*, 2023.

[5] F. Lotfi and F. Afghah, "Open RAN LSTM traffic prediction and slice management using deep reinforcement learning," in *Proc. 2023 57th Annual Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, USA, 2023.

[6] S.-P. Yeh, S. Bhattacharya, R. Sharma, and H. Moustafa, "Deep learning for intelligent and automated network slicing in 5G Open RAN deployment," *IEEE Open Journal of Communications Society*, 2024.

[7] J. Shen, W. Zhang, H. Liu, and Y. Chen, "AI-assisted network-slicing-based next-generation wireless networks," in *Proc. IEEE International Conference on Communications (ICC)*, 2023.

[8] D. Bega, M. Gramaglia, A. Garcia-Saavedra, M. Fiore, A. Banchs, and X. Costa-Perez, "Network slicing meets artificial intelligence: An AI-based framework for slice management," *IEEE Communications Magazine*, 2020.

[9] M. Polese, L. Bonati, S. D'Oro, S. Basagni, and T. Melodia, "ColORAN: Developing machine learning-based xApps for Open RAN closed-loop control on programmable experimental platforms," *IEEE Transactions on Mobile Computing*, pp. 1–14, July 2022.

[10] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[11] A. Alshagri, N. A. Jaafar, A. F. Alsuhaim, H. Alnukhaylan, and A. F. Albeechi, "Evaluation of AI models for intelligent network slicing in Open RAN: A comprehensive benchmarking approach," in *Proc. 2024 International Conference on Intelligent Computing and Next Generation Networks (ICNGN)*, IEEE, 2024, doi:10.1109/ICNGN63705.2024.10871261.

Dynamic Optimization of Input Context Length for Reducing Computational Cost and Power Consumption in Decision Transformers

Taehyun Han[1], Janghwan Kim[2], Chanyeol Yang[3], Jihoon Kong[4], Sangho Lee[5], Chaeyun Seo[6], R. Young Chul Kim[7*]
[1,2,4,6,7*]*Dept. of Software and Communication Engineering, Hongik University, Sejong, Republic of Korea*
[3]*Daon I&C, Gwangmyeong, Republic of Korea*
[4]*Rastech, Daejeon, Republic of Korea*
Emails: [1]*taehyun3172@g.hongik.ac.kr*, [2]*lentoconstante@hongik.ac.kr*, [3]*yangcy2000@da-on.com*,
[4]*go400@gmail.com*, [5]*project@rastech.co.kr*, [6]*chaeyun@hongik.ac.kr*, [7*]*bob@hongik.ac.kr*

***Abstract*— Traditional machine learning models involve two main types of time complexity: training time and inference time. Most previous studies have focused primarily on reducing training time. However, this paper argues that inference time should also be considered an important part of overall complexity. In the context of offline reinforcement learning, the Decision Transformer successfully reframed the problem as sequence modeling and achieved strong performance. Nevertheless, the standard Decision Transformer architecture uses a fixed context length for input sequences at every time step, regardless of the complexity of the current state. This leads to unnecessary computational waste and higher power consumption. To address this issue, we propose a dynamic input length optimization mechanism. Our approach introduces a Multi-Layer Perceptron Selector that predicts the minimum required context length in real time based on the agent's current state. Additionally, to prevent performance degradation caused by excessive information reduction, we ensure control stability using a cumulative probability-based mechanism. The proposed method is expected to maintain the control performance of the Decision Transformer while significantly reducing computational costs and power consumption.**

Keywords*— *Decision Transformer, Sequence Modeling, Offline Reinforcement Learning, Input Length Optimization, Computational Efficiency

I. INTRODUCTION

Reinforcement learning has developed into offline reinforcement learning, which uses fixed datasets to solve the high cost and risk of exploring the real world [1]. In this context, the Decision Transformer (DT) appeared [2]. It rethinks reinforcement learning as a sequence modeling problem. By using past trajectory information, it predicts future actions in an autoregressive way. This has shown better results than traditional methods that rely on value functions.

However, the original DT structure has a fixed input context length, no matter how complex the current state is. This is a structural limit. The attention operation in Transformers increases in computation proportionally to the square of the input length [3]. Therefore, keeping the maximum length even when a long sequence is not needed causes serious inefficiency in computation. This problem becomes bigger as the input sequence in DT gets longer.

This paper proposes a mechanism to improve this inefficiency by dynamically optimizing the input context length of DT. The proposed method introduces a lightweight Multi-Layer Perceptron (MLP) before the DT input. This MLP outputs an appropriate input length for the current state and dynamically trims the sequence. To prevent wrong predictions from the MLP model, we add a safety mechanism. It uses the cumulative sum of the output probability distribution to ensure a stable length is finally chosen. Through this, we aim to maximize inference efficiency by removing unnecessary computations while keeping the performance of the DT model.

In chapter 2, we remark about related works. Afterwards, we will remark proposed mechanism in chapter 3. And then we conclude this paper with future works and expectations.

II. RELATED WORKS

A. Research on Variations Decision Transformers

The Decision Transformer successfully reframed reinforcement learning as a sequence modeling problem [2]. Later, variant models such as the Elastic Decision Transformer were introduced to improve robustness across trajectories of different lengths [4]. However, these mainly focused on enhancing performance by stitching together incomplete trajectories and did not directly address computational efficiency during inference. In contrast, this study differentiates itself from prior work by concentrating on reducing unnecessary computational costs in the inference stage, rather than on improving overall performance.

B. Research on Transformer Efficiency

Research aimed at reducing the computational cost of Transformer models in Reinforcement Learning has been actively conducted [5]. In particular, in the vision domain, techniques such as DynamicViT have been proposed, which dynamically prune unimportant tokens within images to accelerate inference speed [6]. This extends the concept of such "adaptive computation" to the field of continuous control in offline reinforcement learning, proposing a novel mechanism that dynamically adjusts the length of the input sequence itself based on the agent's state.

III. PROPOSED METHOD

We propose a dynamic input length optimization mechanism that adaptively adjusts the input context length of the Decision Transformer depending on the situation. The overall architecture consists of two parts: an offline phase that builds the training data, and an online phase that safely determines the context length during actual inference. Using our proposed method, we extract the optimal input context length for each state to minimize the performance loss of the DT. We then train an MLP selector using a dataset constructed from these extracted states and context lengths. Finally, the determined context length is fed into the DT as input.

A. Definition of Optimal Context Length

First, we need to define the minimum context length that can maintain control performance for each state. To do this, we use a pre-trained Decision Transformer (Frozen Decision Transformer). We compare the baseline action vector(a_{base}) produced when the full sequence is given as input with the action vector produced when the context length is reduced to K_{opt} (a_{masked}).

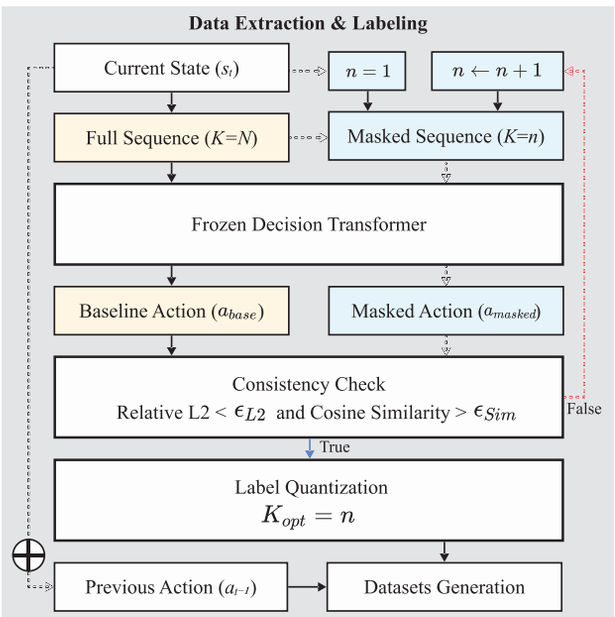


Fig. 1. Data Extraction and Labeling Process for Training MLP Selector

Fig. 1 illustrates the data extraction and labeling process. To evaluate the consistency between two vectors, we apply the following criteria based on user-defined thresholds for the L2 error ϵ_{L2} and cosine similarity ϵ_{Sim}:

$$\frac{||a_{base} - a_{masked}||_2}{||a_{base}||_2} < \epsilon_{L2} \;\&\; sim(a_{base}, a_{masked}) > \epsilon_{Sim} \quad (1)$$

The minimum context length n that satisfies Equation (1) is labeled as the optimal length K_{opt} for the given state. To facilitate the training of the MLP model, the input vector mapped to this optimal length comprises both the current state and the previous action vector.

B. Lightweight MLP Selector and Safety Mechanism

In the real-time inference phase, a lightweight MLP Selector positioned preceding the Transformer model is activated. This selector takes the current state s_t and the previous action a_{t-1} as inputs to predict the probability $P(k), k \in K$ for each context length class. At this stage, simply selecting the length with the highest probability (argmax) can be risky, as it fails to account for the model's uncertainty.

Therefore, this paper proposes a safety mechanism based on cumulative probability. This method iterates from the shortest length $k = 1$, accumulating the predicted probabilities, and selects the length when the sum exceeds a predefined safety threshold (θ).

For instance, assume a total context length of $K = 20$ and that the output probability distribution of the MLP Selector is uniform across all k (i.e., $P(k = 1) \approx \cdots \approx P(k = 20) \approx 0.05$). Using a simple argmax approach would result in an unstable output driven by negligible differences in probability. In contrast, the proposed approach ensures stability by calculating the cumulative probability as follows:

$$\sum_{i=1}^{k} P(i) \geq \theta \rightarrow 0.05 \times k \geq \theta \quad (2)$$

In other words, higher uncertainty in the output probability leads the algorithm to identify a longer sequence as the optimal length. This serves as a core principle for ensuring system safety, encouraging the model to conservatively reference as much past information as possible when the Selector lacks confidence.

IV. Conclusion

In this paper, we propose a dynamic input length optimization mechanism to address the structural inefficiency caused by the fixed input length of Decision Transformer for offline reinforcement learning. The proposed mechanism combines a lightweight MLP Selector with a cumulative probability-based safeguard to eliminate unnecessary computations while maintaining control performance. While existing DTs always process sequences of fixed length K, resulting in a computational complexity of $O(K^2)$, our method processes only the optimized length K_{opt}, significantly reducing the computational load to $O(K_{opt}^2)$. In general environments such as robot control, agents spend most of their time in stable states like driving or idling, where short sequences are sufficient for control; therefore, significant computational cost savings can be achieved on an overall episode average. Furthermore, the safeguard algorithm optimizes the balance between efficiency and control stability by automatically expanding the input length in high-uncertainty intervals. We expect this study to serve as a key lightweight methodology for applying large-scale Transformer models to resource-constrained systems, such as mobile robots, in the future.

References

[1] S. Levine, A. Kumar, G. Tucker, and J. Fu, "Offline reinforcement learning: Tutorial, review, and perspectives on open problems," arXiv preprint arXiv:2005.01643, 2020.

[2] L. Chen, K. Lu, A. Rajeswaran, K. Lee, A. Grover, M. Laskin, P. Abbeel, A. Srinivas, and I. Mordatch, "Decision transformer: Reinforcement learning via sequence modeling," in *Proc. of the Advances in Neural Information Processing Systems (NeurIPS)*, 2021, pp. 15084–15097.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. of the Advances in Neural Information Processing Systems (NIPS)*, Long Beach, CA, 2017, pp. 5998–6008.

[4] Y. Wu, X. Wang, and M. Hamaya, "Elastic decision transformer," in Proc. of the Advances in Neural Information Processing Systems (NeurIPS), New Orleans, LA, 2023, pp. 18532–18550.

[5] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, "Efficient transformers: A survey," ACM Comput. Surv., Vol. 55, No. 6, pp. 109:1–109:28, 2023.

[6] Y. Rao, W. Zhao, B. Liu, J. Lu, J. Zhou, and C. J. Hsieh, "DynamicViT: Efficient vision transformers with dynamic token sparsification," in Proc. of the Advances in Neural Information Processing Systems (NeurIPS), 2021, pp. 13937–1394

Sustainable Loops: Mapping the Scientific Landscape of Digital Circular Economy Solutions in Textile Waste

1st Edison M. Esberto *, 2nd Mideth B. Abisado †
* *College of Computing and Information Technologies (CCIT), National University, Philippines, Manila, Philippines*
† *College of Computing and Information Technologies (CCIT), National University, Philippines, Manila, Philippines*
Emails: *emesberto@nu-laguna.edu.ph, mbabisado@national-u.edu.ph*

***Abstract*— Digital technologies are increasingly recognized as key enablers of circular economy transitions and the achievement of the United Nations Sustainable Development Goals (SDGs). This study systematically maps the scientific landscape at the intersection of circular economy, digitalization, and Industry 4.0 using bibliometric and network analysis. A total of 92 peer-reviewed journal articles published between 2014 and 2024 were analyzed following a structured screening and filtering process. The results show a marked increase in publication output after 2019, indicating accelerating academic and policy interest in digitally enabled circular systems. Thematic mapping identifies three dominant research clusters, focusing on Industry 4.0 technologies, sustainability transitions, and industrial circular applications. In these areas, digital tools are primarily viewed as mechanisms for enhancing resource efficiency, promoting closed-loop production, and facilitating data-driven decision-making. These themes align most strongly with SDG 9 (Industry, Innovation and Infrastructure), SDG 12 (Responsible Consumption and Production), and SDG 13 (Climate Action). Geographical analysis reveals that research output is concentrated in a limited number of technologically advanced countries, while contributions from developing economies remain comparatively low. Citation network results further indicate a strong emphasis on technological optimization, with digital governance, social inclusion, and policy coordination receiving limited scholarly attention. Overall, the findings demonstrate that while Industry 4.0 dominates circular economy research, achieving SDG-oriented impact requires stronger governance frameworks and more inclusive research and policy approaches.**

Keywords—Circular economy, Textile waste management, Digital circular economy, Industry 4.0, Sustainable Development Goals (SDGs), Bibliometric analysis, Digital governance, Resource efficiency

I. Introduction

The modern world is characterized by unprecedented levels of consumption, driven by population growth, urbanization, and globalization. While these developments have elevated living standards, they have also intensified the planet's ecological burden. Among the growing forms of environmental degradation, textile waste has emerged as a particularly pressing concern—often overlooked in mainstream sustainability policies.

The global textile industry operates predominantly on a linear economic model, characterized by resource depletion and significant waste generation, which is increasingly recognized as unsustainable given finite global resources and mounting environmental concerns (Ütebay et al., 2020). This traditional "take-make-dispose" approach leads to substantial quantities of textile waste that are neither recycled nor reused, alongside the extensive use of toxic dyes and chemicals, contributing to widespread environmental degradation such as water body depletion and ecosystem diversity reduction (Schnatmann et al., 2023). This study sees imperative to transition from this linear paradigm to a circular economy framework has gained considerable traction as a viable pathway to mitigate these adverse environmental impacts by decoupling economic activity from finite resource consumption and designing waste out of the system (Körlü & Körlü, 2020).

This model emphasizes resource efficiency and aims to eliminate waste by promoting practices such as reducing, reusing, and recycling materials throughout the entire product lifecycle (Edirisinghe et al., 2024; Rahaman & Pranta, 2025). In response to the environmental challenges posed by the linear economy, the circular economy proposes new business models that optimize resource utilization and limit environmental impacts throughout manufacturing and consumption phases (Das & Dwivedi, 2024). This involves designing products for longevity, ease of disassembly, and recyclability, thereby keeping materials in use for as long as possible. This paradigm shift from a linear to a circular system is essential, as the current global economy is less than 10% circular, indicating a significant waste of valuable resources that could otherwise be recovered and repurposed. A concerted effort is required to develop new technologies that improve the recovery of valuable materials from textile waste streams, facilitating their reuse for consumers (Biyada & Urbonavičius, 2025). Implementing such a model in the textile industry is crucial for sustainability and minimizing environmental impacts, especially considering the vast quantities of textile waste generated annually. This shift necessitates a reevaluation of production and consumption patterns to mitigate environmental damage and conserve natural resources within the Technosphere (Riba et al., 2020). The fast fashion phenomenon, characterized by rapid design turnover and increased production volumes, exacerbates these issues, necessitating a fundamental transformation in both consumer behavior and corporate practices to embrace circularity (Ramírez-Escamilla et al., 2024).

Worldwide, the textile industry is the second-largest industrial polluter, generating an estimated 92 million tons of waste annually and contributing to 10% of carbon emissions and 20% of global wastewater production (Voukkali et al., 2023; Gue et al., 2022). Projections suggest this figure could rise to 134 million tons by 2030 (Poon et al., 2024). In the Philippines, the problem is compounded by widespread fast fashion imports and a weakly structured waste management

system. Only 10–20% of textile waste is formally reused or recycled, while the rest accumulates in landfills or flows into informal markets like “ukay-ukay” (Figueroa et al., 2021).

Although global strategies such as Extended Producer Responsibility (EPR), sustainable fashion campaigns, and national recycling frameworks are gaining momentum, their implementation is uneven. In the Philippines, for instance, the Ecological Solid Waste Management Act (RA 9003) does not explicitly address textile waste. Policy implementation remains fragmented, with informal markets operating in silos and community participation generally low (Juanga-Labayen et al., 2022).

The pressing need for systemic transformation calls for a paradigm shift away from the “take-make-dispose” linear model. The Circular Economy (CE) model offers a viable alternative, emphasizing reuse, upcycling, regeneration, and responsible consumption. CE strategies applied to textiles—such as repair, second-hand trade, and recycling—can significantly mitigate environmental impacts and resource depletion (Smol et al., 2020).

Digital technologies offer further promise by enabling the operationalization of circular models. Platforms such as blockchain-enabled E-barter systems, AI-powered sorting tools, and digital tracking systems have the potential to democratize sustainability and decentralize textile flows (Hussain et al., 2021). These innovations can advance the Sustainable Development Goals (SDGs), particularly SDG 12 (Responsible Consumption and Production) and SDG 13 (Climate Action), by reducing production-related emissions and promoting textile reuse.

However, a country-level co-authorship analysis and publication trends show a significant concentration of research outputs in Europe and high-income countries, while Southeast Asia—including the Philippines—remains severely underrepresented. This disparity presents a research opportunity and an urgent policy gap, as digital circular solutions that succeed in high-income contexts may not easily translate into regions with informal economies and limited infrastructure.

Therefore, this study conducts a bibliometric analysis of Scopus-indexed literature to map the scientific landscape of digital CE solutions for textile waste. The objective is to identify knowledge clusters, geographic gaps, and underexplored innovations. This research contributes a timely and context-aware perspective, highlighting the need for localized and inclusive CE strategies tailored to the unique challenges and opportunities of countries like the Philippines.

II. Related Studies

The convergence of circular economy (CE) frameworks, digital innovation, and textile waste management has become a focal point in sustainability research. As environmental degradation intensifies, scholars have increasingly turned to digital circular models as viable solutions for resource efficiency and pollution reduction. This study builds upon a growing body of literature that explores how technological tools—from AI to blockchain—can operationalize circularity in textiles.

The transition towards a circular textile economy involves a comprehensive shift from the traditional "take-make-dispose" model to a regenerative framework, integrating strategies like recycling, upcycling, and promoting product durability to significantly reduce virgin resource extraction and mitigate pollution (Alves et al., 2024; Silobrit & Jureviciene, 2023). This paradigm emphasizes maintaining products and materials at their highest utility and value for as long as possible, fundamentally redesigning industrial processes to eliminate waste and pollution (Schumacher & Forster, 2022). Indeed, the circular economy in textiles necessitates a transformative lifecycle that prioritizes design for durability, extended product life, and enhanced reuse or recycling capabilities (Din et al., 2025). This approach effectively converts textile waste into valuable raw material resources, thereby minimizing environmental impact and promoting sustainable production and consumption patterns within the industry (Dziubaniuk et al., 2023). This strategic shift encourages innovations in material science and processing technologies, fostering closed-loop systems where textile materials continuously cycle through usage and recovery phases without significant degradation (Leal Filho et al., 2024).

Studies on textile waste within the circular economy framework in the Philippines remain limited (Balanay & Halog, 2019). Balanay and Halog examine the applicability of CE tools, including life cycle assessment and material flow analysis, to the Philippine textile industry, highlighting their potential for textile waste reduction amid fragmented supply chains and weak regulatory capacity. At the policy level, Bueta et al. identify governance barriers affecting circular waste systems nationwide (Bueta et al., 2024). However, empirical, textile-specific analyses of waste flows and social impacts remain underexplored and is being provided in this paper.

Within the context of sustainable waste management, automated and technology-driven segregation systems play a critical role in enabling circular economy practices by ensuring materials are recovered and reintegrated into productive use. Prior studies highlight the effectiveness of deep learning–based object detection, particularly YOLO models, in improving the accuracy and real-time performance of garbage classification, which is essential for efficient recycling workflows (Zhang & Blancaflor, 2025). Complementing this, digital platforms that incentivize proper waste segregation—such as reward-based web applications—have been shown to motivate user participation and strengthen recycling behaviors at the community level (E. B. Blancaflor et al., 2022). Building on these foundations, the Sortino system operationalizes circular economy principles by combining automated waste classification with user engagement mechanisms, thereby supporting closed-loop material flows and reducing reliance on landfills (E. Blancaflor et al., 2025).

The environmental toll of the fashion industry is well documented. Voukkali et al. (2023) report that the sector contributes significantly to global carbon emissions and water consumption. Smol et al. (2020) argued that circular textile systems, such as recycling and reuse, can mitigate these effects if effectively scaled and integrated. Yet, despite these potential gains, implementation in the Global South remains sluggish. In Southeast Asia, especially in the Philippines, reuse and recycling systems are largely informal, with little integration into national policy (Figueroa et al., 2021; Juanga-Labayen et al., 2022).

A bibliometric study by Awan and Sroufe (2022) revealed a strong geographic concentration of research outputs and

collaborations in Europe and North America, particularly in the domain of digital sustainability and CE innovation. This finding is reinforced by the present study's analysis of Scopus-indexed literature, which shows that high-income countries such as the United Kingdom, Italy, and China dominate publication output and collaborative networks. The underrepresentation of Southeast Asian countries suggests a need to expand research and policy engagement in these regions, especially given their unique challenges related to informality and infrastructure.

Within the Philippine context, Figueroa et al. (2021) highlighted the role of ukay-ukay (secondhand clothing) markets in supporting reuse but noted their exclusion from formal waste governance frameworks. This points to a broader challenge: the need to contextualize CE strategies to suit localized conditions. Integrating digital platforms—such as E-barter systems or mobile recycling apps—into such informal ecosystems could provide a more inclusive and scalable pathway toward sustainable textile management.

Overall, these related studies expose the structural imbalance in CE research and practice, affirming the significance of this paper's focus: mapping the digital circular economy landscape in textile waste and identifying geographic and thematic gaps. In doing so, it offers a foundation for targeted innovation, policymaking, and academic collaboration in underrepresented regions like the Philippines.

III. Methodology

This research employs a bibliometric analysis framework to systematically map the global research landscape surrounding digital circular economy (CE) solutions in textile waste management. Bibliometric analysis is a robust quantitative method that enables the identification of intellectual patterns, influential publications, and collaborative structures in scientific literature. To ensure rigor and transparency, the study adopts the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) as a guiding framework for document selection and filtering.

The methodology is structured into three stages: Data Collection, Bibliometric Analysis and Visualization, and Supplementary Scopus Analytics.

A. Data Collection

Relevant literature was retrieved from the Scopus database using an advanced Boolean search strategy targeting the intersection of textile waste, circular economy, digital solutions, and policy. The search was conducted using the following advanced query string:

```
((TITLE-ABS-KEY(textile AND waste) AND TITLE-ABS-KEY(circular AND economy) AND TITLE-ABS-KEY(policy)) AND PUBYEAR > 2013 AND PUBYEAR < 2025 AND ( LIMIT-TO ( SRCTYPE,"j" ) ) AND ( LIMIT-TO ( EXACTKEYWORD,"Circular Economy" ) OR LIMIT-TO ( EXACTKEYWORD,"Waste Management" ) OR LIMIT-TO ( EXACTKEYWORD,"Textile Waste" ) OR LIMIT-TO ( EXACTKEYWORD,"Textile Recycling" ) OR LIMIT-TO ( EXACTKEYWORD,"Textile Waste Management" ) OR LIMIT-TO ( EXACTKEYWORD,"Extended Producer Responsibility" ) OR LIMIT-TO ( EXACTKEYWORD,"Environmental Policy" ) OR LIMIT-TO ( EXACTKEYWORD,"Circular Waste Management" ) OR LIMIT-TO ( EXACTKEYWORD,"Innovation" ) OR LIMIT-TO ( EXACTKEYWORD,"Laws And Legislation" )) AND ( LIMIT-TO ( DOCTYPE,"ar" ) OR LIMIT-TO ( DOCTYPE,"re" ) OR LIMIT-TO ( DOCTYPE,"cp" )) AND ( LIMIT-TO ( SUBJAREA,"ENVI" ) OR LIMIT-TO ( SUBJAREA,"COMP" ) OR LIMIT-TO ( SUBJAREA,"SOCI" ) OR LIMIT-TO ( SUBJAREA,"DECI" )) AND ( LIMIT-TO ( LANGUAGE,"English" )))
```

Filters were applied to restrict results to peer-reviewed journal articles, reviews, and conference papers published between 2014 and 2024, written in English, and belonging to the subject areas of Environmental Science, Computer Science, Decision Science, and Social Sciences. From an initial set of 93 documents, 44 publications met the inclusion criteria after applying PRISMA filters. These documents formed the final dataset for analysis.

B. Bibliometric Analysis and Visualization

The metadata from the final 44 records—comprising titles, abstracts, authors, affiliations, keywords, and citations—were exported in CSV format and processed using VOSviewer, a tool specialized in constructing and visualizing bibliometric networks. Three types of network maps were generated:

1) Co-authorship networks, which identified profile authors and collaborative linkages;
2) Keyword co-occurrence networks, to reveal dominant themes and emerging research areas;
3) Citation networks, highlighting the most influential publications and their interconnections

Additionally, country-level co-authorship analysis was conducted using "Country" as the unit of analysis, revealing geographic disparities in research contributions.

C. Supplementary Scopus Analytics

To enrich the interpretation, Scopus Analyze Results were reviewed to identify leading countries, domains, and institutions in textile CE research. This added a macro-level context to the bibliometric findings and informed the identification of thematic and geographic gaps.

This mixed bibliometric and visual analytics approach allowed the researchers to uncover not only the structural contours of the research landscape but also underexplored opportunities—such as the integration of digital barter models and policy-based innovations for developing countries. By applying this methodology, the study generates evidence-based insights to guide future research, collaborations, and policymaking in sustainable textile management.

IV. Results and Discussion

A. Co-authorship Networks

The co-authorship network analysis reveals a dense and highly interconnected collaboration structure among a core group of researchers in the domain of digital circular economy and textile waste management. As visualized in Figure 1, prominent authors such as Esposito, Mark, Papamichael, Iliana, Demetriou, Giorgos, and Zorpas, Antonis A. form a tightly knit cluster with high link strength, indicating frequent co-publication and strong academic partnerships. The visualization shows near-complete interlinkages among the 11 key authors, suggesting a collaborative research ecosystem where interdisciplinary inputs—ranging from environmental science to waste policy—are being shared effectively.

This concentration of linkages points to the existence of a cohesive and possibly Europe-centered research group focused on circular economic strategies, often addressing textile waste through policy-embedded solutions. The density of the connections also implies high knowledge exchange and possibly the presence of consortium-led publications. However, the relatively small size of the author network compared to the full dataset suggests that outside this core, collaboration is still fragmented. This highlights an opportunity to foster broader international partnerships, particularly with researchers in Southeast Asia and the Global South.

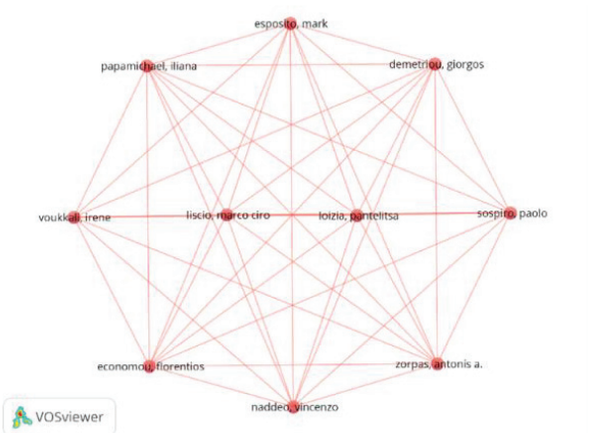


Fig. 1. Co-authorship Network of Researchers in Digital Circular Economy and Textile Waste Management

B. Citation Structures

The citation network map, shown in Figure 2, presents a clustered structure that highlights key authors and influential works published between 2018 and 2024. Central figures such as Solis (2020), Qu (2019), Bukhari (2018), and Alhola (2019) emerge as highly cited contributors, each represented by large nodes and bold text indicating citation weight and link strength.

Solis (2020) is positioned as the most influential work in the network, suggesting foundational contributions to circular economy modeling or sustainability frameworks related to textile or waste sectors. The prominence of Qu (2019) and Alhola (2019) reflects the methodological or policy-relevant nature of their contributions, likely addressing lifecycle analysis, stakeholder engagement, or regulatory mechanisms. More recent contributions from Stanescu (2021) and Dagilienė (2021) are also gaining traction, possibly due to the growing attention on policy integration and digital platforms in circularity.

Interestingly, citation clusters appear to form around publication years, indicating thematic evolution across time. For example, 2019–2021 appears to focus on theoretical and economic modeling, while 2023–2024 entries from authors like Shakib, Herrador, and Emami may reflect emerging technologies or case-based applications in sustainability. The visualization confirms a temporal expansion of research and signals the increasing complexity and diversification of topics within this domain.

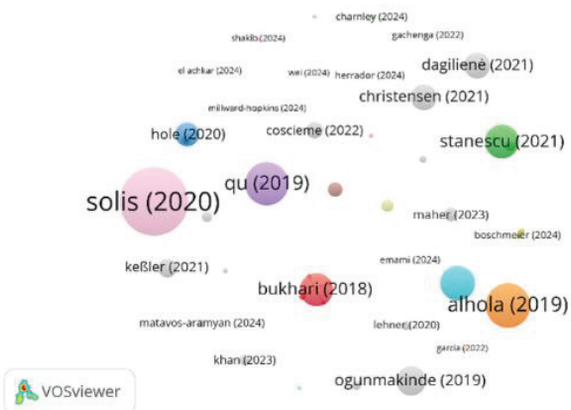


Fig. 2. Citation Network of Influential Authors and Publications (2014-2024)

C. Keyword Co-occurrence Patterns

The keyword co-occurrence network (Figure 3) reveals the thematic breadth of research that intersects with circular economy and textile waste, as well as the central concepts driving scholarly discourse. At the core of the map is the term "circular economy," which is the most dominant keyword, both in size and linkage density. Closely connected to these terms are concepts such as "recycling," "sustainability," and "sustainable development," which collectively underscore the environmental and policy-oriented focus of the field.

Clusters also form around "textile sector," "climate change," "microplastic pollution," and "public policy," indicating the intersectionality of textile waste with broader ecological and governance challenges. Keywords such as "plastic recycling," "cost benefit analysis," "supply chains," and "additives" suggest a growing research interest in material-level analysis and economic feasibility, which are essential for justifying and implementing circular business models.

Notably, the inclusion of "data gaps" and "social learning" reflects the evolving nature of the literature, where behavioral and information access aspects are beginning to influence system-level sustainability discussions. The strong visual clustering of keywords also reveals the multi-disciplinary character of the field, bridging environmental science, economics, materials engineering, and decision science.

Overall, the keyword analysis confirms that digital solutions for textile waste within the circular economy are gaining traction, particularly those addressing climate action, regulatory frameworks, and stakeholder participation. However, niche areas such as public engagement, barter systems, and digital traceability remain underrepresented and may offer promising directions for future research.

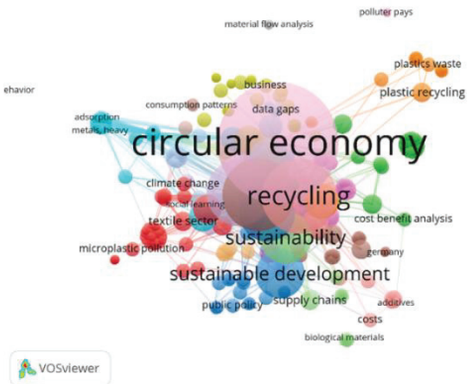


Fig. 3. Keyword Co-occurrence Map Highlighting Thematic Clusters in Circular Economy and Textile Waste Research

D. Country-Level Research Collaboration and Publication Trends

To complement the author-level collaboration insights, a country-level co-authorship analysis was conducted using VOSviewer with "countries" as the unit of analysis. The resulting network visualization (see Figure 4) revealed key international players contributing significantly to textile waste and circular economy literature. Countries such as the United Kingdom, Italy, Australia, Denmark, and China form prominent collaborative clusters, suggesting strong institutional research linkages in these regions. These clusters also reflect shared policy interests and funding investments in environmental sustainability and digital innovation.

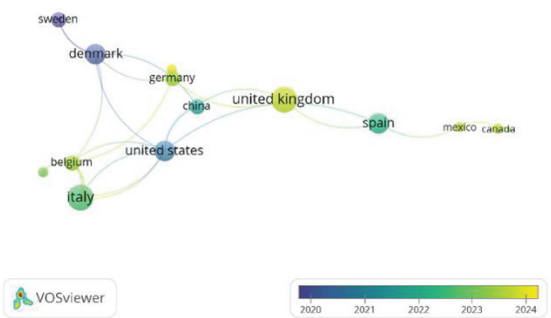


Fig. 4. Country-Level Co-authorship Network on Circular Economy and Textile Waste Research

Supporting this, the Scopus Analyze Results further show that publications in this domain (see Figure 5) are predominantly classified under Environmental Science (48.6%), followed by Social Sciences (16.8%), Computer Science (10.3%), and Decision Sciences (6.5%).

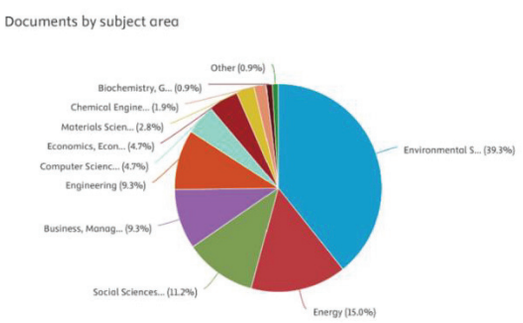


Fig. 5. Publications Domain

The leading countries (see Figure 6) in terms of publication output mirror those found in the co-authorship map, with the United Kingdom, China, and Italy producing the most research outputs related to the circular economy and textile waste. Notably, Southeast Asian countries, including the Philippines, are underrepresented in this global discourse.

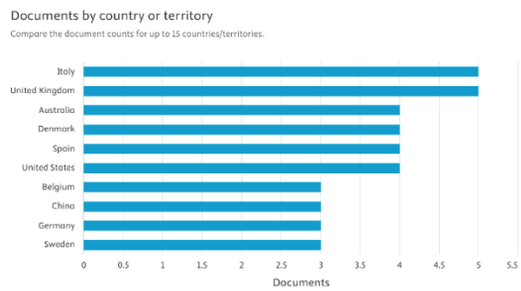


Fig. 6. Leading Countries in terms of Publication Output

This finding highlights a critical research gap and presents an opportunity for Filipino scholars and institutions to establish a stronger presence in the field. Collaborative engagements with leading countries and regional alliances in Asia could not only enhance research visibility but also facilitate the contextualization of global CE frameworks to fit local systems and challenges. Addressing this geographic imbalance in scholarships may be vital to achieving more inclusive and regionally adaptable sustainability solutions.

V. Conclusion and Recommendations

This study presents a comprehensive bibliometric analysis of digital circular economy (CE) research in the context of textile waste, aiming to uncover structural, thematic, and geographic patterns that inform future directions for sustainability interventions. Drawing from 44 Scopus-indexed publications between 2014 and 2024, the analysis revealed a concentrated scholarly ecosystem led by high-income countries, particularly in Europe, with significant gaps in research contributions and collaborations from the Global South, especially Southeast Asia.

The co-authorship network highlighted a cohesive group of influential researchers with strong linkages, mostly within European institutions. Citation analysis underscored the prominence of foundational theoretical works and more recent shifts toward applied digital solutions and policy integration. The keyword co-occurrence map confirmed the dominance of environmental sustainability themes, while also identifying underexplored topics such as social learning, barter systems, and microplastic pollution—suggesting evolving interest in behavioral and localized digital interventions.

Country-level collaboration patterns and Scopus domain analytics further validated the centralization of research in specific regions and disciplines, with Environmental Science, Social Science, and Computer Science emerging as the top research categories. The Philippines and neighboring Southeast Asian countries remain underrepresented in both research outputs and collaborative networks, presenting a compelling case for policy and academic reform.

In light of these findings, several strategic directions are recommended. First, there is a pressing need to foster inclusive and interdisciplinary research environments in underrepresented regions, particularly by empowering Filipino researchers and institutions through cross-border

collaborations and national funding incentives. Strengthening institutional partnerships between local universities and global research leaders can bridge the current disparity in scholarly output and innovation capacity. Second, investment in digital infrastructure must be accelerated to support localized CE platforms. Tools such as blockchain-enabled traceability systems, mobile E-barter platforms, and AI-based textile classification solutions can empower communities and small enterprises to participate meaningfully in sustainable waste practices. Government support in scaling these innovations is essential to their long-term impact. Third, policymakers should revise and expand national waste management frameworks, such as the Ecological Solid Waste Management Act, to explicitly include textile waste and circular practices. Embedding Extended Producer Responsibility (EPR) mechanisms, supported by incentives for circular business models, will be critical for mainstreaming sustainability across the textile value chain. Fourth, community-level engagement must not be overlooked. Digital technologies should be harnessed to facilitate behavioral transformation through educational campaigns, gamified platforms, and participatory recycling programs. These initiatives can increase awareness and participation, particularly in informal or underserved communities. Finally, future research should expand bibliometric methods to track emerging sustainability domains such as circular labor markets, informal barter networks, and digital inclusion in CE transitions. Such evidence-based tracking can guide national development plans and funding agencies in aligning investments with high-impact sustainability outcomes.

By identifying geographic, thematic, and institutional gaps, this study contributes not only to academic understanding but also to actionable roadmaps for digital and circular transformation in textile waste management. With the right policy support, interdisciplinary collaboration, and technological investments, countries like the Philippines can transform from passive observers into active contributors to the global sustainability agenda.

ACKNOWLEDGMENT

The researcher wishes to acknowledge National University, Philippines, for providing the institutional support and academic environment necessary to carry out this study. Special thanks are extended to the College of Computing and Information Technologies (CCIT) Graduate School for their academic guidance and encouragement throughout the development of this research. The researcher also acknowledges the Research and Development Office for its unwavering support and commitment to scholarly excellence, which have been instrumental in the successful completion of this work.

REFERENCES

[1] Voukkali, I., Zorpas, A. A., & Lasaridi, K. (2023). The environmental impact of textile and clothing production and the role of the circular economy. Environmental Science and Pollution Research, 30, 45278–45294

[2] Gue, Y., Zhou, L., & Chen, F. (2022). Environmental footprints of the fast fashion industry: A global perspective. Environmental Pollution, 304, 119223.

[3] Poon, T. K., Zhang, L., & Lin, X. (2024). Forecasting textile waste trajectories in Asia: Policy scenarios and impact analysis. Resources, Conservation & Recycling Advances, 22, 200089.

[4] Figueroa, A., de la Cruz, J., & Soriano, D. (2021). Textile reuse in Philippine ukay-ukay markets: A pathway to sustainability or a blind spot in waste governance? Philippine Journal of Environmental Studies, 14(1), 33–47.

[5] Juanga-Labayen, J., Mendoza, C. M., & Antonio, P. (2022). Challenges and prospects of integrating textile waste into Philippine waste management policy. Environmental Policy Journal of Southeast Asia, 10(2), 145–163.

[6] Smol, M., Marcinek, P., Duda, J., & Szołdrowska, M. (2020). Circular economy indicators in relation to eco-innovation in European regions: A bibliometric approach. Sustainability, 12(6), 2488.

[7] Hussain, S., Malik, M., & Ahmed, M. (2021). Blockchain-based E-barter systems: Enabling circular economy in digital textile trading. Sustainable Technology and Entrepreneurship, 2(4), 100040.

[8] Awan, U., & Sroufe, R. (2022). Sustainable supply chain innovation: Bibliometric review and network analysis. Journal of Cleaner Production, 370, 133303.

[9] Biyada, S., & Urbonavičius, J. (2025). Circularity in textile waste: Challenges and pathways to sustainability. *Cleaner Engineering and Technology*, *24*, 100905. https://doi.org/10.1016/J.CLET.2025.100905

[10] Alves, D. I., Barreiros, M., Fangueiro, R., & Ferreira, D. P. (2024). Valorization of textile waste: non-woven structures and composites. Frontiers in Environmental Science, 12, 1365162. https://doi.org/10.3389/fenvs.2024.1365162

[11] Balanay, R., & Halog, A. (2019). Tools for circular economy: Review and some potential applications for the Philippine textile industry. Circular Economy in Textiles and Apparel: Processing, Manufacturing, and Design, 49–75. https://doi.org/10.1016/B978-0-08-102630-4.00003-0

[12] Biyada, S., & Urbonavičius, J. (2025). Circularity in textile waste: Challenges and pathways to sustainability. Cleaner Engineering and Technology, 24, 100905. https://doi.org/10.1016/J.CLET.2025.100905

[13] Blancaflor, E. B., Neyra, J. M. M., Ramel, K. A. S., Tan, S. M. G., & Rey, W. P. (2022). A Design of an E-Waste Collection Web Application to Motivate Waste Segregation & Collection Practices. ACM International Conference Proceeding Series, 493–499. https://doi.org/10.1145/3568834.3568857

[14] Blancaflor, E., Doydoy, J. D., Rodriguez, Y. K., Tomagos, G., Moreno, D. E., & Abisado, M. (2025). Design and Implementation of 'Sortino': An IoT based Smart Waste Sorting System. 2025 8th International Conference on Control, Robotics and Informatics (ICCRI), 41–45. https://doi.org/10.1109/ICCRI67201.2025.11262266

[15] Bueta, G. R., Domingo, S., & Manejar, A. J. (2024). Study on Circular Economy Pathways for Waste Management in the Philippines. https://doi.org/10.62986/DP2023.46

[16] Das, B., & Dwivedi, S. (2024). Post-Consumer Textile Waste Management Practices and Challenges in India: A Systematic Literature Review. Journal of Scientometric Research, 13(2), 419–429. https://doi.org/10.5530/JSCIRES.13.2.33

[17] Din, T. U., Zaman, N., Uddin, M. N., Khan, W., Qayum, N., Qayum, S., Iqbal, S., Liaqat, I., Mumtaz, A., Safi, S. Z., Tahir, A., & Arshad, M. (2025). The Circular Economy Concept in the Textile Industry. 331–361. https://doi.org/10.1007/978-981-97-8058-7_13

[18] Dziubaniuk, O., Aarikka-Stenroos, L., & Pohls, E. L. (2023). Catalysing The Textile Industry Toward A Circular Economy: An ecosystem approach. The Routledge Handbook of Catalysts for a Sustainable Circular Economy, 67–87. https://doi.org/10.4324/9781003267492-5

[19] Edirisinghe, L. G. L. M., de Alwis, A. A. P., & Wijayasundara, M. (2024). Sustainable circular practices in the textile product life cycle: A comprehensive approach to environmental impact mitigation. Environmental Challenges, 16, 100985. https://doi.org/10.1016/J.ENVC.2024.100985

[20] Körlü, A., & Körlü, A. (2020). Waste in Textile and Leather Sectors. Waste in Textile and Leather Sectors. https://doi.org/10.5772/INTECHOPEN.90014

[21] Leal Filho, W., Dinis, M. A. P., Liakh, O., Paço, A., Dennis, K., Shollo, F., & Sidsaph, H. (2024). Reducing the carbon footprint of the textile sector: an overview of impacts and solutions. Textile Research Journal , 94(15–16), 1798–1814. https://doi.org/10.1177/00405175241236971

[22] Rahaman, Md. T., & Pranta, A. D. (2025). Transitioning from Linear to Circular Economy in Textiles and Apparel: Significance, Case Studies, Innovations, and Research Directions. Materials Circular Economy 2025 7:1, 7(1), 48-. https://doi.org/10.1007/S42824-025-00204-6

[23] Ramírez-Escamilla, H. G., Martínez-Rodríguez, M. C., Padilla-Rivera, A., Domínguez-Solís, D., & Campos-Villegas, L. E. (2024). Advancing Toward Sustainability: A Systematic Review of Circular Economy Strategies in the Textile Industry. Recycling 2024, Vol. 9, Page 95, 9(5), 95. https://doi.org/10.3390/RECYCLING9050095

[24] Riba, J. R., Cantero, R., Canals, T., & Puig, R. (2020). Circular economy of post-consumer textile waste: Classification through infrared spectroscopy. Journal of Cleaner Production, 272, 123011. https://doi.org/10.1016/j.jclepro.2020.123011

[25] Schnatmann, A. K., Schoden, F., Ehrmann, A., & Schwenzfeier-Hellkamp, E. (2023). R principles for circular economy in the textile industry – a mini-review. Communications in Development and Assembling of Textile Products, 4(2), 295–305. https://doi.org/10.25367/cdatp.2023.4.p295-305

[26] Schumacher, K. A., & Forster, A. L. (2022). Textiles in a circular economy: An assessment of the current landscape, challenges, and opportunities in the United States. Frontiers in Sustainability, 3, 1038323. https://doi.org/10.3389/frsus.2022.1038323

[27] Silobrit, I., & Jureviciene, D. (2023). Assessing Circular Textile Industry Development. Economics and Culture, 20(1), 55–67. https://doi.org/10.2478/JEC-2023-0005

[28] Ütebay, B., Çelik, P., Çay, A., Ütebay, B., Çelik, P., & Çay, A. (2020). Textile Wastes: Status and Perspectives. Waste in Textile and Leather Sectors. https://doi.org/10.5772/INTECHOPEN.92234

[29] Zhang, J., & Blancaflor, E. B. (2025). A Review of YOLO Algorithm and The Research in Garbage Classification Target Detection. Proceedings of the 2024 4th International Joint Conference on Robotics and Artificial Intelligence, JCRAI 2024, 182–187. https://doi.org/10.1145/3696474.3702878

An Automatic Transfer Switch Scheme on a 7.7 kWp Solar Panel System with Metaheuristic Approximation

1st Ryan Yudha Adhitya *, 2nd Rendra Cahya Putra *, 3rd Dimas Pristovani Riananda *, 4th M. Abu Jami'in *, 5th Yudi Andika *, 6th Aminatus Sa'diyah †

* *Marine Electrical Engineering, Politeknik Perkapalan Negeri Surabaya, Surabaya, Indonesia*

† *Marine Engineering, Politeknik Perkapalan Negeri Surabaya, Surabaya, Indonesia*

Emails: *1st ryanyudhaadhitya@ppns.ac.id, 2nd m.rendra@student.ppns.ac.id, 3rd dimaspristovani@ppns.ac.id, 4th jammy@ppns.ac.id, 5th yudi.andika@ppns.ac.id, 6th am.sadiyah@ppns.ac.id*

***Abstract*— Renewable energy, especially solar power plants (SPP), is increasingly being used as an alternative solution to meet electricity needs. One of the main obstacles in the implementation of SPP is the instability of the electricity supply due to changes in weather conditions. To overcome this condition, an Automatic Transfer Switch (ATS) DC system is implemented as an automatic mechanism to transfer electricity between the SPP power source and the Indonesian State Electricity Company (ISEC) grid. This study aims to design a DC ATS control scheme on a 7.7 kWp Solar Panel System using a metaheuristic approach. This method is used to determine power transfer decisions based on parameters such as voltage, current, and battery capacity. The results show that the proposed switching system can operate optimally in transferring resources with a high level of accuracy. The application of the metaheuristic method improves the speed and accuracy of the decision-making process, thereby supporting the stability and sustainability of electricity supply.**

***Keywords*— Automatic Transfer Switch DC, SPP, ISEC, Solar Charge Inverter Hybrid, Metaheuristic.**

I. INTRODUCTION

The increasing demand for electrical energy encourages the development of renewable energy sources, one of them is Solar Power Plant (SPP). SPP is an environmentally friendly and sustainable alternative solution for reducing dependence on fossil fuels [1-10]. Currently the use of Solar Power Plants (SPP) is increasing. Some of the SPPs use the ON Grid system with the Indonesian State Electricity Company (ISEC) and some use the Off Grid system [11]. However, the reliability of the electricity supply from SPP is still a major challenge because it is influenced by uncertain factors, such as varying sunlight intensity throughout the day. To ensure continuity of power supply, the Automatic Transfer Switch (ATS) DC system is needed as an automatic mechanism that allows switching of electricity sources between SPP and ISEC network [5-7].

To demonstrate the metaheuristic based control scheme, this study is using 7.7 kWp Solar Panel System, with a maximum of each PV power of 555 watts, a maximum PV voltage of 150 VDC, nominal frequency of 50 to 60 Hz. And the inverter output voltage is 230 VAC With this system, when the power from SPP is insufficient, the electricity supply can immediately switch to ISEC without interruption [13], thus maintaining the stability of the power received by the load. To improve the efficiency of decision making in power transfer, the metaheuristic method is applied. This method is able to determine switching decisions based on parameters such as voltage, current, and battery charging status quickly and accurately. The main objective of this study is to analyze the performance of metaheuristic based switching method ensuring optimal power transition between SPP and ISEC [14]. The results of this study are expected to contribute to the development of a more reliable and efficient hybrid energy system, especially in household and small industry applications.



Fig. 1. 7.7 kWp Solar Panel System.

II. LITERATURE REVIEW

A. Automatic Transfer Switch (ATS)

ATS is an essential component in hybrid energy systems enabling automatic switching between different power sources. This automatic control system is usually called Automatic Transfer Switch (ATS). Automatic Transfer Switch is a series of controls for Power Inverter switches with ISEC that work automatically [14-15]. Research such as that conducted by Maryanto and Sikki [1] discusses the integration of ATS with Automatic Main Failure (AMF) systems and highlights its effectiveness in maintaining power continuity during supply disruptions [17]. ATS systems are particularly relevant in campus environments where uninterrupted power supply is critical to support academic and operational activities.

B. Metaheuristic Method

Metaheuristic algorithm is widely used for decision-making processes in various fields, including energy management. Robianto et al. [2] highlighted its ability to classify data and optimize decision making through a structured tree model. By analyzing parameters such as voltage and current, the metaheuristic method enables efficient switching of energy sources, thus ensuring system stability [18]. The Metaheuristic methods is a classification method that builds a model to predict the class of an object.

Metaheuristic algorithms was chosen because it is able to classify and it can explain the characteristics of the data that has been modeled [16-18]. Adaptive Neuro Fuzzy Inference System (ANFIS) and Recurrent Neural Network (RNN) are popular algorithms method in machine learning [20-22]. This algorithm can be used to decide the optimal resource (solar panels or ISEC) based on the solar panel voltage and current value.

C. Related Case Studies

Research by Fauzy et al. [3] focuses on an IoT-based telemetry system for a solar power plant, providing insights into monitoring and optimization of renewable energy systems. Similarly, [4] uses simulation tools to design a solar power system for an academic facility, highlighting the importance of efficient energy management solutions. In nonlinear decision support systems, the decision tree method has several significant drawbacks. One of the main ones is a tendency to overfit, where the model overfits the training data. This leads it to create highly specific and complex rules, failing to make accurate predictions on new, unseen data. Furthermore, decision trees are highly unstable; small changes to the training data can produce a completely different tree structure, leading to inconsistent decisions. They are also less effective at capturing complex relationships where outcomes depend on the interaction or combination of multiple variables simultaneously. Finally, decision trees have limitations in extrapolation, as they cannot predict values outside the range of the learned data. Consequently, for nonlinear cases, methods such as backpropagation neural networks are often recommended because they can overcome these drawbacks.

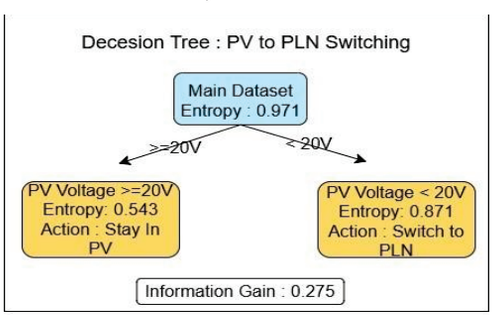


Fig. 2. Decision Tree Scheme (Case Study) [15].

III. METHODOLOGY

A. System Design

The proposed ATS system is designed to automatically switch between energy from solar panels (PV) and ISEC electricity based on voltage conditions.

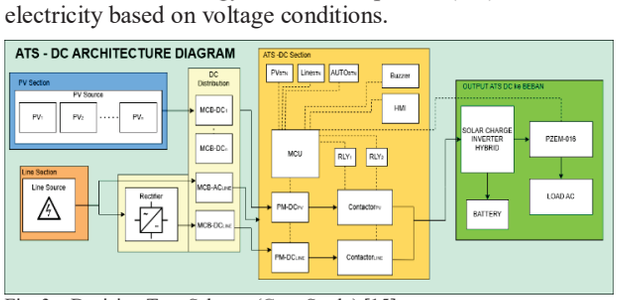


Fig. 3. Decision Tree Scheme (Case Study) [15].

B. System Working Principle

This MPPT (Maximum Power Point Tracking) charging system utilizes energy sources from solar panels (PV) and ISEC automatically based on the voltage conditions generated by the PV. This system uses ESP32 as the main microcontroller to control the entire process, including reading data from PZEM and CT sensors, analyzing voltage conditions, and regulating the activation of contactors, switches, and bells. The process begins with the PV active, where the sensor reads the voltage and current generated by the PV.

The data is analyzed by the ESP32 to determine whether the voltage generated reaches the minimum value of MPPT charging condition. If the voltage is sufficient, the system will activate contactor 1, and sound the buzzer as an indication that the power source comes from the PV. The MPPT charger is then activated to charge using the energy from the PV. Conversely, if the PV voltage is insufficient, the system will automatically turn off and switch to the power source from ISEC. In this process, it will change AC power from ISEC to DC, then contactor from ISEC will be active to activate the load. After that, the bell sounds as an indication that the system is using a source from ISEC, and the MPPT charger continues to run to charge energy from ISEC. This system is designed to ensure that charging continues to run optimally by prioritizing the use of

Renewable energy from PV able to change to ISEC as a backup if needed. With ESP32-based automatic control, this system ensures efficiency and provision in energy management for charging power needs.

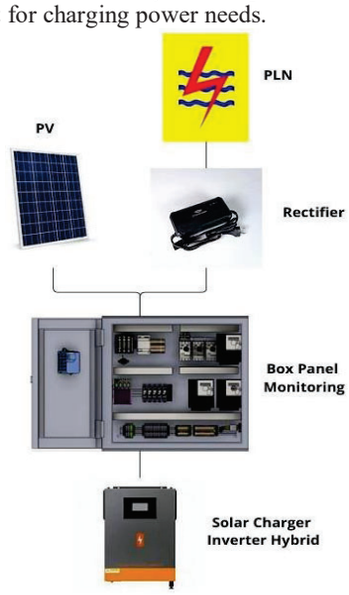


Fig. 4. System Implementation.

This MPPT (Maximum Power Point Tracking) charging system utilizes energy sources from solar panels (PV) and ISEC automatically based on the voltage conditions generated by the PV. This system uses ESP32 as the main microcontroller to control the entire process, including reading data from PZEM and CT sensors, analyzing voltage conditions, and regulating the activation of contactors, switches, and bells. The process begins with the PV active, where the sensor reads the voltage and current generated by the PV. The data is analyzed by the ESP32 to determine

whether the voltage generated reaches the minimum value of SOC, PV Voltage and PV Current. If those parameters are sufficient, the system will activate contactor 1, and sound the buzzer as an indication that the power source comes from the PV. The MPPT charger is then activated to charge using the energy from the PV. Conversely, if the PV voltage is insufficient, the system will automatically turn off and switch to the power source from ISEC. In this process, it will change AC power from ISEC to DC, then contactor from ISEC will be active to activate the load. After that, the bell sounds as an indication that the system is using a source from ISEC, and the MPPT charger continues to function to charge using energy from ISEC. This system is designed to ensure that charging continues to run optimally by prioritizing the use of renewable energy from PV, but can still switch to ISEC as a backup if needed. With ESP32-based automatic control, this system ensures efficiency and provision in energy management for charging power needs. The working principle of this research explained in the block diagram below.

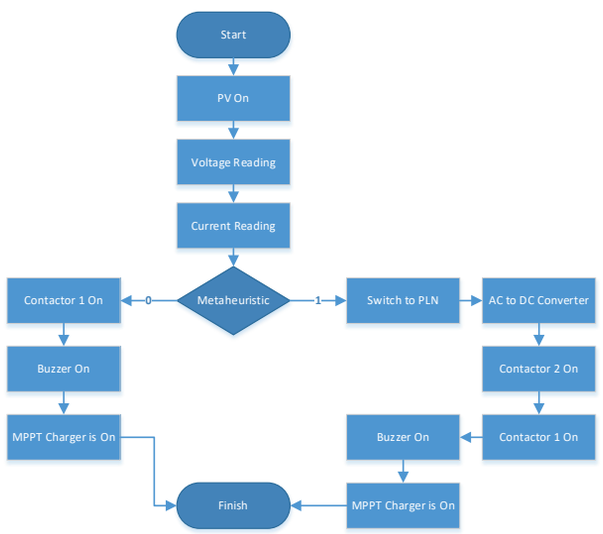


Fig. 5. Flowchart System.

C. Metaheuristic Integration

A metaheuristic algorithm is used to analyze voltage and current data in real time, facilitating optimal switching decisions. Parameters such as measured voltage and current are fed into the algorithm to determine the appropriate power source. The algorithm is implemented on an ESP32 microcontroller. This study uses a comparison of two metaheuristic methods, namely the adaptive neurofuzzy inference system (ANFIS), which has the main advantage of robust nonlinear system modeling, and the recurrent neural network (RNN), which is a development of the basic structure of artificial neural networks [20-22]. Adaptive Neuro-Fuzzy Inference System (ANFIS) is a hybrid system that combines the principles of neural networks and fuzzy logic to model complex input-output relationships. Mathematically, ANFIS is based on a first-order sugeno fuzzy inference model. The ANFIS structure consists of several layers, each with its own specific function and set of mathematical equations. Here is the complete ANFIS equation:

$$f = \frac{\omega 1(p1x+q1y+r1)+\omega 2(p2x+q2y+r2)}{\omega 1+\omega 2} \quad (1)$$

This equation shows that the ANFIS output is a weighted average of each consequent function (f_i), where the weights are the normalized firing strength (ω_i). The parameters of the ANFIS (the membership function parameters a_i, b_i, c_i and the consequent parameters p_i, q_i, r_i) are adjusted through a hybrid training process to minimize the error between the ANFIS output and the target data. While The core idea of an RNN is that it has a memory, called the hidden state (h_t) which is updated at each time step t using the current input (x_t) and the previous hidden state (h_{t-1}). As the baseline dataset for modeling both methods, the data collection scenario was conducted on a 7.7 kWp solar PV system that had been built. Data was collected with a dynamic resistive load to determine the decrease in the solar PV battery capacity response and as a basis for determining the switching state. The following is the response data for solar PV usage in this study:

TABLE I. DATA RESPONSE OF SOLAR PANEL SYSTEM

Id	I_PV(A)	V_PV(V)	SOC_batt(%)	Switching State
1	0.0	1.0	53.0	1
2	0.0	1.3	53.0	1
3	0.0	1.0	53.0	1
4	0.0	0.9	51.0	1
5	0.0	1.4	53.0	1
6	0.0	1.1	53.0	1
7	0.0	1.0	51.0	1
8	0.0	1.0	53.0	1
9	0.0	1.4	53.0	1
10	0.0	1.1	53.0	1
11	0.0	1.0	53.0	1
12	0.0	1.4	53.0	1
13	0.0	1.3	53.0	1
14	0.0	1.0	53.0	1
…	…	…	…	…
4058	0.0	48.4	39.0	0
4059	0.0	30.8	39.0	0
4060	0.0	13.6	36.0	0
4061	0.0	6.2	36.0	0
4062	0.0	3.7	34.0	0
4063	0.0	3.0	29.0	0
4064	0.0	4.1	19.0	0

Description:
ATS output 0: PV is used.

ATS output 1: ISEC is used

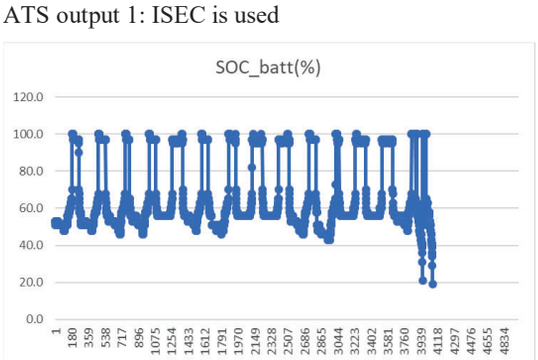
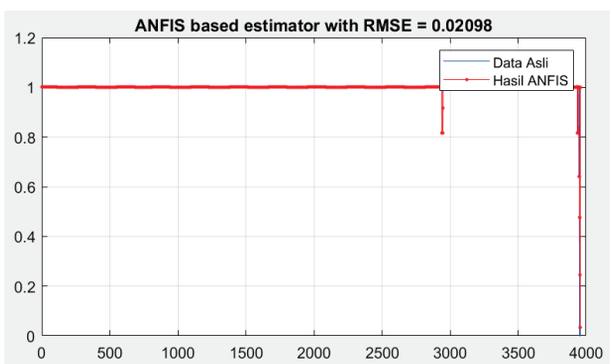


Fig. 6. SOC response in September 2025.

IV. RESULT AND DISCUSSION

A. Training stage using ANFIS

In designing a decision support system, there are several data-based techniques, namely through a statistical approach and also through a metaheuristic approach. In some cases, the statistical approach is limited to modeling linear data and has weaknesses in nonlinear modeling cases with high data fluctuation characteristics. Therefore, two metaheuristic-based approaches were carried out. Namely, the fuzzy logic-based method or better known as ANFIS (Adaptive Neuro Fuzzy Inference System) and the artificial neural network-based method or better known as RNN (Recurrent Neural Network). In the first scenario, the ANFIS method design was carried out with the number of membership functions for the voltage input and current input variables being 3, the membership form is triangular (trapmf), the membership output is a constant 0/1 with 0 being the normal condition and 1 being the minimum capacity condition which is the basis for the switching process, and the number of iterations is 100 iterations. The training process uses the help of MATLAB 2024a software with the Matlab file facility which calls the file with the extension .xls (dataset file in Table II for further training stages until the desired conditions are achieved. The following are the training results using ANFIS for 3957 pairs of training data and 107 pairs of testing data:



(a)

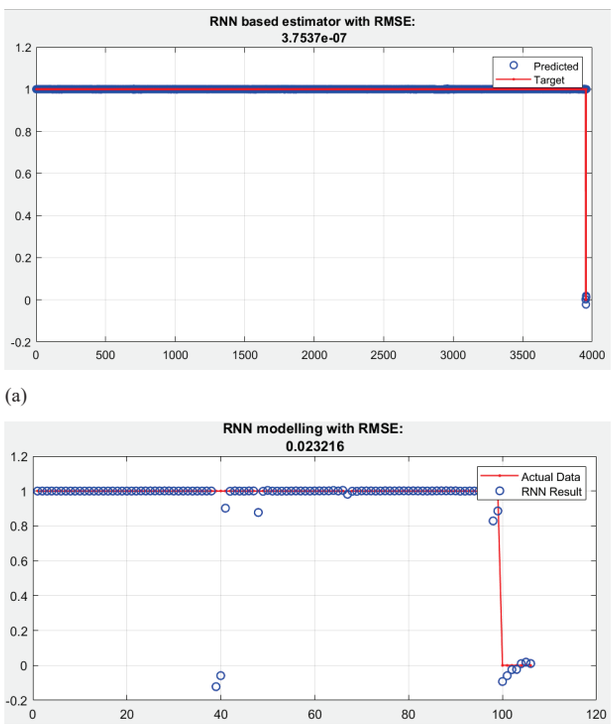
(b)

Fig. 7. (a) Training stage and (b) testing age.

ANFIS modeling shows an RMSE (mean difference) value of 0.135. It can be concluded that ANFIS-based modeling can be applied as a decision support system in this study.

B. Training stage using RNN

A Recurrent Neural Network (RNN) is a type of artificial neural network specifically designed to process sequential data, such as text, speech, or time series. Unlike conventional neural networks, RNNs have an internal memory that allows them to use information from previous inputs to influence the current input. The mechanism involves a feedback loop in which the output from the previous time step is fed back as input to the next time step, stored in a hidden state. This makes RNNs highly effective at understanding context and temporal dependencies in data. The RNN structure used in this study uses two hidden layers, each with three hidden neurons and nine hidden neurons. The weight update method uses Bayesian regression because this method reaches the target condition faster than other methods such as Lavenberg-Marquardt, gradient descent, and gradient ascent. The iteration limit is set at 100 iterations. The following is a comparison curve between the actual data (switching state) and the RNN prediction data.



(a)

(b)

Fig. 8. (a) Training stage and (b) testing age.

Figure 8 shows that RNN produces a smaller RMSE value when compared to ANFIS, with a value reaching 0.023. so that RNN is more visible to apply. These two methods have shown a representative curve and follow the actual data pattern with a relatively small RMSE value. So RNN can be applied to the basic rules of the switching process with dynamic (non-linear) input parameter references. The implementation of RNN in the next stage can be done by converting the RNN structure using a specific programming language (such as the C language used in microcontrollers), so that the real-time auto switching process can be applied. In its application, softmax-based thresholding can be added to round the value of 0 or 1 with the information 0 is the use of PV and 1 is the use of ISEC.

V. CONCLUSION

This research has successfully modeled a switching control scheme for dynamic or nonlinear input parameters. It

is proven by ANFIS and RNN-based metaheuristic modeling that is able to achieve minimal RMSE values and follow actual data (switching state), where the switching state is based on PV voltage and current conditions. With this switching arrangement, it can become an intelligent relay to determine the use of renewable electricity sources and optimize the reliability of solar panel batteries.

REFERENCES

[1] S. M. I. Maryanto I, "Sistem Automatic Transfer Switch (Ats) Automatic Main Failure (Amf) Menggunakan Sms," Jrec (Journal Of Electrical And Electronics, Vol. 6, No. No. 1, P. 19–32, 2018.

[2] Robianto, "Penerapan Metode Decision Tree Untuk Mengklasifikasikan Mutu Buah Jeruk Berdasarkan Fitur Warna Dan Ukuran," Coding Jurnal Komputer Dan Aplikasi, Vol. 9, No. No. 01,

[3] Fauzy, "Rancang Bangun Alat Telemetri Parameter Pembangkit Listrik Tenaga Surya Berbasis Iot," Jurnal Eksitasi, Vol. 1, No. 1, Pp. 14-21, 2022.

[4] Zambak, "Rancang Bangun Pembangkit Listrik Tenaga Surya (SPP) Pada Laboratorium Teknik UMSU Menggunakan Simulasi Pvsyst," Jurnal Teknologi Elektro, Vol. 14, No. 2, P. 72, 2023.

[5] Slamet Han, "Solar Power Plant Design Using Automatic Transfer Switch on ISEC-SPP Network," ETJ (Engineering and Technology Journal), Vol. 9, No. 7, pp. 4527-4532, 2024.

[6] Adi Pratama Putra dan Adi Mulyadi, "Design an Automatic Transfer Switch for Solar Power Plant," Jurnal Teknologi Elektro, Vol. 15, No. 1, pp. 10-15, 2023.

[7] Muhammad Agil Haikal, "Solar Power Plant Optimization using Automatic Transfer Switch (ATS) and Low Voltage Disconnect (LVD)," Jurnal Energi Terbarukan, Vol. 2, No. 3, pp. 45-50, 2024.

[8] Y&H, "1000W Solar Hybrid Inverter DC12V to AC230V, Off-Grid Pure Sine Wave Inverter with 50A MPPT Solar Charger Controller," Amazon.com, 2023.

[9] VoltaconSolar, "Motorised and Automatic Transfer Switches," VoltaconSolar.com, 2023.

[10] Clean Energy Reviews, "Guide to Designing Off-Grid and Hybrid Solar Systems," CleanEnergyReviews.info, 2024.

[11] DIY Solar Power Forum, "Hybrid Inverter or Inverter with ATS for Small Scale Installation," DIYSolarForum.com, 2022.

[12] J. Jamaaluddin, S. D. Ayuni, and I. A. S. Wulandari, "Design of Automatic Transfer Switch System Solar Power Plant – ISEC," J. Electr. Technol. UMY, vol. 7, no. 2, pp. 57–64, 2024, doi: 10.18196/jet.v7i2.14651.

[13] E. P. Febtiawan, L. A. Syamsul, I. Akbar, and A. S. Rachman, "Forecasting Produksi Energi Photovoltaic Menggunakan Algoritma Random Forest Classification," J. Inf. Syst. Res., vol. 5, no. 4, pp. 1053–1062, 2024, doi: 10.47065/josh.v5i4.5514.

[14] H. Nur, K. Ningrum, B. Winarno, A. Dharma, and R. Gusta, "Design of Pure Sine Wave Inverter with Automatic Transfer Switch for Hybrid Solar Power Plant," vol. 08, no. 02, 2024.

[15] W. Safitri, S. N. Anisa, A. Al-habsyi, T. Elektro, and P. N. Jakarta, "Implementasi Algoritma Decision Tree dalam rangka Peningkatan Efisiensi Energi Penggunaan Beban Listrik dalam Ruangan Abstrak SNIV : SEMINAR NASIONAL INOVASI VOKASI," vol. 3, no. 1, pp. 506–515, 2022.

[16] S. Hariyanto, D. Lobo, F. Mahmuddin, and S. Klara, "Automatic Transfer Switch (ATS) for Floating Lift Net Powered by Hybrid Energy Source," E3S Web Conf., vol. 601, pp. 1–12, 2025, doi: 10.1051/e3sconf/202560100088.

[17] S. Jain and V. Sharma, "Decision Trees in Intrusion Detection : A Comparative Analysis of Machine Learning Techniques," no. 1, 2025.

[18] M. Abdillah, T. Mutia, T. A. Nugroho, and N. I. Pertiwi, "Design of Automatic Transfer Switch on A Renewable Energy Hybrid Grid System at PT Lentera Bumi Nusantara," J. Adv. Technol. Multidiscip., vol. 1, no. 2, pp. 38–44, 2022, doi: 10.20473/jatm.v1i2.40293.

[19] L. A. Venancio and A. Y. Chua, "A novel low cost automation of transfer switch control for a hybrid solar power system with simulation," Int. J. Mech. Eng. Robot. Res., vol. 9, no. 7, pp. 987–997, 2020, doi: 10.18178/ijmerr.9.7.987-997.

[20] D. P. Riananda et al., "System Protection Analysis on 1kW Solar Charge Hybrid Inverter Using Adaptive Neuro-Fuzzy Inference System (ANFIS)," 2025 International Conference on Smart Computing, IoT and Machine Learning (SIML), Surakarta, Indonesia, 2025, pp. 1-6, doi: 10.1109/SIML65326.2025.11080910.

[21] Adhitya, R.Y., Rinanto, N., Syafiuddin, Resmi, F., Arumsari, N., Sarena, S.T. (2025). Advanced Data Analytics for Maritime Energy Efficiency. In: Elsisi, M., Rinanto, N., Su, CL. (eds) Maritime Infrastructure for Energy Management and Emission Reduction Using Digital Transformation. Studies in Infrastructure and Control. Springer, Singapore. https://doi.org/10.1007/978-981-96-4438-4_4.

[22] Ryan Yudha Adhitya, "Sistem Pengendalian Robot KRSRI Menggunakan Logika Fuzzy Sugeno Orde Nol", Jurnal Elkolind Volume 11, Nomor1, Mei 2024 DOI: http://dx.doi.org/10.33795/elkolind.v11i1.4057.

Optimization of the YOLOv12s Architecture with the Addition of 160x160 and C3k2 Heads for Floating Plastic Bottle Detection

1st Alief Akbar Agung El-Mahirah
Department of Informatics
Hasanuddin University
Makassar, Indonesia
elmahirahaaa22d@student.unhas.ac.id

2nd Novy N.R.A Mokobombang*
Department of Informatics
Hasanuddin University
Makassar, Indonesia
novy@unhas.ac.id
Corresponding Author

***Abstract*— Plastic waste pollution in aquatic ecosystems poses critical environmental challenges, with floating plastic bottles being a major contributor. Automatic detection using computer vision offers an efficient monitoring solution, however, it faces significant challenges including very small object sizes, colors similar to the water environment, and variations in lighting and water reflection. This study aimed to optimize the YOLOv12s architecture to improve the detection capability of small floating plastic bottles by adding a 160x160 high-resolution detection head and integration of C3K2 modules. The 160x160 detection head operates on the P2 feature map to maintain detailed spatial information, whereas the C3K2 module amplifies spatial feature extraction with maintained parameter efficiency. The evaluation was conducted on 27 test images with variations in shooting height (5, 6, and 7 m) using precision, recall, F1-score, and mAP metrics. The results showed that the modified model achieved a precision of 0.79, an increase in mAP50 from 0.88 to 0.92 (an increase of 4%), an increase in recall from 0.94 to 0.97 (an increase of 3%), and an increase in the F1-Score from 0.86 to 0.89 (an increase of 3%). The modified model detected 30 objects compared to 25 objects at baseline with faster convergence. Increased recall is crucial because it minimizes missed objects in environmental monitoring. This study contributes to the development of an automatic floating waste monitoring and cleaning system for the preservation of aquatic environments.**

Keywords — YOLOv12s, object detection, floating plastic bottles, head 160×160, C3K2, computer vision

I. INTRODUCTION

Plastic waste, especially plastic bottles, has become a critical urban environmental problem as population growth and development activities have intensified. This source of marine debris comes from community activities in coastal areas as well as land-based waste flows through rivers, with macro and microplastic accumulations increasing consistently over the past four decades[1]. Although plastic bottles play an important role in daily life, their short lifespan makes them a significant source of post-consumer pollution[2]. In Jakarta, about 74% of anthropogenic waste on the river surface consists of plastic, with a leakage rate of 9.9 grams per capita per day during the rainy season [3]. The most harmful impact of this pollution is the resistance of plastics to natural decomposition, where the material is simply fragmented into microplastics and nanoplastics that continue to pollute the environment in the long term [4]. Therefore, the evaluation and management of plastic bottle waste is an urgent priority in efforts to control water pollution and protect aquatic ecosystems .

As an efficient monitoring solution, Computer Vision is a key field of artificial intelligence that allows systems to mimic human visual perception to recognize objects and patterns from digital imagery[5]. Historically, this technology has evolved from a rules-based approach to a modern paradigm dominated by deep learning, primarily through Convolutional Neural Networks (CNNs). CNN architecture is capable of extracting spatial features with high precision for a variety of complex tasks such as image classification, object detection, semantic segmentation, and motion tracking [6].

The application of an automated approach using computer vision offers an effective solution for detecting floating debris, regardless of visual challenges such as small object size, color similarity to the environment, and variations in water lighting. In this context, the YOLO (You Only Look Once) algorithm is widely adopted due to its advantages in speed and real-time accuracy. Several studies have attempted to optimize this method; [7]Proposed YOLOv5_CBS that excels in increased recall over the standard, but is still limited to small object detection because it uses only three detection scales. Furthermore, [8] developed YOLOv8-MSS by adding a 160 x 160 scale, C2f_MLCA module, and a SIoU loss function, which managed to achieve a mAP@0.5% increase over standard YOLOv8.

With the latest version of YOLO, YOLOv12, it offers a variety of architectural enhancements specifically designed to improve performance on small objects. YOLOv12 uses the R-ELAN backbone to amplify feature flow, an A2C2F (Area Attention Cross-Channel Fusion) neck to focus detection on critical areas and improve feature fusion, and a FlashAttention module that speeds up the inference process without sacrificing accuracy[9].

Based on the description above, this study aims to optimize YOLOv12s in detecting small waste objects such as plastic bottles on the surface of the water. One potential approach is to add a scale such as 160×160, and add C3k2 to the head. By evaluating model performance with metrics such as precision, recall, FI-Score, and mAP, before and after optimization. The results of this research are expected to support the development of an automatic monitoring system for floating waste and contribute to efforts to preserve the aquatic environment.

II. RELATED WORK

YOLO (You Only Look Once) has revolutionized real-time object detection by constantly balancing speed and accuracy through continuous architectural innovation. Each version introduces significant improvements that make up the model, following the development of the yolo architecture based on the innovations of each version [10].

YOLOv12 is the latest iteration in the You Only Look Once (YOLO) series of object detection algorithms, which will be released in February 2025. This version marks a significant advance in real-time object detection by integrating attention mechanisms into the YOLO framework, while maintaining competitive inference speeds. This attention-centered approach allows YOLOv12 to surpass other popular real-time detectors in terms of accuracy, while achieving advanced performance through a combination of innovative attention methods, Residual Efficient Layer Aggregation Networks (R-ELAN), and various architectural optimizations. The YOLOv12 architecture is a follows:

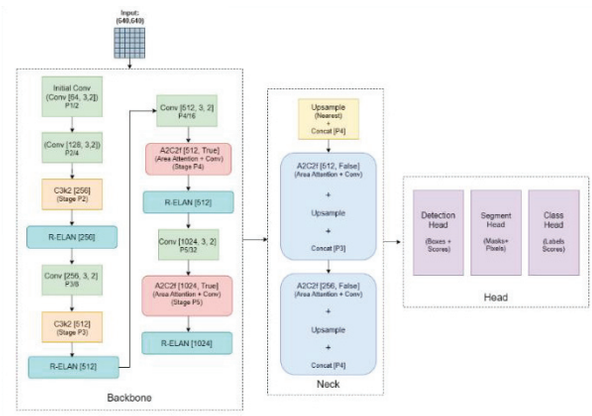


Fig 1. YoloV12 Architecture

YOLOv12 introduces an Area Attention (A^2) module that reduces computational complexity by up to 50% while maintaining large receptive fields, as well as a Residual ELAN (R-ELAN) hierarchy that reduces the number of parameters by 18% and FLOPs by 24%. With architectural optimizations such as depth-wise convolution 7×7, adaptive MLP ratio, and Mosaic-9/Mixup augmentation, YOLOv12 achieved a 12.8% mAP increase on the COCO dataset, making it a new benchmark in real-time object detection[11].

The addition of a 160 × 160 layer head is an important strategy for improvinging the detection accuracy of small objects by working on a high-resolution feature map of the initial layer of the backbone (P2) prior to further downsampling. Because small objects tend to lose spatial information as they pass through layers that reduce the size of the feature (1/16 or 1/32 resolution), the 160×160 head allows the model to focus detection on fine spatial details with a narrower receptive field, capturing more texture information and object boundaries that are important for classification and localization [12].

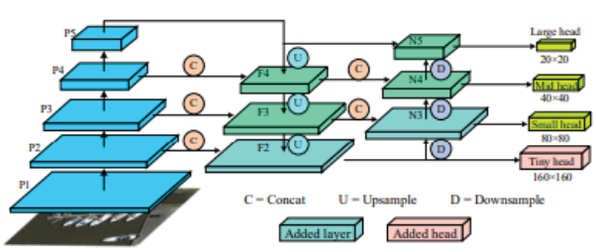


fig 2. Layer 160 x 160

C3k2 is a lightweight convolutional block consisting of three convolutional layers with two shortcut paths, allowing for smooth gradient flow while maintaining parameter efficiency. Developed from the C3 (YOLOv5) and ELAN (YOLOv7) modules, C3k2 is simplified to improve training stability in more complex YOLOv12 architectures, extracting fine as well as rough spatial features with high accuracy especially for the detection of small, complex, or overlapping objects [13].

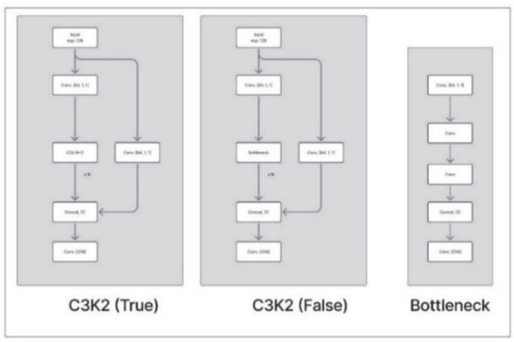


Fig 3. C3K2

Previous studies have validated the effectiveness of YOLO architecture modifications in addressing the challenges of floating waste detection. Research by[8] on YOLOv8, [7] on YOLOv5, and [14] on YOLOv7 collectively shows that the addition of high-resolution heads and the amplification of feature extraction modules are able to mitigate the loss of spatial information on small objects. While this strategy has proven successful in previous versions, exploration of specific optimizations on the latest YOLOv12s architecture for the detection of plastic bottles on the surface of the waterwhich has the challenge of dynamic background noisehas not been widely done.

In contrast to previous research, this study focused on optimizing YOLOv12s through structural modification by adding a 160 x 160 detection scale to the P2 feature map and integrating the C3k2 module on the head. The addition of the detection layer aims to process features at a high resolution to capture the intrinsic details of objects that are often lost in the downsampling process. Meanwhile, the C3k2 module is applied to enrich the representation of spatial features and improve the efficiency of feature extraction, which is crucial for distinguishing small objects from the visual disturbances of the water surface.

There are several fundamental challenges in detecting small objects in aquatic environments that stem from the principles of signal theory and processing:

1. Receptive Field Mismatch

Effective Receptive Field (ERF) measures the actual input region that significantly affects network prediction, which has a Gaussian distribution and is much smaller than the theoretical receptive field. However, in baseline models (such as YOLOv12s) that do not have adaptive mechanisms, the resulting ERF is often still much larger than a small target object [15]. This scale and rigidity mismatch creates a fundamental challenge, which is Receptive Field Mismatch. This scale mismatch leads to two major problems: (1) Feature Dilution, where the features of small objects are averaged against the context of an extensive background, reducing the discriminating

power, and (2) Vulnerability to Clutter, which increases interference from water reflections or other backgrounds, thereby hampering detection performance.

2. Loss Features through Downsampling

Feature Loss through Downsampling is an inherent weakness of Convolutional Neural Networks (CNNs), where the process of progressively reducing spatial dimensions (feature maps) which is essential for expanding the receptive field and reducing computational costs,
leads to a loss of granularity or crucial fine detail. This compression creates an information bottleneck, especially for small objects, as the maximum discriminative information that can be encoded becomes limited, and important details of objects can be completely lost after multiple layers of downsampling[16]. This extreme compression limits the maximum discriminative information that can be encoded, resulting in a significant loss of features and making it difficult to distinguish the object particularly under conditions of occlusion or variation in appearance.

III. RESEARCH METHODOLOGY

A. Dataset

The dataset used in this study is plastic bottles, taking plastic bottle datasets using drones with various altitude distances. There are 3 categories of height for taking objects used, the first is a distance of 5 meters from the water level in the river, the second is 6 meters and the last is a distance of 7 meters. The images were taken in lighting environments as diverse as light and slightly dark, each image containing plastic bottles made up of different shapes and real conditions as found in the field.

B. Modification of YOLOv12

This study proposes two major modifications to the baseline YOLOv12s architecture to improve the detection performance of small floating plastic bottles.

1. Addition of Head Detection 160×160

The first modification was to add a new detection head layer at 160×160 resolution that worked on the high-resolution feature map (P2) of the initial layer of the backbone. This additional head allows the model to focus detection on fine spatial details, retain information on small objects that are likely to be lost due to downsampling and capture more information and boundaries of small objects as shown in figure 5.

2. The Role of Feature Map P2 in Small Object Detection

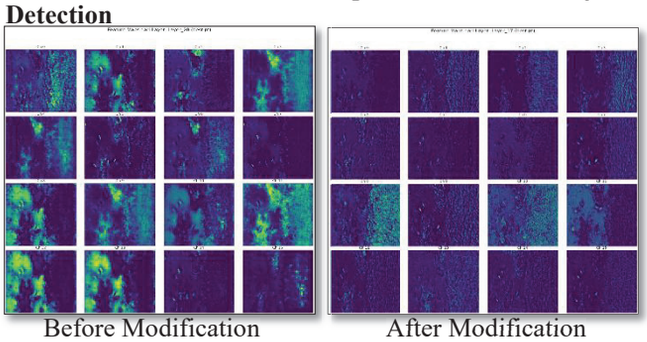


Fig 4. Futuremap/Heatmap visualization

resolution of 1/4 of the input image (for example, 160×160 for a 640×640 input). The scientific basis for the use of P2 in the detection of small objects is based on three fundamental principles: preservation of spatial resolution, receptive field alignment, and semantic-spatial trade-off. Here's a heatmap/future map visualization to show the difference in visualization before and after being modified.

3. Addition of C3k2 Module to the Head

The second modification is to integrate the C3k2 module on the head to strengthen feature extraction. C3k2 is a lightweight convolutional block consisting of three convolutional layers (3 Conv) and, two shortcut paths (k2) to maintain a smooth gradient flow, parameter efficiency. By placing C3k2 on the head, the model can extract fine and rough spatial features with high accuracy, particularly when detecting small, complex or overlapping objects.

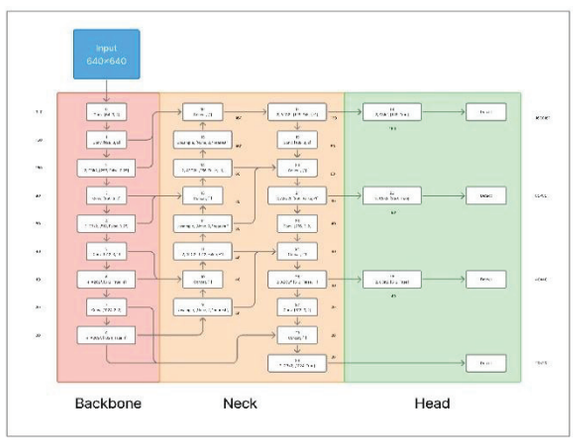


Fig 5. Modified Architecture Yolov12s

C3k2 convolutional blocks are lightweight architectures designed for computing efficiency and small object detection performance. Its effectiveness lies in the use of two skip connections ("k2") that critically optimize the gradient flow to counter the vanishing gradient so that weak small object features can be trained effectively as well as create implicit ensembles to increase model resilience. In addition, C3k2 combines three different 3 x 3 convolutional layers for multi-scale feature extraction, resulting in a dense feature representation that is rich in detail and semantics. This advantage is achieved at very low computing costs, making it a superior choice over the very expensive Transformer-based attention mechanism at high feature resolution.

C. Limitations of the Yolov12 Baseline in Object Detection

1. Limitations on Small Object Detection

Although YOLOv12 has significant architectural updates, the baseline variant still faces crucial obstacles in detecting small objects such as floating bottles. The main problem lies in the loss of spatial information owing to the aggressive downsampling process, where small object details are eliminated before reaching the detection layer. In addition, the absence of a high-resolution detection head (160 x 160 or P2) causes the subtle features of the initial layer of the backbone to be underutilized, thus limiting the model's ability to localize micro-sized objects.

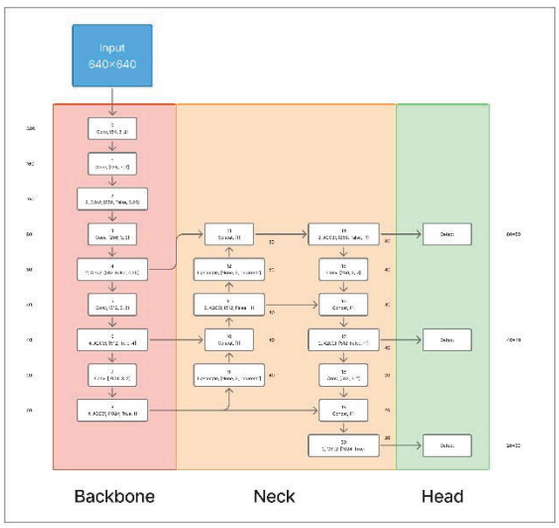


Fig 6. Architecture Yolov12s Base

These challenges are compounded by the complexity of the aquatic environment, such as the interference of water ripples and light reflections that create visual ambiguity between objects and backgrounds. In standard models, too wide a receptive field tends to cause over-smoothing, where small object features become blurred and blend in with environmental noise. Therefore, specific structural modifications are required, namely the addition of the P2 detection scale and the integration of the C3k2 module to maintain the resolution of the feature and increase the sensitivity of the model to such visual interference.

D. Proposed Method

The method described in this study requires customizing the architecture of YOLOv12s, specifically emphasizing changes to the structure of the detection head and strengthening of the feature extraction module. This architectural improvement was achieved by combining two important components: the addition of a detection head at 160×160 resolution and the integration of the C3k2 module on the head. The purpose of this integration is to address the limitations of the YOLOv12s baseline regarding spatial information loss when successive downsampling occurs on very small objects, by utilizing the high-resolution feature representation of the initial layer of the backbone (P2).

The 160×160 detection head operates on high-resolution feature maps with a narrower receptive field, allowing the model to capture fine spatial details and small object texture information without over-smoothing. The C3k2 module integrated into the head serves as a lightweight convolutional block with three convolutional layers and two shortcut paths to maintain the gradient flow while maintaining parameter efficiency. The combination of these two components was designed to enrich the representation of spatial features, resulting in improved detection performance of small plastic bottles compared to the baseline model YOLOv12s.

E. Evaluation

In the field of small-object detection, most methods evaluate the detection performance separately. At this stage, the main evaluation metrics used were Precision, Recall, F1-score, and mean Average Precision (mAP). These metrics are used to quantitatively assess the performance of detection algorithms [17]. The following is an explanation and formula for the evaluation metrics used:

$$P = \frac{TP}{TP+FP} \tag{1}$$

$$R = \frac{TP}{TP+FN} \tag{2}$$

$$F1 = 2 \cdot \frac{P \cdot R}{P+R} \tag{3}$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{4}$$

IV. RESULTS AND DISCUSSION

A. Experimental setup

The entire model training and evaluation process was performed using the Google Colab platform with the T4 GPU as the computing accelerator. The framework used was YOLOv12 Ultralytics Tools, which provides a complete infrastructure for the training, validation, and inference of YOLO-based object detection models. The model was trained with standard configurations that included optimizers, learning rate schedulers, and data augmentation, such as Mosaic-9 and Mixup to enhance modern generalizations.

B. Result of the YOLOv12 model performance evaluation

The performance evaluation result of the object detection model were developed using the YOLOv12s architecture. The evaluation was conducted to compare two model configurations, the Baseline YOLOv12s and Optimized YOLOv12s, which incorporated the addition of a 160×160 resolution detection head and a C3K2 module on the head. The aim was to determine the extent to which the integration of high-resolution heads and C3K2 can improve object detection performance, particularly in the context of detecting floating plastic bottles on the surface of water.

Metric	Model			
	Baseline YOLOv12	**Layer 160x160**	**C3K2**	**YOLOv 12 Modifiy**
Precision (%)	0.78	0.75	**0.84**	0.79
Recall (%)	0.94	**0.97**	0.94	**0.97**
F1-Score (%)	0.86	0.85	0.86	**0.89**
mAP@50 (%)	0.88	0.91	0.88	**0.92**
mAP@55 (%)	0.88	0.91	0.89	**0.92**
mAP@60 (%)	0.88	0.90	0.89	**0.92**
mAP@65 (%)	0.88	0.90	0.89	**0.92**
mAP@70 (%)	0.88	0.89	0.89	**0.92**
mAP@75 (%)	0.87	0.89	0.89	**0.91**

mAP@80 (%)	0.87	0.87	0.88	**0.90**
mAP@85 (%)	0.84	0.83	0.85	**0.88**
mAP@90 (%)	0.77	0.70	**0.78**	0.77
mAP@95 (%)	**0.59**	0.45	**0.59**	0.51
mAP(50-95)	0.45	0.47	0.46	**0.50**

TABLE. I. Comparison Table of Performance of YOLOv12s Baseline and Modified Model in detecting plastic bottle objects

Table I summarizes the results of the ablation studies on the validation dataset. The initial implementation of the 160 x 160 detection head succeeded in drastically increasing the sensitivity of the model, as evidenced by the achievement of 97% Recall and an increase in mAP@50 to 0.91. However, this configuration degrades the precision of strict localization, where mAP@90 plummets to 0.70, implying that the addition of resolution alone causes the bounding box to be less accurate.

This flaw was successfully addressed in the final model (YOLOv12 Modify) through the integration of the C3K2 module, which improved feature representation. This synergy successfully restored mAP@90 to 0.77 while maintaining high sensitivity, resulting in the best performance with an mAP(50-95) of 0.50. This improvement confirms that the combination of a high-sensitivity head and a powerful feature module is crucial for applications that demand few false negatives without sacrificing object-positioning accuracy.

Although these modifications affect computing costs, they can be justified for mission-critical environmental applications. In this context, minimizing missed objects takes precedence over a slight decrease in inference speed, as long as the system remains operational within real-time limits.

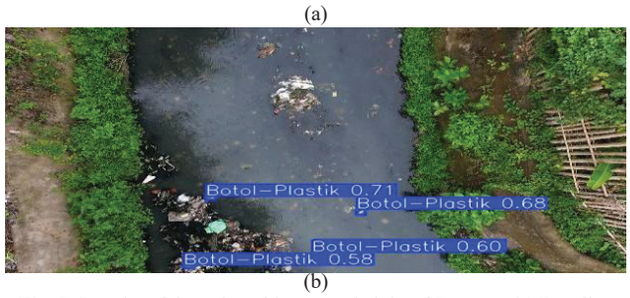
C. Result and Discussion

The results of the performance evaluation on the validation dataset showed that integrating the 160 x 160 detection head and C3K2 module into the architecture (YOLOv12 Modify) significantly improved the capabilities of the model compared to the baseline. As seen in the comparison of the main metrics, the proposed model balances the precision performance and sensitivity. Although the Precision was recorded as stable at 0.79 (slightly below the C3K2-specific variant which reached 0.84), the model showed a clear improvement in the recall (up from 0.94 to 0.97) and F1-Score (up from 0.86 to 0.89). In addition, mAP@50 experienced a significant spike from 0.88 to 0.92, with mAP(50-95) reaching 0.50.

This shift in the value of this metric indicates the characteristics of a much more sensitive model (high sensitivity). In the context of environmental waste detection, the top priority is to minimize the number of missed objects (i.e., false negatives). With a Recall of 0.97, the model was proven to be very effective in capturing almost all target objects, including those that were small or difficult to see. The increase in mAP(50-95) from 0.45 to 0.50 further proves that the proposed architecture is not only more sensitive, but also more accurate overall than the base model, making it a solution for environmental monitoring applications.



(a)



(b)

Fig. 7. Results of detection objects at a height of 7 meters (a) Baseline YOLOv12s and (b) Modif YOLOv12s

A comparison of the detection results at a distance of 7 m showed the superiority of the modified model, although there was a trade-off in the form of a 2% decrease in the average confidence score (0.62 to 0.60). This marginal decrease in precision was compensated for by a significant increase in recall, where the proposed model successfully detected four objects compared to only three objects at baseline.

Comparison of detection results at a distance of 7 meters showed the superiority of the modified model, although there was a trade-off in the form of a 2% decrease in the average confidence score (0.62 to 0.60). This marginal decrease in precision was compensated by a significant increase in Recall, where the proposed model successfully detected 4 objects compared to only 3 objects at baseline.

The visual results of these individual images reflect patterns consistent with the overall quantitative performance of the model. To provide a more comprehensive evaluation, the next section discusses the robustness analysis of the model in handling a wider range of environmental condition scenarios.

Model	**Objek**	**Height**	**Number of detected objects**	**Avg. confidance score**
Baseline Yolov12s	Plastic Bottles	5 Meter	8	0.64
		6 Meter	6	0.74
		7 Meter	11	0.62
Modification Yolov12s	Plastic Bottles	5 Meter	10	0.65
		6 Meter	8	0.61
		7 Meter	12	0.60

Tabel. 2 Base and modify performance comparison table

Table 2 shows a complete aggregation of the detection results on 9 test images with different sampling height variations (5 meters 6 meters, and 7 meters), providing a broader quantitative perspective on the comparison of baseline performance and modification of YOLOv12s.

This systematic evaluation included 27 test images taken under varying lighting conditions, water surfaces

with different turbidity levels, and heterogeneous object densities, such that the results obtained represented the performance of the model under real operational conditions. A comprehensive analysis of this aggregate data showed that the modified model consistently detected more plastic bottle objects at all variations in retrieval altitude, with a total of 30 objects detected compared to 25 objects at baseline, confirming the proposed model's superiority in improving recall capabilities for small object detection while maintaining acceptable accuracy.

This performance improvement is directly correlated with the research goal, which is to optimize the floating plastic bottle waste detection system to support efforts to clean and monitor the aquatic environment automatically and in real-time.

V. CONCLUSION

This research successfully optimized the YOLOv12s architecture for detecting small floating plastic bottles through two key architectural modifications: the addition of a 160×160 high-resolution detection head operating on P2 feature maps and the integration of C3K2 modules in the detection head. The modified model demonstrated significant performance improvements across multiple metrics, achieving a mAP50 of 0.92 (1% increase), recall of 0.97 (2% increase), and F1-Score of 0.89 (2% increase) compared with the baseline.

Evaluation of 27 test images captured at varying heights (5, 6, and 7 m) showed that the modified model detected 30 objects versus 25 objects in the baseline model, with faster convergence. Although the precision slightly decreased from 0.83 to 0.79, the substantial recall improvement is more critical for environmental monitoring applications, as it minimizes missed objects. The high-resolution detection head effectively preserved fine spatial details, whereas the C3K2 module enriched the feature representation with maintained parameter efficiency.

These results confirm that the proposed modifications successfully address the challenges of small-object detection in aquatic environments, including visual interference from water reflections and dynamic backgrounds. This study contributes to the development of automated floating waste monitoring and cleaning systems for aquatic environment preservation.

REFERENCES

[1] Rafsanjani, S. Werorilangi, W. Samad, A. Saru, and A. Faizal, *Identifikasi Sampah Laut Terapung (Floating Marine Debris) Berdasarkan Identification of floating marine debris by monsoon season in the waters of Barranglompo Island, Makassar City*. 2021.

[2] Y. A. Kusuma, D. Hatining, and A. Sudarni, "Pengenalan Pengolahan Sampah Botol Plastik Sebagai Bahan Baku Energi Alternatif," 2022.

[3] M. M. Sari *et al.*, "Plastic pollution in the surface water in Jakarta, Indonesia," *Mar Pollut Bull*, vol. 182, Sep. 2022, doi: 10.1016/j.marpolbul.2022.114023.

[4] N. Ulhasanah *et al.*, "Characterization of Microplastics in Jakarta's Urban Downstream and Estuary Water Bodies," *Ecological Engineering and Environmental Technology*, vol. 25, no. 11, pp. 260–273, 2024, doi: 10.12912/27197050/192678.

[5] R. Szeliski, "Computer Vision: Algorithms and Applications 2nd Edition," 2021. [Online]. Available: https://szeliski.org/Book,

[6] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object Detection in 20 Years: A Survey," Jan. 2023, [Online]. Available: http://arxiv.org/abs/1905.05055

[7] X. Yang *et al.*, "Detection of River Floating Garbage Based on Improved YOLOv5," *Mathematics*, vol. 10, no. 22, Nov. 2022, doi: 10.3390/math10224366.

[8] J. Wang and H. Zhao, "Improved YOLOv8 Algorithm for Water Surface Object Detection," *Sensors*, vol. 24, no. 15, Aug. 2024, doi: 10.3390/s24155059.

[9] Y. Tian, Q. Ye, and D. Doermann, "YOLOv12: Attention-Centric Real-Time Object Detectors Latency (ms) MS COCO mAP (%)," 2025, doi: 10.48550/arXiv.2502.12524.

[10] R. Khanam and M. Hussain, "A Review of YOLOv12: Attention-Based Enhancements vs. Previous Versions," Apr. 2025, [Online]. Available: http://arxiv.org/abs/2504.11995

[11] R. Sapkota *et al.*, "YOLOv12 to Its Genesis: A Decadal and Comprehensive Review of The You Only Look Once (YOLO) Series," 2025. [Online]. Available: https://www.researchgate.net/publication/381650388

[12] M. Yue, L. Zhang, Y. Zhang, and H. Zhang, "An Improved YOLOv8 Detector for Multi-scale Target Detection in Remote Sensing Images," no. 1, 2022, doi: 10.1109/ACCESS.2022.Doi.

[13] M. A. R. Alif and M. Hussain, "YOLOv12: A Breakdown of the Key Architectural Features," Feb. 2025, [Online]. Available: http://arxiv.org/abs/2502.14740

[14] K. Li, Y. Wang, and Z. Hu, "Improved YOLOv7 for Small Object Detection Algorithm Based on Attention and Dynamic Convolution," *Appl Sci (Switzerland)*, vol. 13, no. 16, Aug. 2023, doi: 10.3390/app13169316.

[15] J. Wang *et al.*, "Adaptive Receptive Field Enhancement Network Based on Attention Mechanism for Detecting the Small Target in the Aerial Image," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–18, 2024, doi: 10.1109/TGRS.2023.3337266.

[16] R. Hesse, S. Schaub-Meyer, and S. Roth, "Content-Adaptive Downsampling in Convolutional Neural Networks." [Online]. Available: https://github.com/visinf/cad/.

[17] L. T. Ramos and A. D. Sappa, "A Decade of You Only Look Once (YOLO) for Object Detection: A Review," Aug. 2025, doi: 10.1109/ACCESS.2025.3630988.

Optimal Control of a Nonlinear Tourism Environment Dynamic System for Sustainable Development

1st Ida Ayu Putu Ari Utari
Departement of Mathematics)
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
7002241001@student.its.ac.id

2nd Subchan
Departement of Mathematics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
subchan@its.ac.id

3rd Nikenasih Binatari
Departement of Mathematics
Universitas Negeri Yogyakarta
Yogyakarta, Indonesia
nikenasih@uny.ac.id

***Abstract*—Sustainable tourism management is expected to achieve a balance between the number of tourists, economic growth, investment, and environmental quality. This research formulated a mathematical model that interprets the interactions between the four main components of the tourism sector. Next, the optimal control method is implemented to design policies in mathematical models in sustainable tourism. The policy is to add two control variables, namely environmental recovery efforts and green investment incentives that influence environmental and investment variables. Solving optimal control problems in models using Pontryagin's Maximum Principle. The results of numerical simulations illustrate that optimal control policies can prevent the decline of environmental quality and maintain sustainable growth of tourists, the economy, and investment compared to conditions without control. These results show that optimal control methods can be used to formulate sustainable tourism policies.**

***Index Terms*—optimal control, sustainable tourism, dynamic systems, mathematical modelling, Pontryagin's Maximum Principle.**

I. Introduction

Tourism is an important aspect of economic development in various countries. In some countries, tourism plays a major role in contributing to economic growth, as a source of foreign exchange, and as a provider of employment. However, significant tourist growth has resulted in major challenges to the environment and infrastructure [1]. High reliance on the tourism sector makes the economy vulnerable to external constraints. The COVID-19 pandemic demonstrated this concretely, as global mobility restrictions led to a significant decline in tourist visits and economic problems. The impact of the COVID-19 pandemic has not only affected the economic sector, but also the ecology, namely the relationship between the balance of tourism activities and environmental quality.

Research related to the application of mathematical models in sustainable tourism is developing more and more. Casagrandi and Rinaldi [2] developed a dynamic system model that depends not only on the number of tourists but also on the quality of the environment and infrastructure. According to Gössling et al. [3], Seraphin et al. [4], and Hjálmarsdóttir et al. [5] it is reinforced that tourism has a complex relationship with the environment and economic growth, where tourism activities can increase income but also accelerate environmental damage. Next, Ying [6] developed a dynamic mathematical model simulated with MATLAB. The simulation results showed asymptotically unstable conditions, where the number of tourists and environmental pressures are unstable if not properly regulated. These results confirm that model development requires more than just describing tourism dynamics, but also stability analysis and optimal control theory. Another research, Delpini et al. [7] examines the *predator-prey-tourist* model, which describes ecological interactions with tourist visitation pressure. This research also demonstrates a control policy in the form of an entrance fee aimed at regulating tourist numbers and protecting conservation areas. This model still uses two variables and does not include other related factors.

Several previous studies have encountered obstacles in integrating components, namely the number of tourists, economic growth, investment, and the environment simultaneously in mathematical models. Many studies focus on the relationship between the components of tourists and the environment or tourists and the economy without paying attention to the role of investment as another important component. It seems that there is still little research that shows optimal policy strategies in a single integrated mathematical model.

This research proposes an approach by developing a mathematical model with four main variables that demonstrate the interactions between tourists, the economy, investment, and environmental quality. The mathematical model is constructed using a system of nonlinear differential equations, providing logistic growth, economic relationships, tourists, investment, and environmental recovery and degradation. Furthermore, this study formulates the optimal control problem by adding two policies: environmental restoration efforts and green investment incentives. The approach used is Pontryagin's Maximum Principle.

Finally, numerical simulations have an important role to describe the dynamic system of sustainable tourism. Numerical simulation on mathematical models in the tourism sector aims

to describe four main variables and provide policy strategies to achieve a sustainable tourism system. Thus, collaboration is obtained between mathematical models, numerical simulations of mathematical models, and optimal control as a means of supporting decision-making in the sustainable tourism sector.

II. Mathematical Models in Sustainable Tourism

In this section, a nonlinear dynamic model is developed that describes the interaction between the number of tourists, economic growth, investment, and environmental quality under conditions of sustainable tourism. The model is constructed using a system of nonlinear differential equations that describe the change in each variable over time. The variables used in the mathematical model for sustainable tourism include,

$T(t)$: number of tourists (million people),
$E(t)$: economic growth (trillion rupiah ADHK),
$I(t)$: tourism sector investment (trillion rupiah),
$L(t)$: environmental quality (index 0–1).

These four variables are related to each other directly and indirectly. Increasing tourist numbers can boost revenue and investment in the tourism sector, but simultaneously can degrade environmental quality. However, when quality is low, it negatively impacts the attractiveness of a destination, reducing tourism, economic revenue, and investment. This relationship is further formulated in four equations.

A. Dynamic Model Equations

The mathematical model in sustainable tourism consists of four nonlinear differential equations that show the dynamics of four main variables. The system of differential equations is formulated as follows,

$$\begin{aligned}
\frac{dT}{dt} &= aT\left(1-\frac{T}{K}\right)+\beta_1 E-\alpha_1 T(1-L)\\
\frac{dE}{dt} &= \delta_1 T+\delta_2 I-\alpha_2 E\\
\frac{dI}{dt} &= \mu E-\alpha_3 I\\
\frac{dL}{dt} &= bL\,(1-L)-(\theta T+\gamma T I)
\end{aligned} \tag{1}$$

with,

T : number of tourists,
E : economic growth,
I : investment,
L : environmental quality,
a : intrinsic growth rate of tourists,
K : maximum tourist carrying capacity,
β_1 : impact of economic growth on tourist arrivals,
α_1 : rate of decrease in tourist arrivals due to environmental degradation,
δ_1 : contribution of tourists to economic growth,
δ_2 : contribution of investment to economic growth,
α_2 : depreciation rate of the economic sector,
μ : impact of economic growth on investment accumulation,
α_3 : depreciation rate of investment,
b : natural environmental recovery rate,
θ : impact of tourists on environmental degradation,
γ : impact of tourist–investment interaction on environmental degradation.

B. Basic Assumptions of the Model

The mathematical model in equation 1 is built based on several main assumptions that describe the dynamics of sustainable tourism. The number of tourists follows the form of logistics with the maximum tourist capacity being K and increases with economic growth, namely β_1 and decreases due to damaged environmental quality, namely α_1. Furthermore, economic growth increases along with the increasing number of tourists and investments, namely δ_1 and δ_2 but experiences natural depreciation of α_2. Meanwhile, investment in the tourism sector increases following economic growth by μ and decreases due to investment depreciation, namely α_3. Finally, environmental quality can recover naturally, namely b but decreases due to the increasing number of tourists and interactions between tourists and investments of θ and γ. These assumptions ensure that the relationships between tourist numbers, economic growth, investment and environmental quality are depicted realistically and are consistent with the basis of a sustainable tourism system.

III. Optimal control

An optimal control approach is needed to design policies to maintain the sustainability of the tourism system. The optimal control problem is formulated using Pontryagin's Maximum Principle [8] [9].

A. Optimal Control Formulation

The mathematical model in sustainable tourism was developed by adding two control variables that describe government policies in the tourism sector. Where, $u_1(t)$ is an environmental restoration effort and $u_2(t)$ represents a green investment incentive that is expected to reduce the intensity of environmental damage due to interactions between tourists and investment.

The addition of these two controls is optimized with the goal of maximizing the objective function by prioritizing tourist numbers and economic growth, while minimizing environmental degradation and policy costs. This results in a system of model equations as follows,

$$\begin{aligned}
\frac{dT}{dt} &= aT\left(1-\frac{T}{K}\right)+\beta_1 E-\alpha_1 T(1-L)\\
\frac{dE}{dt} &= \delta_1 T+\delta_2 I-\alpha_2 E\\
\frac{dI}{dt} &= \mu E-\alpha_3 I+\sigma u_2\\
\frac{dL}{dt} &= bL\,(1-L)-\left(\theta T+\gamma^{\text{eff}}(u_2)TI\right)+\kappa u_1
\end{aligned} \tag{2}$$

where, $\gamma^{\text{eff}}(u_2) = \frac{\gamma}{1+\eta u_2}$.
The objective function is formulated as follows,

$$J(T, E, L, u_1, u_2) = \int_0^{t_f} \Big[\psi_1 T + \psi_2 E - \psi_3 (1-L)^2 - \frac{c_1}{2}u_1^2 - \frac{c_2}{2}u_2^2\Big]\, dt. \tag{3}$$

The boundary conditions are given by $0 \le u_1(t) \le \bar{u_1}$ and $0 \le u_2(t) \le \bar{u_2}$.
with,

- σ: effectiveness of investment incentives,
- η: effectiveness of investment-induced greening measures,
- κ: effectiveness of environmental restoration,
- ψ_1: weight associated with tourist arrivals in the objective function,
- ψ_2: weight associated with economic growth in the objective function,
- ψ_3: weight associated with environmental degradation, represented by $(1-L)^2$,
- $\frac{c_1}{2}u_1^2$: cost of environmental restoration,
- $\frac{c_2}{2}u_2^2$: cost of green investment incentives.

B. Optimal Control Solutions

Solving the optimal control problem on the system of equations 2 using Pontryagin's Maximum Principle. The Hamiltonian function is obtained,

$$\begin{aligned}\mathcal{H} &= \mathcal{H}(T(t), E(t), I(t), L(t), u(t), \lambda) \\ &= \psi_1 T + \psi_2 E - \psi_3 (1-L)^2 - \frac{1}{2}c_1 u_1^2 - \frac{1}{2}c_2 u_2^2 \\ &\quad + \lambda_T \Big[aT\Big(1 - \frac{T}{K}\Big) + \beta_1 E - \alpha_1 T(1-L)\Big] \\ &\quad + \lambda_E \Big[\delta_1 T + \delta_2 I - \alpha_2 E\Big] \\ &\quad + \lambda_I \Big[\mu E - \alpha_3 I + \sigma u_2\Big] \\ &\quad + \lambda_L \Big[bL(1-L) - \big(\theta T + \gamma^{\text{eff}}(u_2)\, TI\big) + \kappa u_1\Big]\end{aligned} \tag{4}$$

Next, the state equations, costate equations, and stationary conditions are determined.

1) State Equations
The state equation is obtained by differentiating the Hamiltonian function $\mathcal{H}$ with respect to each λ is obtained,

$$\frac{\partial \mathcal{H}}{\partial \lambda_T} = aT\Big(1 - \frac{T}{K}\Big) + \beta_1 E - \alpha_1 T(1-L) \tag{5a}$$
$$\frac{\partial \mathcal{H}}{\partial \lambda_E} = \delta_1 T + \delta_2 I - \alpha_2 E \tag{5b}$$
$$\frac{\partial \mathcal{H}}{\partial \lambda_I} = \mu E - \alpha_3 I + \sigma u_2 \tag{5c}$$
$$\frac{\partial \mathcal{H}}{\partial \lambda_L} = bL(1-L) - \big(\theta T + \gamma^{\text{eff}}(u_2)\, TI\big) + \kappa u_1 \tag{5d}$$

2) Costate Equations
In this costate equation, the Hamiltonian function is derived with respect to each variable to obtain,

$$-\frac{\partial \mathcal{H}}{\partial T} = -\psi_1 - \lambda_T\Big(a\Big(1 - \frac{2T}{K}\Big) - \alpha_1(1-L)\Big) - \lambda_E \delta_1 + \lambda_L \theta, \tag{6a}$$
$$-\frac{\partial \mathcal{H}}{\partial E} = -\psi_2 - \lambda_T \beta_1 + \lambda_E \alpha_2 - \lambda_I \mu, \tag{6b}$$
$$-\frac{\partial \mathcal{H}}{\partial I} = -\lambda_E \delta_2 + \lambda_I \alpha_3 + \lambda_L \gamma^{\text{eff}}(u_2)\, T, \tag{6c}$$
$$-\frac{\partial \mathcal{H}}{\partial L} = -2\psi_3(1-L) - \lambda_T \alpha_1 T - \lambda_L b(1-2L). \tag{6d}$$

3) Stationary Conditions
The stationary condition is obtained by differentiating the Hamiltonian function $\mathcal{H}$ with respect to the control variable $u(t)$ as follows,

$$\frac{\partial \mathcal{H}}{\partial u_j} = 0, \qquad j = 1, 2$$

subject to the bounds $0 < u_j < U_j$. So, it is obtained,

a) For u_1

$$\begin{aligned}\frac{\partial \mathcal{H}}{\partial u_1} = -c_1 u_1 + \lambda_L \kappa &= 0 \\ -c_1 u_1 &= -\lambda_L \kappa \\ \bar{u_1}(t) &= \frac{\kappa}{c_1}\lambda_L(t)\end{aligned}$$

since $0 < u_1 < U_1$ the optimal control is obtained as follows,

$$\begin{aligned}u_1^*(t) &= \min\left(U_1, \max(0, \bar{u}_1(t))\right) \\ &= \min\left(U_1, \max(0, \frac{\kappa}{c_1}\lambda_L(t)\right)\end{aligned}$$

b) For u_2

$$\frac{\partial \mathcal{H}}{\partial u_2} = -c_2 u_2 + \lambda_I \sigma + \lambda_L \frac{\gamma\eta}{(1+\eta u_2)^2} TI$$

So that the following stationary conditions are obtained,

$$-c_2 u_2 + \lambda_I \sigma + \lambda_L \frac{\gamma\eta}{(1+\eta u_2)^2} TI = 0$$

since $0 < u_2 < U_2$ the optimal control is obtained as follows,

$$u_2^*(t) = \min\left(U_2, \max(0, \bar{u}_2(t))\right)$$

The optimal system is obtained by substituting $u^*(t)$ into the state and costate equations and then solving it numerically.

IV. Numerical Simulation Results

Numerical simulations were performed to evaluate the effectiveness of the recommended optimal control strategy. These simulations compared the uncontrolled and controlled conditions. The parameter values used in the simulation are presented in Table I.

TABLE I: State variables, control variables, and parameters used in numerical simulations

Symbol	Type	Meaning	Value / Unit
T	State variable	number of tourists	T_0
E	State variable	economic growth	E_0
I	State variable	investment level	I_0
L	State variable	environmental quality	L_0
u_1	Control variable	environmental restoration effort	$0 < u_1 < U_1$
u_2	Control variable	green investment incentive	$0 < u_2 < U_2$
a	Parameter	intrinsic tourist growth rate	0.10
K	Parameter	tourist carrying capacity	7.2
β_1	Parameter	effect of economic growth on tourism	0.30
α_1	Parameter	impact of environmental degradation on tourism	0.35
δ_1	Parameter	contribution of tourism to economic growth	1.90
δ_2	Parameter	contribution of investment to economic growth	0.16
α_2	Parameter	depreciation rate of economic sector	0.08
μ	Parameter	effect of economic growth on investment	0.09
α_3	Parameter	depreciation rate of investment	0.25
b	Parameter	natural environmental recovery rate	0.10
θ	Parameter	tourist-induced environmental degradation	0.0018
γ	Parameter	tourism–investment interaction effect	8.9×10^{-7}

The simulation is initialized with $T(0) = T_0$, $E(0) = E_0$, $I(0) = I_0$, and $L(0) = L_0$, which represent the baseline conditions of the tourism system before policy intervention. The following are the results of numerical simulations of each variable with and without control over a period of 10 years,

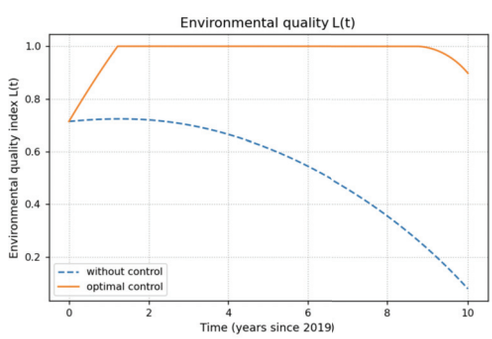
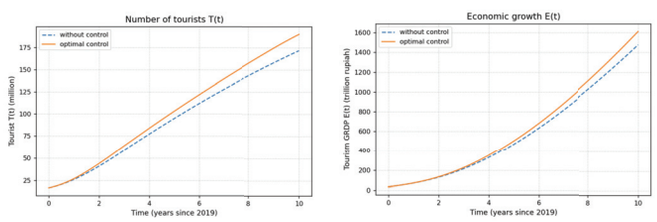
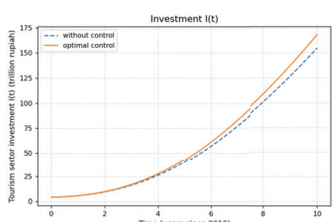


Fig. 1: Comparison of environmental quality $L(t)$ with and without optimal control.

Fig. 1 shows the environmental quality $L(t)$ with and without optimal control. In uncontrolled conditions, environmental quality tends to decline gradually due to the accumulation of tourist pressure and investment activities, which in the medium-term horizon causes significant environmental degradation. Conversely, conditions with optimal control of environmental quality can be maintained at a higher level throughout the simulation period. This indicates that optimal implementation of environmental restoration policies and green investment incentives effectively suppresses the rate of environmental degradation, despite increasing tourist numbers and economic activity. Thus, optimal control plays an important role in maintaining environmental sustainability without hindering the development of the tourism sector.



(a) Number of tourists $T(t)$ (b) Economic growth $E(t)$



(c) Investment $I(t)$

Fig. 2: Comparison of system dynamics without and with optimal control: (a) number of tourists $T(t)$, (b) economic growth $E(t)$, and (c) investment level $I(t)$.

Fig. 2 shows a comparison of dynamics system with and without optimal controls for number of tourists $T(t)$, economic growth $E(t)$, and investment $I(t)$. For all variables, the implementation of optimal controls resulted in trajectories that were consistently higher than those without controls, with the difference becoming more pronounced over time. This indicates that optimal policies through environmental management and investment incentives can simultaneously improve tourism performance, encouraging economic growth and investment without compromising system sustainability.

Overall, the numerical simulation results illustrate the implementation of optimal control that can improve the sustainable tourism system compared to conditions without control. The policy is implemented optimally and not only encourages tourism, economic and investment growth but can also maintain sustainable environmental quality in the long term.

V. Conclusion

This study concludes that the developed mathematical model of sustainable tourism, combined with the implementation of optimal control, is effective in balancing economic and

environmental objectives. The analysis and numerical simulation results indicate that the optimal policy can simultaneously increase tourist numbers, economic growth, and investment, while maintaining long-term environmental quality. Thus, the optimal control approach provides a relevant and applicable policy framework for sustainable tourism management.

ACKNOWLEDGMENT

The author would like to thank the Sepuluh Nopember Institute of Technology (ITS) for its institutional support in conducting this research. He also thanks the Supervisor and Co-Supervisor for their invaluable guidance, direction, and scientific discussions throughout the research process.

REFERENCES

[1] UNWTO and UNEP, "Tourism and the Sustainable Development Goals: Journey to 2030," World Tourism Organization, Madrid, 2019.

[2] R. Casagrandi and S. Rinaldi, "A theoretical approach to tourism sustainability," *Conservation Ecology*, vol. 6, no. 1, p. 13, 2002.

[3] S. Gössling, D. Scott, and C. M. Hall, "Global trends in length of stay: Implications for destination management and climate change," *Journal of Sustainable Tourism*, vol. 26, no. 12, pp. 2087–2101, 2018, doi: 10.1080/09669582.2018.1529771.

[4] H. Seraphin, P. Sheeran, and A. Yallop, "Assessing the vulnerability of tourist destinations to overtourism: A multidimensional framework," *Tourism Economics*, vol. 30, no. 1, pp. 112–130, 2024, doi: 10.1177/13548166251349240.

[5] H. B. Hjálmarsdóttir and G. K. Óskarsson, "Navigating growth and sustainability: Analysing the economic impact of tourism in Iceland," *Tourism and Hospitality*, vol. 6, no. 2, p. 119, 2025, doi: 10.3390/tourhosp6020119.

[6] F. Ying, "The dynamic analysis and mathematic modeling for regional tourism system," in *Proc. 2014 Sixth Int. Conf. Measuring Technology and Mechatronics Automation (ICMTMA)*, 2014, pp. 696–699, doi: 10.1109/ICMTMA.2014.172.

[7] D. Delpini, R. Melis, and P. Russu, "Dynamical analysis of a prey–predator–tourist model: Environmental preferences and optimal fee control," *Chaos, Solitons & Fractals*, vol. 191, p. 115845, 2025, doi: 10.1016/j.chaos.2024.115845.

[8] S. Lenhart and J. T. Workman, *Optimal Control Applied to Biological Models*, Chapman & Hall/CRC, Boca Raton, FL, 2007.

[9] D. S. Naidu, *Optimal Control Systems*, CRC Press, Boca Raton, FL, 2002.

GAT-Assisted Risk-Aware Isolation in a Bio-Inspired IoT Botnet Defense Simulation

Vaishali Ravi, Mohd Anuaruddin Ahmadon
and Shakirah Mohd Taib
Department of Computing
Universiti Teknologi PETRONAS
Seri Iskandar, Malaysia
vaishali_24007113@utp.edu.my,
.ahmadon@utp.edu.my, shakita@utp.edu.my

Iswardi Yuzrin, Che Azizi Che Rus
and Ahmad Hafiz Ghadzali
Department of Digital Innovation & Technology
Universiti Teknologi PETRONAS
Seri Iskandar, Malaysia
iswardi.yuzrin@utp.edu.my, cheazizi_cherus@utp.edu.my,
ahmadhafiz@utp.edu.my

***Abstract*—The scale of Internet of Things (IoT) networks is continuing to be difficult to contain when it comes to botnet propagation due to high connectivity density and increased lateral propagation of malicious traffic over large-scale networks. This paper is based on the Myco-Barrier deceptive demilitarized zone (DMZ) architecture and suggests a Graph Attention Network (GAT)-enhanced isolation mechanism that could be used to predictively isolate botnets. In contrast to traditional reactive defenses, the proposed GAT mechanism uses inductive learning to predict high-risk nodes on dynamic network topologies, using features such as resource utilization and trust scores. Despite a high degree of class imbalance in training, the model achieves a prediction accuracy of 99.36% and an ROC-AUC of 0.98, hence successfully differentiating between benign and malignant propagation of stress. When implemented in Myco-Barrier as an isolation stimulus trigger, the GAT-enhanced defense significantly reduces the rate of peak infection, and it increases service availability compared to reactive baseline approaches. These results demonstrate the usefulness of graph neural networks in facilitating adaptive, topology-based security controls of an IoT network.**

***Index Terms*—bio-inspired, botnet, IoT security, isolation, GAT**

I. Introduction

The high density of connectivity and resource restrictions of modern Internet of Things (IoT) settings are an adequate environment facilitating a rapid spread of botnets. These weak links are used by adversaries to build large swarms which can be used to cause distributed denial-of-service (DDoS) attacks and lateral flows and then later bypass perimeter defenses which cannot keep up with the scale and dynamic traffic patterns of heterogeneous networks. Although moving target defense (MTD) and deceptive environments are promising, even simple isolation policies, like those based on static isolation, can hardly contain an outbreak in an infrastructure without substantially undermining authentic service provision [1].

Myco-Barrier architecture was first introduced as the bio-inspired architecture of the defense, using the demilitarized zone (DMZ) scheme to establish the separation between the public Internet and the assets that are being safeguarded [2]. This defence system has been inspired by fungal mycorrhizal networks (FMN) and the blood-brain barrier (BBB), that decentralize signalling in order to isolate suspicious nodes. But the use of fixed heuristic thresholds showed some fatal weaknesses: the defense was fundamentally reactive, and they only initiated isolation once compromised, and could not utilize the more detailed topological relationships, such as bridge nodes, that support the spread of botnets.

To address these deficiencies, the current paper incorporates Graph Attention Networks (GAT) into the Myco-Barrier decision loop, which will convert the system into an anticipatory and topologically focused defense as opposed to a reactive one. Graph neural networks (GNNs) exploit the natural graph topology of the IoT network, unlike traditional deep learning models, which consider devices in isolation, to learn the relationship between risk propagation via connectivity. GATs are also used due to their inductive learning ability, which also allows risk forecasting of dynamically growing proxy nodes without the need to retrain them [3].

This study suggests a closed loop structure where a Python simulator is able to produce snapshots of labelled graphs under varying attack scenarios encompassing command-and-control (C2) and peer-to-peer (P2P) topologies. Different snapshots are fed into a GAT model to predict infection probability in the future given local features and neighbour states. In the process of running, this risk score activates pre-emptive isolation which is the concept of removing attack pathways before they fully materialise.

The main contributions of this work are:

- **Dynamic topology learning by inductive graph learning:** Implementation of a GAT-based risk predictor that can inductively learn to improve risk assessment.
- **Anticipatory supervision approach:** Suggestion of a future-based branding scheme which learns topological antecedents to infection and shifts the defense paradigm of reactionary containment to proactive segregation.
- **Class imbalance Resilience:** It is shown that the models are effective in a highly imbalanced scenarios, as in the case when the combination of graph normalisation and weighted loss functions leads to an ROC-AUC of 0.98.

II. Related Works

This study explores the existing literature related to application of techniques for intrusion detection and isolation mechanism.

A. Graph Neural Networks for IoT Intrusion Detection

The integration of Graph Neural Networks in the context of IoT security has gained significance faster due to the ability to encode non-Euclidean dependencies between devices. The recent studies have shown that GNNs outperform outdated classifiers, including Random Forests, by modeling network traffic in form of graph structures. As an example, hybrid networks combining Graph Convolutional Networks (GCNs) with the solution proposed in [4] and attention-based models (GAT) [5] have demonstrated high detection rates on test data sets such as CICIoT-2023 and BoT-IoT. However, most of these paradigms work in transductive regime where the whole graph structure is accessible when making inferences [3], [6]. This limitation makes them inappropriate in dynamic defence space where network topology is changing as operation proceeds.

B. Moving Target Defense in IoT

Moving Target Defense is a set of strategies that should neutralize reconnaissance by the attacker by dynamically changing the configuration of systems. The MTD surveys carried out in [1] elaborate methods such as IP shuffling, port hopping and protocol switching that are used to increase the complexity of possible attacks. Although suitable in breaking the kill chain between statics, modern MTD models generally rely on randomized or schedule-based events (e.g., shuffling after every N seconds) rather than risk-based decision making [7]. In addition, existing MTD strategies are not topology sensitive; they randomly rearrange features without considering the structural importance of a node [8].

C. Bio-Inspired Cyber Defense

Biological systems have served as inspiration in the creation of decentralized security paradigm. Studies dedicated to Artificial Immune Systems (AIS) [9] and Swarm Intelligence have produced powerful algorithms of anomaly detection and distributed patching [10]. Such systems mimic the adaptability of biological systems; but often they make use of shallow learning heuristics and not deep representation learning [10].

Even though bio-inspired algorithms can be used to facilitate decentralised resilience, they do not generally have the predictive acuity that modern deep-learning methods provide. As a result, a paucity of literature that combines the regenerative rationale of biological networks with the structural aspect extraction potential of GNNs, persistently exists. Existing solutions address IoT defense in isolation: GNN-based detectors are often transductive and limited to static graphs; MTD frameworks typically lack risk-aware triggers; and bio-inspired mechanisms often lack predictive depth. This work bridges these gaps by introducing an inductive, topology-aware GAT within the Myco-Barrier architecture. Unlike prior approaches, our model learns generalized risk patterns that persist across dynamic topological mutations, enabling a defense strategy that is both structurally adaptive and predictively robust.

III. Methodology

This section describes how a Graph Attention Network is trained on synthetic botnet traces from the Myco-Barrier simulator and later embedded as a risk oracle inside the defense loop. We formulate anticipatory botnet detection as a node-level binary classification problem: given the current graph snapshot and node features, the model predicts whether a node is likely to become infected within a short future horizon.

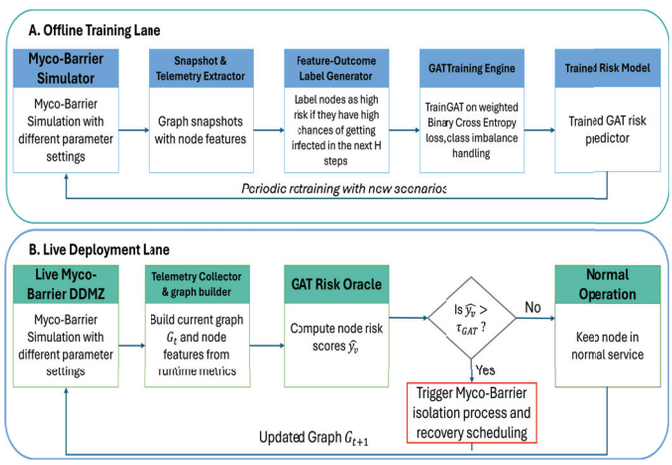


Fig. 1. System overview. Offline: the Myco-Barrier simulator generates labeled graph snapshots to train the GAT risk model. Live: the trained GAT continuously scores nodes in the live Myco-Barrier defense layer and triggers bio-inspired isolation and proxy spawning.

A. Overview of the GAT-Assisted Myco-Barrier

Fig. 1 illustrates the complete pipeline. The process has two phases:

1) Offline training phase. The Python-based Myco-Barrier simulator is executed under diverse botnet scenarios (different topologies, sizes, and infection aggressiveness). At each simulation step, the current IoT graph G_t and node attributes are captured, and the future evolution of each node is tracked over a horizon H. These snapshots are converted into graph samples for GAT training.

2) Live deployment phase. Once trained, the GAT model is embedded back into the simulator. At every defense cycle, the current graph G_t is fed to the GAT, which outputs a risk probability $\hat{y}_v$ for each node v. Nodes with $\hat{y}_v \geq \tau$ are treated as *high-risk* and immediately passed to the Myco-Barrier isolation logic (first-degree quarantine and optional proxy insertion), even if they have not yet been observed as infected.

Fig. 2 shows the local view around a target node. Rather than averaging all neighbors, the GAT computes an attention weight α_{ij} for each neighbor j, indicating how relevant that neighbor is to node i's future compromise. Warm colors (yellow/red) in the legend correspond to high attention magnitude

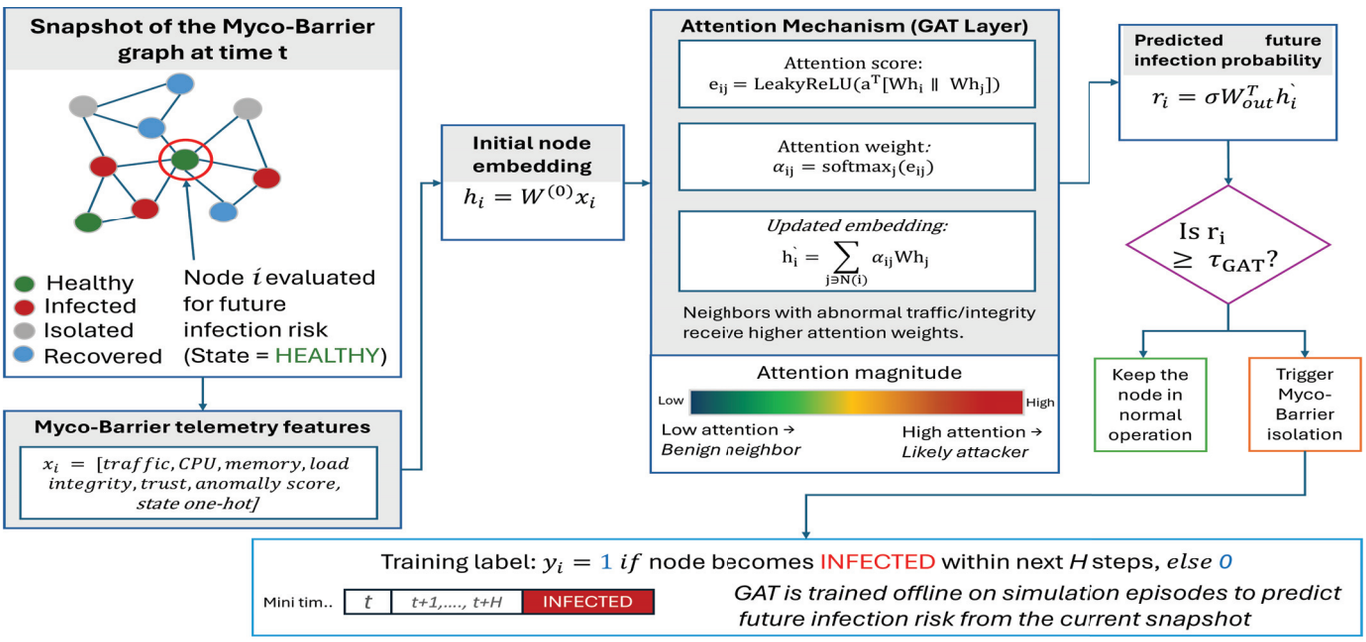


Fig. 2. GAT-based node risk estimation and training pipeline in Myco-Barrier. A graph snapshot at time t is encoded into telemetry features x_i, passed through the GAT attention layer to obtain a risk score $\hat{y}_i$, and compared against a threshold τ_{GAT} to trigger anticipatory isolation. The same snapshot is paired with a future-outcome label y_i (infection within the next H steps) for offline training.

and typically reflect suspicious neighbors (e.g., high traffic, low integrity). The final embedding h_i' is a weighted sum of these neighbor features and is used to predict $\hat{y}_i$.

B. Problem Formulation

We model the IoT defense surface at time t as a directed graph

$$G_t = (V_t, E_t), \tag{1}$$

where V_t is the set of nodes (core routers, edge gateways, end devices, and proxy nodes) and E_t is the set of communication links. Each node $v \in V_t$ is associated with a feature vector $x_v^{(t)} \in \mathbb{R}^F$.

The GAT learns a mapping

$$f : (G_t, x_v^{(t)}) \to \hat{y}_v \in [0, 1], \tag{2}$$

where $\hat{y}_v$ denotes the predicted probability that node v will become infected within the next H steps. During deployment, $\hat{y}_v$ is compared to a decision threshold τ to trigger anticipatory isolation.

C. Dynamic Dataset Generation with Myco-Barrier Simulator

To avoid relying on static attack traces, we generate a large, labeled graph corpus using the Myco-Barrier simulator. Each run of the simulator produces a sequence of graphs $\{G_t\}_{t=0}^{T}$ with evolving node states.

The dataset covers a range of operating conditions:

- **Botnet architectures:** Centralized Command-and-Control (C2) and Peer-to-Peer (P2P).
- **Network scales:** $|V_t| \in \{50, 100, 150\}$ nodes.
- **Attack aggressiveness:** infection probabilities $p_{\text{inf}} \in \{0.25, 0.35, 0.50\}$ and initial infection ratios $\rho_0 \in \{0.25, 0.30, 0.40\}$.

For each scenario, the simulator runs for $T = 60$ steps. At every step t, we record the current graph G_t and then look ahead over a horizon H (e.g., $H = 3$) to see which nodes become infected. This produces a set of labeled samples $(G_t, x^{(t)}, y^{(t)})$ that jointly capture resource dynamics, botnet spreading, and mitigation responses.

D. Node Feature Representation and Future-Outcome Labeling

The feature vector for node v at time t combines resource usage, local security indicators, and structural context:

$$x_v^{(t)} = \left[\mathcal{R}_v^{(t)} \parallel \mathcal{S}_v^{(t)} \parallel \mathbf{1}_{\text{type}} \parallel \mathbf{1}_{\text{state}}\right], \tag{3}$$

where

- $\mathcal{R}_v^{(t)} \in \mathbb{R}^4$ captures CPU utilization, memory usage, traffic volume, and composite capacity/load.
- $\mathcal{S}_v^{(t)} \in \mathbb{R}^3$ encodes integrity score, trust score from neighbors, and anomaly score.
- $\mathbf{1}_{\text{type}}$ and $\mathbf{1}_{\text{state}}$ are one-hot encodings of the node's layer (core, edge, device, proxy) and operational state.

Instead of labeling nodes by their *current* infection status, we use a future-outcome label:

$$y_v^{(t)} = \mathbb{I}\big(\exists k \in [1, H] \text{ such that } \text{State}_v(t + k) = \text{INFECTED}\big), \tag{4}$$

where $\mathbb{I}(\cdot)$ is the indicator function. Nodes that are currently healthy but become infected within the horizon are labeled as positive examples, enabling the GAT to learn patterns that

precede compromise. Nodes already infected at time t are excluded from training via a mask to prevent trivial labels.

E. Graph Attention Architecture

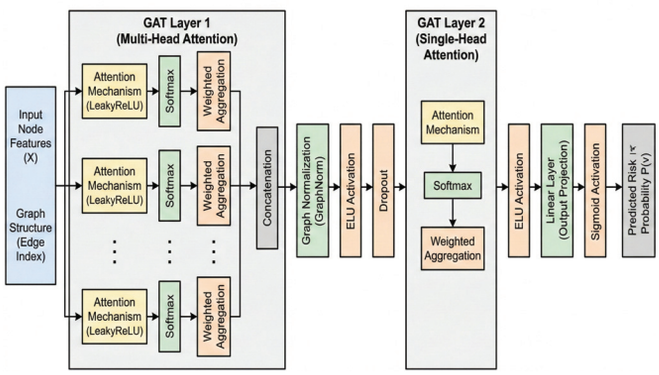


Fig. 3. Two-layer GAT used for node-level risk prediction. Multi-head attention aggregates neighbor signals into a risk-aware embedding, followed by a sigmoid output for binary classification.

To handle the dynamic graph structure (nodes join/leave due to proxy creation and isolation), we adopt an inductive two-layer GAT, shown in Fig. 3. For each layer l, a shared linear transformation $\mathbf{W}^{(l)}$ is applied to node features, and the raw attention between node i and neighbor j is

$$e_{ij}^{(l)} = \text{LeakyReLU}\Big(\mathbf{a}^{(l)T}\big[\mathbf{W}^{(l)}h_i^{(l-1)} \,\|\, \mathbf{W}^{(l)}h_j^{(l-1)}\big]\Big), \quad (5)$$

where $\mathbf{a}^{(l)}$ is a learnable attention vector and $h_i^{(0)} = x_i$. These scores are normalized via softmax over the neighborhood $\mathcal{N}(i)$:

$$\alpha_{ij}^{(l)} = \frac{\exp\left(e_{ij}^{(l)}\right)}{\sum_{k\in\mathcal{N}(i)}\exp\left(e_{ik}^{(l)}\right)}. \quad (6)$$

The updated embedding for node i is

$$h_i^{(l)} = \sigma\left(\sum_{j\in\mathcal{N}(i)}\alpha_{ij}^{(l)}\mathbf{W}^{(l)}h_j^{(l-1)}\right), \quad (7)$$

with multi-head attention in the first layer and a single head in the second layer. To improve separability between noisy yet benign nodes and truly compromised nodes, we apply GraphNorm and Dropout ($p = 0.5$) after the first layer.

The final layer produces a scalar logit z_v for each node, which is converted into a probability $\hat{y}_v = \sigma(z_v)$ using the sigmoid function.

F. Loss Function, Class Imbalance and Optimization

Because most nodes in the simulator remain healthy, the dataset is highly imbalanced (Safe $\gg$ High-risk). To prevent the model from collapsing to the majority class, we employ a weighted binary cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N}\sum_{v\in V}\big[w_{\text{pos}}\, y_v \log(\hat{y}_v) + (1-y_v)\log(1-\hat{y}_v)\big], \quad (8)$$

where $w_{\text{pos}} = N_{\text{safe}}/N_{\text{risk}}$ up-weights the contribution of high-risk nodes. In our experiments, w_{pos} is approximately 4.

The model is optimized using Adam with learning rate $\eta = 10^{-3}$ for 200–300 epochs. A validation split is used to monitor convergence via loss and ROC-AUC. At inference time, we use a risk threshold τ tuned to favor high recall on high-risk nodes, and the resulting $\hat{y}_v \geq \tau$ events are fed directly into the Myco-Barrier isolation and proxy-creation routines.

IV. Results and Discussion

This section evaluates the performance of the proposed GAT-assisted isolation mechanism. We first analyze the predictive capability of the GAT on the dynamic Myco-Barrier dataset, and then compare the closed-loop behavior of the GAT-assisted defense against the heuristic-only baseline in terms of infection containment and service availability.

A. Experimental Setup

The GAT model was trained on a dataset generated by the Myco-Barrier simulator, comprising multiple infection scenarios across varying topologies (Hybrid, C2, and P2P) and network sizes ($N \in \{50, 100, 150\}$). The resulting corpus contains 333,198 node-level samples. The class distribution is highly imbalanced, with benign nodes outnumbering high-risk nodes by approximately 200:1, reflecting a realistic IoT setting where most devices remain uncompromised. The validation set preserves this imbalance to accurately assess performance under deployment conditions.

B. Training Convergence

Fig. 4 illustrates the training and validation loss over 300 epochs. Both curves decrease rapidly before stabilizing, with no significant divergence between training and validation loss. This behavior indicates that the GAT learns meaningful patterns from the synthetic traces without overfitting to particular graph instances or topologies, which is a common risk in transductive GNN setups.

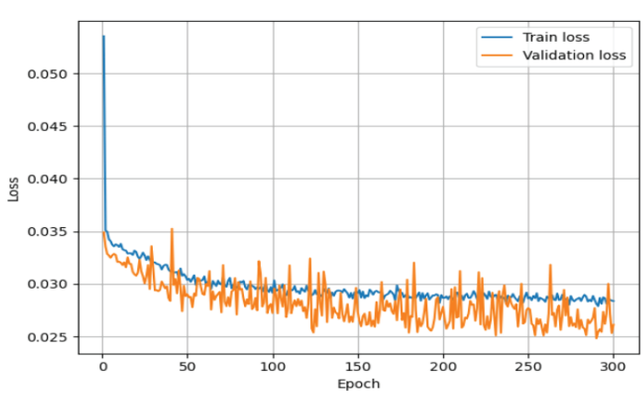


Fig. 4. Training and validation loss of the GAT risk predictor over 300 epochs. Both curves converge and remain close, indicating good generalization without overfitting.

C. Predictive Performance

The node-level classification performance of the trained GAT model is summarized in Table I. The model attains an overall accuracy of 99.36%, highlighting its stability in distinguishing normal traffic stress from emerging botnet propagation.

TABLE I
PERFORMANCE METRICS OF GAT RISK PREDICTOR

Metric	Value
Accuracy	0.9936
ROC–AUC	0.9826
Recall (Class 1: High-risk)	0.7486
Precision (Class 1: High-risk)	0.4050
F1-score (Class 1)	0.5256

Fig. 5 reports the ROC curve, with an AUC of 0.983. The curve closely follows the upper-left boundary, confirming that the GAT reliably ranks truly high-risk nodes above benign nodes in most pairwise comparisons.

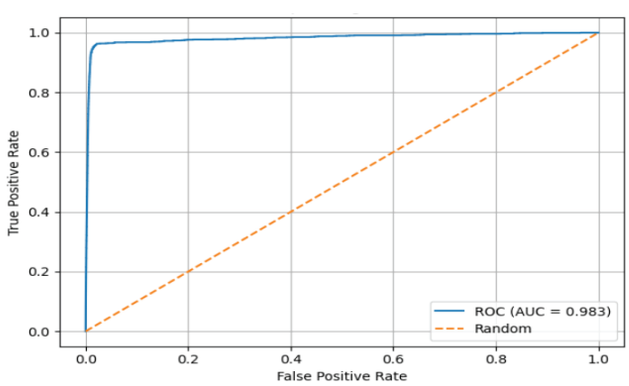


Fig. 5. ROC curve of the GAT risk predictor. The AUC of 0.983 confirms strong discriminative power.

At the operating threshold $\tau = 0.6$, the model achieves a Recall of 74.86% on the high-risk class, meaning that nearly three-quarters of impending infections are anticipated before they fully manifest. The Precision of 40.50% (and PR–AUC of 0.444) indicates that some benign nodes are also flagged as high-risk. However, in the deceptive DDMZ context, the cost of isolating a decoy node is significantly lower than the cost of missing an actual compromise. The class-weighted loss (with $w_{\text{pos}} \approx 4$) deliberately biases the model towards high recall on the high-risk class.

D. Inductive Learning Capability

A key requirement for Myco-Barrier is the ability to score nodes that did not exist during initialization (e.g., dynamically spawned proxies). Fig. 8 shows the empirical distribution of predicted risk scores $P(v)$. Benign nodes are strongly concentrated near $P(v) \approx 0$, while high-risk nodes form a secondary mode above $P(v) \geq 0.6$. This separation indicates that the GAT does not memorize node identifiers; instead, it learns structural and behavioral motifs (such as "high-traffic node with multiple infected neighbors") that transfer to new nodes and altered topologies.

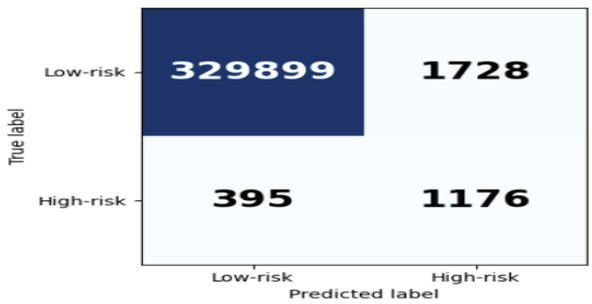


Fig. 6. Confusion matrix for the GAT risk predictor. Out of 1,571 high-risk nodes, 1,176 are correctly identified, with 1,728 false alarms among approximately 330k benign nodes.

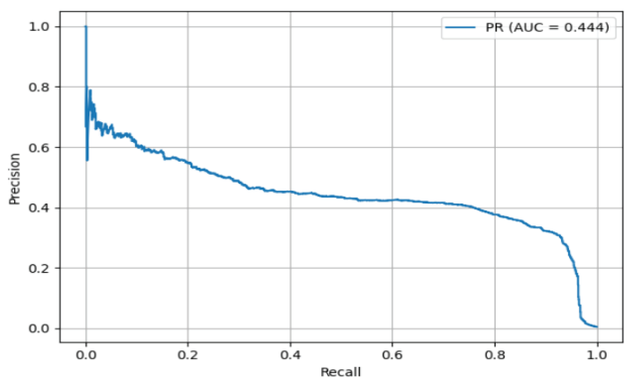


Fig. 7. Precision–recall curve for the high-risk class on the validation set (PR-AUC = 0.444). Precision gradually decreases as recall approaches 1.0 due to the strong class imbalance between benign and high-risk nodes.

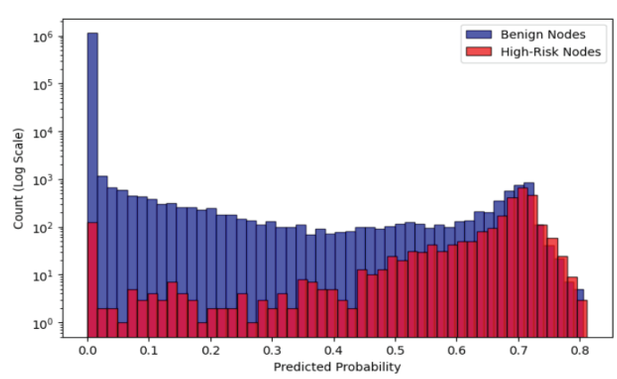


Fig. 8. Log-scale probability distribution of GAT-predicted infection risk for benign nodes (blue) and high-risk nodes (red) on the validation set. Benign nodes concentrate near $P(v) \approx 0$, while high-risk nodes form a separate mode around $P(v) \approx 0.6$–0.8.

E. Operational Impact

From the confusion matrix in Fig. 6, the model anticipates 1,176 out of 1,571 actual infection events, while 395 remain undetected. These missed events typically correspond to early, stealthy propagation stages where local signals are weak. At

the same time, the true-negative rate exceeds 99.4%, so the vast majority of benign nodes are correctly left in normal operation. This balance supports Myco-Barrier's objective of maintaining high service availability while still providing aggressive containment of emerging botnet activity.

F. Threshold Sensitivity

Because the probability distribution in Fig. 8 exhibits a clear gap between the bulk of benign nodes and the high-risk tail, the decision threshold τ can be tuned to trade precision for recall. Lowering τ from 0.6 to 0.4 substantially increases recall (more early-stage infections are captured) at the cost of a moderate increase in false positives. This tunability allows operators to choose a more conservative or more aggressive operating point depending on current threat levels.

G. Comparison with Baseline Myco-Barrier

To assess the impact of GAT integration on end-to-end defense, we compare the proposed framework against the original heuristic-only Myco-Barrier in closed-loop simulations ($N = 150$, C2 topology).

1) Infection Containment: Fig. 9 shows the time evolution of infected nodes. The heuristic-only defense (orange) exhibits repeated infection spikes, indicating that isolation is triggered only after significant damage has already occurred. In contrast, the GAT-assisted defense (green) rapidly drives the infection curve towards zero and suppresses secondary waves of lateral movement by isolating bridge nodes pre-emptively.

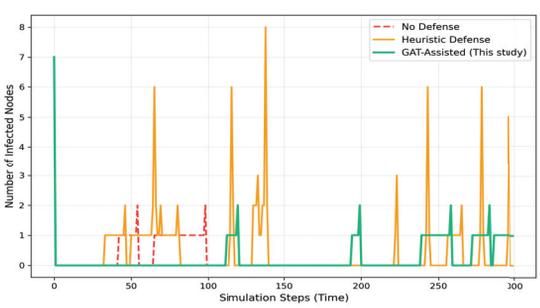


Fig. 9. Infection containment dynamics. The GAT-assisted defense (green) suppresses secondary infection waves compared to the heuristic baseline (orange).

2) Service Availability: Fig. 10 reports the number of operational (non-isolated) nodes over time. The heuristic baseline exhibits periods of over-isolation during high-traffic phases, where benign overload is misinterpreted as malicious behavior. The GAT-assisted defense (blue) maintains a consistently higher level of service availability by selectively isolating only those nodes with high predicted risk, leveraging its strong discriminative performance (ROC–AUC 0.98) to minimize collateral damage.

V. Conclusion and Future Work

This paper introduces a GAT-based Myco-Barrier defense system to curb the activity of IoT botnets, shifting the emphasis away towards reactive and rule-based detection tools

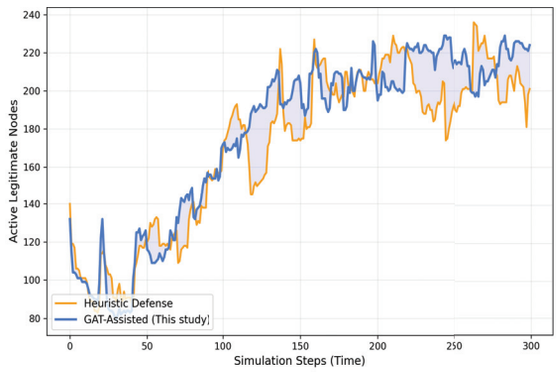


Fig. 10. Service availability over time. The GAT-assisted approach (blue) avoids over-isolation and preserves more legitimate nodes than the heuristic baseline.

to forward-looking risk prediction. The proposed model can detect high-risk nodes in dynamic topologies with an accuracy of 99.36% and an ROC-AUC of 0.98, by using inductive graph attention. The findings highlight the effectiveness of graph neural networks in designing high-level isolation policy which is dynamic and topology-aware, and which blocks most of the attacks before they can propagate at a large scale without sacrificing a high false-alarm rate. In the future, this framework will be improved by incorporation of reinforcement-learning-based policy modification and testing its functionality on hardware testbeds and real-world IoT implementations.

References

[1] R. Sun, Y. Zhu, J. Fei, and X. Chen, "A survey on moving target defense: Intelligently affordable, optimized and self-adaptive," *Appl. Sci. (Basel)*, vol. 13, no. 9, p. 5367, Apr. 2023.

[2] V. Ravi, M. A. Ahmadon, S. Shuhidan, and S. Yamaguchi, "Bionexus defense: The myco-barrier mechanism for adaptive IoT security," in *Proc. 2025 IEEE Global Conference on Consumer Electronics (GCCE)*, Osaka, Japan, Sep. 2025, to be published.

[3] A. S. Ahanger, S. M. Khan, F. Masoodi, and A. O. Salau, "Advanced intrusion detection in internet of things using graph attention networks," *Sci. Rep.*, vol. 15, no. 1, p. 9831, Mar. 2025.

[4] O. Ceran, E. Özdoğan, and M. Uysal, "Leveraging graph neural networks for IoT attack detection," *Sakarya Univ. J. Comput. Inform. Sci.*, vol. 8, no. 2, pp. 223–244, Jun. 2025.

[5] M. Zhong, M. Lin, C. Zhang, and Z. Xu, "A survey on graph neural networks for intrusion detection systems: Methods, trends and challenges," *Comput. Secur.*, vol. 141, no. 103821, p. 103821, Jun. 2024.

[6] T. Ngo, J. Yin, Y.-F. Ge, and H. Wang, "Optimizing IoT intrusion detection—a graph neural network approach with attribute-based graph construction," *Information (Basel)*, vol. 16, no. 6, p. 499, Jun. 2025.

[7] Swati, S. Roy, J. Singh, and J. Mathew, "Securing iiot systems against ddos attacks with adaptive moving target defense strategies," *Scientific Reports*, vol. 15, 2025.

[8] D. P. Sharma, "Evaluating moving target defense methods using time to compromise and security risk metrics in iot networks," *Electronics*, 2025.

[9] I. E. Agbehadji, A. Abayomi, O. J. Aroba, E. Freeman, and R. N. Nketsiah, "Bio-inspired computational intelligence (BioCOIN) for cyber security during COVID-19 pandemic: A bibliometric review," *ijirss*, vol. 8, no. 7, pp. 418–441, Oct. 2025.

[10] K. T. Chui, R. W. Liu, M. Zhao, and X. Zhang, "Bio-inspired algorithms for cybersecurity - a review of the state-of-the-art and challenges," *Int. J. Bio Inspired Comput.*, vol. 23, pp. 1–15, 2024.

Integrating Motivation and Organizational Context: A Conceptual Model for Gamification-Based Cybersecurity Awareness in SMEs

1st Qinthara Putrie Zuvianne
School of Economic and Business
Telkom University
Bandung, Indonesia
qintharapz@student.telkomuniversity.ac.id

2nd Puspita Kencana Sari
School of Economic and Business
Telkom University
Bandung, Indonesia
puspitakencana@telkomuniversity.ac.id

***Abstract*— Gamification offers a promising approach for enhancing cybersecurity awareness in small and medium enterprises (SMEs), which often struggle with limited resources and varying employee digital literacy levels. Through a narrative review, existing literature on gamification, motivation, and organizational learning is synthesized. The analysis identifies Self-Determination Theory (SDT), Theory of Planned Behavior (TPB), and Technology-Organization-Environment (TOE) Framework as key theoretical foundations for understanding user motivation and behavioral intention in gamified training. Findings reveal that leaderboards, challenges, points, badges, feedback, and narratives are the most effective in boosting engagement and self-efficacy. However, existing research predominantly focuses on large organizations, indicating a gap in gamification applications for SMEs. A conceptual model integrating Self-Determination Theory, Theory of Planned Behavior, and Technology-Organization-Environment organizational factors is proposed to guide gamified cybersecurity awareness training design in small and medium enterprises context. The framework provides direction for future empirical research and practical implications for developing more engaging training programs.**

Keywords*— *Gamification, cybersecurity, awareness, training, small and medium enterprises, employee motivation

I. INTRODUCTION

The lack of cybersecurity awareness among employees remains one of the most persistent challenges organizations face in mitigating evolving digital threats, such as social engineering attacks, data breaches, and human errors [1]. Traditional cybersecurity learning has shown limited effectiveness in fostering sustained motivation and long-term behavioral change among employees [2]. Organizations are increasingly interested in exploring more interactive and motivational approaches, including the incorporation of gamification elements within cybersecurity training programs [3]. Meta-analytic evidence showed that gamification can increase learner engagement and lead to better knowledge retention in professional training environments[4].

It has been found that gamification increases employees' intrinsic motivation for organizational learning by satisfying basic psychological needs for competence, autonomy, and relatedness [5]. Previous research also proved that it affects creativity and commitment to work in an intrinsically motivational way, and Capatina (2024) further underlined that intrinsic motivation may be increased by gamified elements used in corporate training, which enhances the motivation of employees [6]. However, in spite of all these positive findings, the practice of gamification in organizations is very diverse, and there does not seem to be a consensus on what game elements are most effective [7].

Self-Determination Theory (SDT) served as a foundational framework in existing literature for understanding how the satisfaction of psychological needs drives intrinsic motivation and cultivates positive behavioral outcomes in the workplace [8]. Research findings demonstrate that need-fulfillment strategies can enhance employee involvement and conformity to security measures in cybersecurity domains. In parallel, the Theory of Planned Behavior (TPB) is widely utilized to analyze the formation of employees' intentions toward security policy adherence through the interplay of attitudes, subjective norms, and perceived behavioral control [9]. The Technology-Organization-Environment (TOE) Framework complements this understanding by emphasizing the role of organizational resources, environmental factors, and technological capabilities in shaping security behavior adoption and cybersecurity policy effectiveness within workplace environments [10].

Gamification has been recognized as an approach that can shift employees' perceptions of cybersecurity training from mandatory task to a more engaging and meaningful learning experience. Empirical studies show that integration elements like leaderboards, badges, and point systems enhance employee engagement and focus and responsible behavioral intentions. Bitrián (2023) shows that integrating game-based components in e-learning environments enhances attitudes toward training, therefore reliably reinforcing the effectiveness of training on shaping desirable security behaviors [11].

Gamification represents an innovative strategy that might help cybersecurity training transcend from a necessary evil to a more interesting and thus motivating means of learning. Elements entailing points, badges, and leaderboards have been found to increase participant engagement and facilitate responsible behavioral intentions in the workplace, based on a growing body of research on gamified cybersecurity training [12]. However, with the exception of a few studies, most of this has been conducted in educational institutions or large organizations, while relatively few studies have focused specifically on employees in small or medium-sized enterprises [13].

Small and medium-sized enterprises (SMEs), whose characteristics and challenges differ substantially from those of large organizations, constitute the focus of this study. SMEs often face limited formal security policies, gaps in employees'

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

IT literacy and constraints in financial and technical resources [14]. Prior research on cybersecurity in SMEs, indicate that low motivation and the absence of strong security policies, even when they possess sufficient awareness [15]. Although gamification has been introduced in some SME-oriented cybersecurity training programs as a motivational strategy, empirical research examining its effectiveness particularly in influencing employee motivation and awareness remains scarce.

Based on this background and gap, this study proposes three research question as follows:

RQ1. How is gamification used in employee-focused cybersecurity training in organizations?

RQ2. Which elements of gamification have the most influence on employee motivation factors in cybersecurity training?

RQ3. How does conceptual model apply gamification to increase cybersecurity awareness in the SMEs sector?

This research is proposed because, despite evidence that shows gamification to be effective in enhancing motivation and cybersecurity awareness training in large organizations, its application within SMEs remains limited. By integrating the analysis of gamification elements with employee motivation factors in the SME context, this study intends to contribute theoretically to the understanding of motivational mechanism in gamified cybersecurity training and practically by offering design recommendations for training program tailored to SME employees, thereby supporting the development of a more resilient and motivated cybersecurity culture.

II. UNDERSTANDING THEORIES

A. Self-Determination Theory

Deci and Ryan introduced the Self-Determination Theory (SDT), which was a conceptual framework explaining the mechanisms through which motivation arises and influences both behavioral patterns and psychological well-being. According to this theory, autonomous kinds of motivation are usually more effective than controlled ones in yielding continuous engagement and positive behavior [16]. SDT posits three fundamental psychological needs such as autonomy, competence, and relatedness that serve as the cornerstone of intrinsic motivation [17]. When these needs are satisfied, engagement becomes higher, while their lack triggers negative effects on motivational processes [18]. For the purposes of this study, these three constructs are measured as perceived autonomy, perceived competence, and perceived relatedness, capturing employees' sense of freedom in decision-making, belief in their capabilities, and feelings of social connection within the organizational context.

B. Theory of Planned Behavior

Icek Ajzen proposed the Theory of Planned Behavior (TPB) by extending the Theory of Reasoned Action for better representation of how individuals develop behavioral intentions and subsequently translate them into observable actions [19]. According to TPB, the formation of behavioral intention depends upon three main components: attitude toward the behavior, representing an individual's perception about the expected consequences involved in performing that behavior; subjective norms, grasping the influence of perceived expectations from significant referent groups; and perceived behavioral control, reflecting one's confidence about possessing the required capabilities and resources to perform the behavior [20], [21]. Furthermore, perceived behavioral control can directly influence actual behavior whenever individuals continue to perceive a sense of control over the accomplishment of their actions [22].

C. TOE Framework

The Technology–Organization–Environment (TOE) Framework states that technological innovations are adopted and implemented in organizations based on three important dimensions, namely technological, organizational, and environmental factors [23]. Ever since it was developed by Tornatzky and Fleischer, the TOE framework has found significant application in information systems research, within contexts such as cloud computing adoption among small and medium enterprises in Bangladesh [24], and e-retailing implementation in Vietnam [25]. This study focuses on the organizational dimension, more specifically investigating manager innovativeness and organizational readiness, which capture leadership capacity for innovation and the sufficiency of organizational resources, respectively. This framework provides a structured means of identifying the determinants that will shape technology adoption processes within organizations.

III. METHODOLOGY

This study follows the approach of a narrative review, which identifies, analyses, and synthesizes different research outcomes related to the subject under review. Materials were retrieved from the Scopus database within the period from 2020 to 2025, using the keyword (“gamification”) AND (“learning” OR “business”) AND (“engagement” OR “motivation) AND (“learning OR training”) AND (“performance” OR “productivity”). The mentioned keywords are combined in order to retrieve relevant literature on the application of gamification in an organizational or business context and relate to aspects regarding motivation, engagement, learning, training, performance, and productivity improvement.

TABLE I. REVIEWED GAMIFICATION STUDIES

Ref	Scope	Country	Theory	Organizational Focus
[26]	Gamified e-training improves employees’ self-efficacy and information security behaviors	Germany	IS Success Model	Engineering and high-tech sector
[5]	Gamified corporate training enhances employees’ learning and performance	European continent especially France and Spain	SDT, Cognitive Load Theory & Social Exchange Theory	Varying sector
[11]	Analyze how gamification influences employees’ enjoyment and	Germany and Europe	TAM	Engineering and high-tech sector

	acceptance of e-training system			
[27]	Impact of SDT on gamified corporate training outcomes	India	SDT	Information technology sector
[28]	Gamification affects employee engagement and productivity	UAE, Saudi Arabia, Oman, Bahrain, Kuwait, Qatar	Flow Theory, SDT, TAM	Information technology sector
[29]	Game dynamics effect motivation, enjoyment, and work engagement	Indonesia	SDT	Technology -based creative sector
[30]	Gamification is implemented in Polish comparing theory with practice	Poland	SDT, Flow Theory	Business enterprises sector
[31]	Gamification improves HR management and employee motivation	Ukraine	N/S*	Energy sector
[32]	Cloud gamification system to enhance employee motivation	N/S*	Design Science Research, 6D Gamification Framework	Information technology sector
[33]	Gamified e-coaching supports employee health and engagement in the workplace	Netherland	Transtheoretical Model, Unified Health Gamification	Occupational health and workplace well-being sector

*N/S: not specified

Cybersecurity awareness can be strengthened through training initiatives that integrate gamification into e-training programs. By incorporating elements such as points, challenges, badges, and real-time feedback, these programs are designed to increase employee engagement and encourage active participation. Such mechanisms not only reinforce self-efficacy but also promote consistent security-related behavior over time. In addition, gamification helps create an interactive and immersive learning environment that keeps cybersecurity training dynamic, reduces learning fatigue, and supports employees in continuously understanding and responding to emerging security risks.

IV. Discussion

It can be analyzed from past literature that gamification enhances the motivation and behavior of employees regarding cybersecurity training by providing certain engaging, competitive elements which enhance self-efficacy and learning experiences. These findings constitute the conceptual basis for designing a gamification approach suitable for cybersecurity awareness in SMEs.

TABLE II. Gamification Elements and Outcomes

Ref	Gamification Elements	Platform Used	Outcome
[26]	Challenges, clear goals, feedback, and narrative context	Online gamified training platform	Enhancing employee engagement and security self-efficacy within gamified cybersecurity training
[5]	Leaderboards, interactive challenges, points, badges, and social collaboration features	Discovery-Innovation-Growth (DIG) Platform	Increased engagement, knowledge retention, and employee performance. Leaderboards and interactivity were found to be the most empirically effective in capturing attention and enhancing participant motivation
[11]	Narrative and challenges	Gamified e-training system	Enhanced enjoyment and positive employee attitudes toward the system
[27]	Leaderboards, challenges, and points	Gamified internal training system	Influenced employee motivation and participation. However, the effectiveness of these elements depends on their alignment with organizational culture and training program design
[28]	Employee recognition	Kahoot, Mentimeter, Scoutible, and AI/ML/IoT-based gamified HR Systems	Enhancing employee productivity by providing recognition and positive feedback that significantly boost motivation and work engagement
[29]	Points, levels, and badges	Corporate gamified internal system	Enhancing intrinsic motivation, which in turn contributes to increased work engagement
[30]	Competition, collaboration, storyline, continuous feedback, personalization, and intrinsic motivation fulfillment	Cloud-based corporate gamification platforms	Enhancing engagement and guiding user behavior toward business goals
[31]	Competition, win-win strategy, visual aesthetics, and quests	Gamified simulation training platform	Enhance employee motivation, collaboration, and productivity
[32]	Points, leaderboards, and badges	Cloud tools	Enhance performance transparency, and provide recognition for individual achievement, collectively strengthening employee motivation and engagement
[33]	Two-level game design, combining personal and social challenges through personalization, social competition, and coach involvement	Mobile apps	Enhancing motivation and promoting positive health behavior changes among employees

RQ1. How is gamification used in employee-focused cybersecurity training in organizations?

The review of the ten articles in Table I proved that most studies concerning the application of gamification in cybersecurity awareness training were set in the context of a large organization. However, no studies have been found to investigate the application of gamification to raise cybersecurity awareness specifically for SME actors. Thus, there is still a research gap regarding how gamification can be adapted to the characteristics and needs of SME actors.

RQ2. Which elements of gamification have the most influence on employee motivation factors in cybersecurity training?

Table II Among the gamification elements in cybersecurity training that can best influence employee motivation are leaderboards, challenges, points and badges, feedback, and narrative [5], [26], [30]. These indeed increase employee participation, focus, and intrinsic motivation if challenged, with progress that can be identified and rewarded in effort. Equally, the establishment of continuous feedback and the use of narration lead to a more captivating and motivating situation to the respondents, which will encourage them to finish the course and remember it for a longer period. The interaction between the above elements plays an eventual role in motivation, participation, and self-efficacy that is widely induced in gamified cybersecurity training.

RQ3. How does conceptual model apply gamification to increase cybersecurity awareness in the SMEs sector?

A. Self-Determination Theory

Perceived autonomy refers to the extent to which learners perceive freedom and control over their learning process [34]. In gamified contexts, the design elements of free-choice learning pathways, extra challenges, and personalized avatar choices allow participants to make choices based on preference, which, in turn, supports intrinsic motivation and engenders long-term engagement in digital contexts [1]. Accordingly, this study formulates the following hypothesis:

H1: Perceived autonomy positively influences intention to use gamified cybersecurity awareness training

Perceived competence reflects an individual's confidence in their capacity to comprehend or accomplish a given task [35]. Thus, in a gamified training system, mechanisms such as earning points, progressing through hierarchical levels, acquiring achievement badges, and receiving immediate feedback about performance enhance participants' self-evaluations of their own abilities, which in turn promote active participation along with continued intention to use the platform [1]. Consequently, the following hypothesis is formulated:

H2: Perceived competence positively influences intention to use gamified cybersecurity awareness training.

Perceived relatedness is defined as the extent to which participants feel connected to others using the intervention [34] Gamification elements of collaborative missions, competitive leaderboard, and interactive social functionalities establish a sense of community that promotes comfort and intensifies commitment, leading to continued intention to use the cybersecurity training platform [36]. On this basis, the following hypothesis is proposed:

H3: Perceived relatedness positively influences intention to use gamified cybersecurity awareness training.

Intrinsic motivation is an inner drive, which encourages individuals to engage in activities they find interesting and gratifying, and which also resonates with basic needs satisfaction according to SDT. Individuals with a high level of intrinsic motivation would likely exhibit heightened levels of autonomy, a positive sense of competence, and heightened levels of relatedness [37], [38]. As such, intrinsic motivation is expected to positively relate to perceived autonomy, perceived competence, and perceived relatedness. Drawing upon this theoretical rationale, the following hypotheses are advanced:

H4: Intrinsic motivation positively influences on perceived autonomy.

H5: Intrinsic motivation positively influences on perceived competence.

H6: Intrinsic motivation positively influences on perceived relatedness.

B. Theory of Planned Behavior

Attitude toward behavior refers to an individual's positive or negative evaluation of performing a certain behavior [20]. In the context of gamified cybersecurity training, positive attitudes are created when the participants view the training as enjoyable, effective, and helpful. Empirical evidence shows that these positive attitudes lead to increased interest and behavioral intention to use a system.. Consistent with this theoretical perspective, the following hypothesis is formulated:

H7: Attitude toward using gamified cybersecurity awareness training positively influences intention to use gamified cybersecurity awareness training.

Subjective norms are the perceived social pressures that guide individuals to engage in certain behaviors [19]. In the context of gamified training systems, the endorsement and encouragement from peers or supervisors reinforce the intention to use the platform, and empirical findings thus far show that social normative influences significantly affect one's intentions to adopt technology [39]. Based on this reasoning, the following hypothesis is proposed:

H8: Subjective norm positively influences intention to use gamified cybersecurity awareness training.

Perceived Behavioral Control (PBC) refers to the degree to which individuals believe they possess the ability and resources to execute a particular behavior [19]. In the context of gamified cybersecurity training programs, PBC can consist of system usability, available time, and institutional support, each of which has been identified as influencing adoption intentions towards digital systems [39] Grounded in this theoretical framework, the following hypothesis is advanced:

H9: Perceived behavioral control positively influences intention to use gamified cybersecurity awareness training

C. TOE Framework Organizational

According to the TOE Framework, internal organizational factors are decisive in technology adoption processes, especially those related to managerial receptiveness toward innovations [40]. Manager innovativeness denotes the degree to which leaders are open to adopting new ideas and emerging

technologies, an attribute that has empirically been shown to facilitate the pace of digital innovation adoption, including gamified training [25]. Innovation-oriented managerial support thus acts as a driver for the effective implementation of gamified cybersecurity training programs. Accordingly, the following hypothesis is formulated:

H10: Manager innovativeness positively influences intention to use gamified cybersecurity awareness training

Organizational readiness refers to the sufficiency of resources, institutional support structures, and technological infrastructure that altogether determine an organization's preparedness for emerging technology adoption [40]. It has been empirically verified that organizations better prepared for the implementation of digitized innovations tend to achieve higher success rates [25]. In the context of gamified cybersecurity training, such organizational preparedness allows for easier processes of adoption and continued use of the platforms. Based on this premise, the following hypothesis is proposed:

H11: Organizational Readiness positively influences intention to use gamified cybersecurity awareness training

This study develops a comprehensive conceptual framework that integrates Self-Determination Theory (SDT), the Theory of Planned Behavior (TPB), and the organizational context of the Technology-Organization-Environment (TOE) Framework together with factors specific to SMEs. By combining these perspectives, the model show how individual psychological needs, behavioral intentions, and organizational readiness simultaneously predict employees' willingness to adopt gamified cybersecurity awareness training. It allows the conceptual model to provide a holistic view of the determinants of intention and to visualize the relationship among all studied variables, as shown in Figure 1.

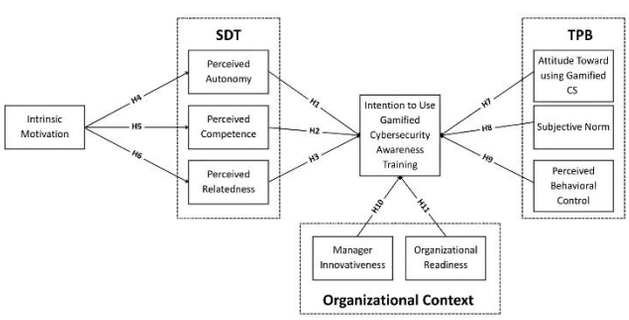


Fig. 1. Research Model

V. Conclusion and Future Research

This research demonstrates that gamification holds significant promise for enhancing cybersecurity awareness, yet its implementation remains predominantly confined to large-scale enterprises and has not been sufficiently tailored to address the specific requirements of SMEs. The literature analysis reveals that mechanisms such as leaderboards, challenges, point systems, achievement badges, real-time feedback, and storytelling elements consistently demonstrate effectiveness in fostering motivation, engagement, and self-efficacy among participants. Drawing upon these insights, this study proposes a conceptual framework that synthesizes constructs from SDT, TPB, and organizational readiness factors to explain the adoption of gamified cybersecurity training within SME contexts. This model offers both theoretical contributions to the existing body of knowledge and establishes a foundation for future empirical validation studies.

References

[1] A. K. Gwenhure and F. S. Rahayu, "Gamification of Cybersecurity Awareness for Non-IT Professionals: A Systematic Literature Review," *International Journal of Serious Games*, vol. 11, no. 1, pp. 83–99, Mar. 2024, doi: 10.17083/ijsg.v11i1.719.

[2] J. Mahat, N. Alias, and F. D. Yusop, "Systematic literature review on gamified professional training among employees," 2023, *Routledge*. doi: 10.1080/10494820.2022.2043910.

[3] H. R. D. Putranti, R. Retnowati, A. A. Sihombing, and D. Danang, "Performance Assessment through Work Gamification: Investigating Engagement," *South African Journal of Business Management*, vol. 55, no. 1, 2024, doi: 10.4102/sajbm.v55i1.4287.

[4] S. Mohanty and B. P. Christopher, "A study on role of gamification elements in training outcomes: comparing the mediating effect of intrinsic and extrinsic motivation," *Learning Organization*, vol. 30, no. 4, pp. 480–500, Jun. 2023, doi: 10.1108/TLO-08-2022-0098.

[5] A. Capatina, D. Juarez-Varon, A. Micu, and A. E. Micu, "Leveling up in corporate training: Unveiling the power of gamification to enhance knowledge retention, knowledge sharing, and job performance," *Journal of Innovation and Knowledge*, vol. 9, no. 3, Jul. 2024, doi: 10.1016/j.jik.2024.100530.

[6] N. Dacre, P. Constantinides, and J. Nandhakumar, "How to Motivate and Engage Generation Clash of Clans at Work? Emergent Properties of Business Gamification Elements in the Digital Economy," *SSRN Electronic Journal*, Apr. 2021, doi: 10.2139/ssrn.3809398.

[7] L. Smirani and H. Yamani, "Analysing the Impact of Gamification Techniques on Enhancing Learner Engagement, Motivation, and Knowledge Retention: A Structural Equation Modelling Approach," *Electronic Journal of e-Learning*, vol. 22, no. 9, pp. 111–124, Sep. 2024, doi: 10.34190/ejel.22.9.3563.

[8] F. Kausar, M. U. Ijaz, M. Rasheed, A. Suhail, and U. Islam, "Empowered, accountable, and committed? Applying self-determination theory to examine work-place procrastination," *BMC Psychol*, vol. 13, no. 1, Dec. 2025, doi: 10.1186/s40359-025-02968-7.

[9] M. M. Alshammari and Y. H. Al-Mamary, "Bridging Policy and Practice: Integrated Model for Investigating Behavioral Influences on Information Security Policy Compliance," *Systems*, vol. 13, no. 8, Aug. 2025, doi: 10.3390/systems13080630.

[10] K. Wasudawan and C. H. Sim, "Shaping a Digital Future: Examining Technology, Organisation and Environment (TOE) Framework," *Journal of Technology Management and Business*, vol. 11, no. 1, Jun. 2024, doi: 10.30880/jtmb.2024.11.01.005.

[11] P. Bitrián, I. Buil, S. Catalán, and S. Hatfield, "The use of gamification strategies to enhance employees' attitudes towards e-training systems," *International Journal of Management Education*, vol. 21, no. 3, Nov. 2023, doi: 10.1016/j.ijme.2023.100892.

[12] T. M. Tran, R. Beuran, and S. Hasegawa, "Gamification-Based Cybersecurity Awareness Course for Self-regulated Learning," *International Journal of Information and Education Technology*, vol. 13, no. 4, pp. 724–730, Apr. 2023, doi: 10.18178/ijiet.2023.13.4.1859.

[13] A. Chidukwani, S. Zander, and P. Koutsakis, "Cybersecurity preparedness of small-to-medium businesses: A Western Australia study with broader implications," *Comput Secur*, vol. 145, Oct. 2024, doi: 10.1016/j.cose.2024.104026.

[14] Y. Hoong, D. Rezania, and R. Baker, "When traditional SME managers encounter cybersecurity: Discourse analysis of opportunities and dilemmas in meeting the demands," in *Technology in Society*, Elsevier Ltd, Sep. 2024. doi: 10.1016/j.techsoc.2024.102650.

[15] H. Taherdoost, "A Critical Review on Cybersecurity Awareness Frameworks and Training Models," in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 1649–1663. doi: 10.1016/j.procs.2024.04.156.

[16] E. L. Deci, A. H. Olafsen, and R. M. Ryan, "Self-Determination Theory in Work Organizations: The State of a Science," *Downloaded from www.annualreviews.org. Guest (guest*, vol. 4, p. 19, 2025, doi: 10.1146/annurev-orgpsych.

[17] R. M. Ryan and E. L. Deci, “Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being Self-Determination Theory,” Ryan, 1985.

[18] R. A. Laguerre and J. L. Barnes-Farrell, “Bringing Self-Determination Theory to the Forefront: Examining How Human Resource Practices Motivate Employees of All Ages to Succeed,” *J Bus Psychol*, vol. 40, no. 1, pp. 1–37, Feb. 2025, doi: 10.1007/s10869-024-09951-w.

[19] I. Ajzen, “The theory of planned behavior,” *Organ Behav Hum Decis Process*, vol. 50, no. 2, pp. 179–211, 1991, doi: 10.1016/0749-5978(91)90020-T.

[20] I. Ajzen, “The theory of planned behavior,” in *Handbook of Theories of Social Psychology: Volume 1*, SAGE Publications Inc., 2012, pp. 438–459. doi: 10.4135/9781446249215.n22.

[21] M. Bosnjak, I. Ajzen, and P. Schmidt, “The theory of planned behavior: Selected recent advances and applications,” Aug. 01, 2020, *PsychOpen*. doi: 10.5964/ejop.v16i3.3107.

[22] S. Suntornsan, S. Chudech, and P. Janmaimool, “The Role of the Theory of Planned Behavior in Explaining the Energy-Saving Behaviors of High School Students with Physical Impairments,” *Behavioral Sciences*, vol. 12, no. 9, Sep. 2022, doi: 10.3390/bs12090334.

[23] I. Ahmed, “Technology organization environment framework in cloud computing,” *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 18, no. 2, pp. 716–725, Apr. 2020, doi: 10.12928/TELKOMNIKA.v18i2.13871.

[24] A. B. N’Dri and Z. Su, “Successful configurations of technology–organization–environment factors in digital transformation: Evidence from exporting small and medium-sized enterprises in the manufacturing industry,” *Information and Management*, vol. 61, no. 7, Nov. 2024, doi: 10.1016/j.im.2024.104030.

[25] T. H. Nguyen, X. C. Le, and T. H. L. Vu, “An Extended Technology-Organization-Environment (TOE) Framework for Online Retailing Utilization in Digital Transformation: Empirical Evidence from Vietnam,” *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 8, no. 4, Dec. 2022, doi: 10.3390/joitmc8040200.

[26] P. Bitrián, I. Buil, S. Catalán, and D. Merli, “Gamification in workforce training: Improving employees’ self-efficacy and information security and data protection behaviours,” *J Bus Res*, vol. 179, Jun. 2024, doi: 10.1016/j.jbusres.2024.114685.

[27] P. Kulkarni, P. Gokhale, Y. M. Satish, and B. Tigadi, “An empirical study on the impact of learning theory on gamification-based training programs,” *Organization Management Journal*, vol. 19, no. 5, pp. 170–188, Nov. 2022, doi: 10.1108/OMJ-04-2021-1232.

[28] H. U. Rahiman, R. Kodikal, and S. Suresh, “Game on: Can gamification enhance productivity?,” *F1000Res*, vol. 12, p. 818, Jul. 2023, doi: 10.12688/f1000research.131579.1.

[29] D. Setiawan RIATMAJA, M. Suyanto, M. Muafi, and W. Prajogo, “The Effect of Using Game Dynamics Towards Employee Work Engagement: An Empirical Study in Indonesia,” *Journal of Asian Finance*, vol. 8, no. 6, pp. 427–0436, 2021, doi: 10.13106/jafeb.2021.vol8.no6.0427.

[30] A. Witoszek-Kubicka, “Implementation of gamification in Polish companies-stages, elements, ethics,” *Information (Switzerland)*, vol. 11, no. 8, Aug. 2020, doi: 10.3390/INFO11080371.

[31] A. Polyanska, M. Andriiovych, N. Generowicz, J. Kulczycka, and V. Psyuk, “Gamification as an Improvement Tool for HR Management in the Energy Industry—A Case Study of the Ukrainian Market,” *Energies (Basel)*, vol. 15, no. 4, Feb. 2022, doi: 10.3390/en15041344.

[32] C. Ribeiro, I. Fernandes, and F. Portela, “Toward an Enterprise Gamification System to Motivate Human Resources in IT Companies,” *Information (Switzerland)*, vol. 15, no. 1, Jan. 2024, doi: 10.3390/info15010026.

[33] C. Zhang *et al.*, “Promoting occupational health through gamification and e-coaching: A 5-month user engagement study,” *Int J Environ Res Public Health*, vol. 18, no. 6, pp. 1–17, Mar. 2021, doi: 10.3390/ijerph18062823.

[34] L. Li, K. F. Hew, and J. Du, “Gamification enhances student intrinsic motivation, perceptions of autonomy and relatedness, but minimal impact on competency: a meta-analysis and systematic review,” *Educational Technology Research and Development*, vol. 72, no. 2, pp. 765–796, Apr. 2024, doi: 10.1007/s11423-023-10337-7.

[35] O. Pahlavanpour and S. Gao, “A systematic mapping study on gamification within information security awareness programs,” Oct. 15, 2024, *Elsevier Ltd*. doi: 10.1016/j.heliyon.2024.e38474.

[36] S. A. Triantafyllou and C. K. Georgiadis, “Gamification of MOOCs and Security Awareness in Corporate Training,” in *International Conference on Computer Supported Education, CSEDU - Proceedings*, Science and Technology Publications, Lda, 2022, pp. 547–555. doi: 10.5220/0011103000003182.

[37] E. L. Deci and R. M. Ryan, “The ‘What’ and ‘Why’ of Goal Pursuits: Human Needs and the Self-Determination of Behavior,” 2000.

[38] M. Gagné and E. L. Deci, “Self-determination theory and work motivation,” *J Organ Behav*, vol. 26, no. 4, pp. 331–362, Jun. 2005, doi: 10.1002/job.322.

[39] G. N. O. Copyrights @ Dwi Fitriani and I. Winda, “Analisis Pengaruh Attitude, Subjektif Norm dan Perceived Behavioral Control pada Game Shopee Tanam Terhadap Purchase Intention dan Actual Behavior,” *Journal Of Social Science Research*, vol. 3, pp. 5313–5330.

[40] O. Ali, P. A. Murray, S. Muhammed, Y. K. Dwivedi, and S. Rashiti, “Evaluating Organizational Level IT Innovation Adoption Factors among Global Firms,” *Journal of Innovation and Knowledge*, vol. 7, no. 3, Jul. 2022, doi: 10.1016/j.jik.2022.100213.

Orbital Intelligence: A Vision for Sustainable Large Language Model Training and Inference in Space

Mohamed Khalil Brik[1]

[1]The American University in Cairo

September 2025

Abstract

The rapid expansion of Large Language Model (LLM) training has led to significant computational requirements and associated carbon emissions. For example, training models such as GPT-4 produces approximately 7,138 metric tons of carbon dioxide equivalent. This vision paper proposes the use of space-based infrastructure for LLM training and, critically, inference, to address these sustainability challenges by utilizing abundant orbital solar energy. Orbital computation offers a critical thermodynamic advantage, accessing solar irradiance that is four times more intense than the terrestrial average. We propose a hybrid approach that combines dedicated orbital data centers (ODCs) in Geostationary Earth Orbit (GEO) for foundational training with distributed edge computing resources across Low Earth Orbit (LEO) constellations for sustainable inference. This reframed strategy tackles the 90% of a model's lifecycle energy consumption attributed to usage. The paper addresses core challenges, including radiation hardening, thermal management using Loop Heat Pipes (LHP), the logistics of data migration, and the strategic mandate for a tenfold reduction in launch emissions, concluding that space represents the next necessary frontier for large-scale, sustainable AI development.

1 Introduction

The unprecedented scaling of Artificial Intelligence (AI) has established Large Language Models (LLMs) as the technological cornerstone of the modern digital economy [1]. However, this transformative growth is underpinned by an accelerating and globally significant environmental footprint. Training state-of-the-art LLMs demands computational resources that place immense strain on terrestrial energy grids and water supplies, necessitating a radical shift in infrastructure deployment.

The migration of high-performance computing (HPC) to space is not merely a novelty but a strategic response to a fundamental energy crisis. The orbital environment provides a decisive advantage: unparalleled access to solar radiation unattenuated by the Earth's atmosphere [2]. This continuous, zero-carbon power source offers the means to fundamentally decouple AI computation from the variable, carbon-intensive constraints of terrestrial grids [3].

This paper details the techno-economic and architectural feasibility of establishing Orbital Data Centers (ODCs). We advocate for a hybrid, multi-orbital architecture designed to optimally place AI workloads: synchronous, high-power training in the stable environment of GEO, and asynchronous, high-volume inference within distributed LEO constellations. We analyze the core engineering solutions required for hardware resilience and thermal management, and we address the strategic mandates related to data logistics, regulatory governance, and the critical challenge of launch sustainability.

2 The Environmental Imperative

The case for orbital deployment is rooted in the alarming quantification of LLM development costs across carbon, energy, and water consumption.

2.1 Quantifying the Terrestrial Footprint

The training of frontier LLMs results in energy and carbon costs that far exceed previous computational models. For example, training OpenAI's GPT-3, possessing 175 billion parameters, required approximately 1,287 megawatt-hours (MWh) of electricity [4] and resulted in an estimated 552 metric tons of CO_2 [5]. The environmental cost scaled dramatically with its successor, GPT-4, which was estimated to produce 7,138 metric tons of CO_2, a footprint comparable to the annual emissions of 1,550 US citizens [5, 6].

Furthermore, terrestrial LLM operations impose a profound strain on water resources. Training GPT-3 alone required over 700 kiloliters (kL) of water for

cooling, a volume large enough to fill two-thirds of an Olympic-sized swimming pool [7]. The orbital shift, by utilizing the deep-space vacuum for passive heat rejection, completely eliminates this reliance on terrestrial water for cooling.

2.2 The Inference-Centric Strategy

Historically, sustainability discussions concentrated on the singular, massive training phase. However, this focus is strategically incomplete. Recent comprehensive benchmarking and analysis indicate that LLM *inference* (continuous, global usage) is the dominant factor in lifetime environmental costs, potentially accounting for up to 90% of a model's total lifecycle energy consumption [7–9].

This paradigm shift necessitates a re-evaluation of the orbital value proposition. While a GEO hub may handle the 10% training problem, the greatest environmental return on investment (EROI) lies in leveraging LEO constellations as a zero-carbon, distributed edge inference network to address the dominant 90% problem [10]. By relying exclusively on clean, captured solar energy, orbital infrastructure operates with a localized operational carbon intensity factor (F_{em}) tending toward zero [9].

3 Thermodynamic Advantages of Orbital Placement

The feasibility of Orbital Intelligence relies entirely on the superior energy continuity and intensity provided by the space environment.

3.1 Solar Irradiance and Attenuation

The critical advantage of space is the solar constant, the intense, unattenuated solar radiation flux, nominally measured at $1367.5\,W/m^2$ outside the Earth's atmosphere [2]. This figure varies seasonally by approximately $\pm 3.5\%$ due to orbital eccentricity [2].

This contrasts sharply with the average solar radiation reaching the Earth's surface (insolation), which is dramatically attenuated by atmospheric scattering, reflection, and absorption, averaging only about $340\,W/m^2$ [11]. Orbital collection, therefore, confirms a foundational physics advantage, accessing a solar energy source that is four times more intense than the terrestrial average, maximizing the power generation potential per unit of solar collector mass.

3.2 Orbital Stability and Power Continuity

The stability required for continuous, synchronous LLM training dictates the choice of orbit.

Table 1: Comparative Orbital Power Continuity

Parameter	**LEO**	**GEO**
Altitude (km)	160–2,000 [12]	35,786 [13]
Eclipses per Year	$\approx 5,500$ [14]	≈ 84 [14]
Max Eclipse Duration	35 minutes [15]	72 minutes [14]
Battery DoD Stress	High (5,000+ cycles) [14]	Low (Seasonal only)
Best Workload	Inference / Edge Computing [16]	Foundational Training [17]

Geostationary Earth Orbit (GEO) is ideal for foundational training due to its near-continuous solar exposure, experiencing predictable eclipses only seasonally around the equinoxes [13,14]. Conversely, LEO's high frequency of eclipses (approximately 15 per day) [14] subjects hardware and battery systems to severe thermal and battery cycling stress (over 5,000 cycles annually), rendering it fundamentally unsuitable for sustained, synchronous training operations [15].

4 Hybrid Architectural Framework

The proposed hybrid architecture must strategically match the computational workload to the optimal orbital environment, leveraging both dedicated infrastructure and repurposing existing assets.

4.1 Dedicated Orbital Data Centers

Dedicated AI satellites are already transitioning from concept to reality. China, for instance, has embarked on constructing a vast orbital supercomputer network, currently achieving 5 POPS (peta operations per second) with a long-term objective of 1,000 POPS [18]. These specialized satellites utilize onboard 8-billion parameter AI models, high-speed laser inter-satellite links (ISLs) up to 100 Gbps, and 30 terabytes of storage to process remote sensing data and bypass terrestrial bandwidth bottlenecks [18].

- **GEO Cluster for Foundational Training:** Foundational LLM training requires high-performance Remote Direct Memory Access (RDMA) networking to perform constant, synchronous, all-to-all model weight reduction operations across distributed GPU clusters [17].

The required instantaneous interconnect bandwidth is estimated at approximately 1 Gbps per GPU [19]. Only the static, stable GEO environment can reliably provide the required stable, high-synchronicity links.

- **LEO Constellations for Distributed Inference:** LEO networks are ideally suited for asynchronous distributed inference [10, 16]. In this model, LEO satellites function as requesters and workers, processing data near the source to reduce communication latency and bandwidth cost [16].

4.2 Constraints of Mission Augmentation

While the hybrid approach includes computational resources on existing missions, their contribution is constrained. The International Space Station (ISS) successfully prototypes AI, notably using platforms like HPE's Spaceborne Computer-3 [20] for edge AI tasks such as assisting astronauts with maintenance procedures [12] and dynamic targeting for Earth observation [21]. However, repurposing Global Navigation Satellite Systems (GNSS), such as GPS, for deep learning is severely limited. Legacy, radiation-tolerant processors on these platforms are several magnitudes below the computing power required for advanced deep learning algorithms [22]. Existing missions can only feasibly support small-scale, asynchronous edge inference, lacking the volume, power budget, and thermal management necessary for sustained foundational LLM training [3, 22].

5 Critical Engineering and Logistics

The deployment of commercial-grade HPC hardware in space is constrained by two fundamental physical challenges: radiation and heat dissipation in a vacuum.

5.1 Radiation Hardening for COTS Hardware

The space environment presents severe risks to modern Commercial Off-The-Shelf (COTS) components, which are essential for LLM performance but are designed for benign terrestrial operation. Radiation from galactic cosmic rays and solar events causes two primary failure modes: Total Ionizing Dose (TID), resulting in cumulative material degradation, and Single Event Effects (SEE), causing functional disturbances or catastrophic failure from single particle strikes [23, 24]. The Geosynchronous Transfer Orbit (GTO), a proxy for the GEO environment, subjects hardware to extreme TID, potentially up to 100K Rad per annum [25].

Since performance necessitates COTS GPUs/TPUs, the strategy relies on advanced mitigation:

- **System-Level Mitigation:** Employing good fault-tolerant design practices is crucial for mitigating SEE at the circuit and system level [24].
- **Hardware Redundancy:** Specialized manufacturers, such as AMD, offer components incorporating Triple Modular Redundancy (TMR) designed to mitigate radiation effects [26].
- **Real-Time Shielding:** Novel concepts like the COTS-Capsule propose using particle detector arrays to encapsulate CSEE-sensitive electronics, enabling real-time detection and characterization of particle strikes for immediate mitigation or error correction [27].
- **Co-Processing:** Leveraging radiation-hardened co-processors, like the High Performance Data Processor (HPDP), as a dependable mathematical backend allows the execution of AI-driven workloads without relying solely on vulnerable COTS components [28].

5.2 Thermal Management in Vacuum

High-performance AI accelerators generate substantial heat loads [3]. The vacuum of space prohibits atmospheric convection, rendering traditional terrestrial cooling impossible [3].

The reliable solution is the Loop Heat Pipe (LHP). LHPs are passive, two-phase heat transfer devices that use the evaporation and condensation of a working fluid and capillary forces to circulate the fluid, transferring heat from the GPU (evaporator) to a large radiator panel (condenser) that rejects the heat into the deep-space heat sink [29, 30]. This passive architecture offers significant advantages critical for orbital deployment: up to 10 times longer maintenance-free lifetime, lower power consumption, and up to 30% less mass compared to active systems [31].

5.3 Data Logistics: The Hidden Uplink Problem

LLM training is a synchronous, compute-bound workload [17, 19]. While existing space infrastructure focuses on solving the downlink bottleneck (space-to-ground data transmission), the initial

training phase introduces the Hidden Uplink Problem: the necessity of migrating petabytes of clean, foundational data corpora from terrestrial ground stations to orbit. This initial data migration requires dedicated, high-capacity ground-to-space links and represents a vast capital expenditure and time investment.

Furthermore, the dynamic orbital motion of LEO satellites creates the LEO Motion Liability. LEO constellations, despite their advantages for inference, exhibit sustained latency spikes caused by dynamic routing and reliance on constantly shifting inter-satellite links (ISLs). These instabilities are catastrophic for the strict, all-to-all synchronization protocols required for large-scale distributed training, confirming the necessity of a static GEO architecture for the most compute-intensive workloads.

6 Strategic Viability and Governance

The success of Orbital Intelligence is ultimately measured by its economic justification and its ability to operate within the emerging, fractured space regulatory framework.

6.1 Comparative Economics and Launch Debt

Global CapEx on terrestrial data center infrastructure is projected to exceed $1.7 trillion by 2030, driven largely by AI expansion [32]. The Horizon Europe-funded Ascend study concluded that space data centers could be economically viable, projecting a return on investment (ROI) of several billion by 2050, primarily by eliminating the operational costs of carbon energy and complex cooling [3].

However, this economic viability is critically vulnerable to the environmental cost of deployment. The launch phase presents the Launch Sustainability Achilles' Heel. The operational carbon savings will be nullified unless the system's launch debt is drastically reduced. The Ascend study mandate requires the development of a launcher ten times less emissive over its entire lifecycle compared to current vehicles [3, 33]. Without this next-generation, low-emissivity launch capability, the project risks merely shifting the environmental burden from electrical OpEx to launch CapEx.

6.2 The Regulatory and Sovereignty Void

The exponential proliferation of commercial mega-constellations (projected to reach up to 100,000 LEO satellites) is rapidly outpacing the international legal framework, which is currently centered on the 1967 Outer Space Treaty (OST) [34].

- **Data Sovereignty and Residency:** Processing personal data in non-sovereign space introduces significant legal risk. Since international space law is nascent regarding data privacy, orbital operators must preemptively design systems to comply with diverse, complex domestic and global privacy systems (e.g., GDPR).
- **Orbital Debris Mitigation:** Computational constellations must adhere rigorously to international guidelines, which typically mandate a 90% success rate for post-mission disposal (PMD) [35, 36]. The sheer number of objects in these mega-constellations increases the computational requirement for continuous, complex propagation and conjunction analyses over multi-century timescales [35].
- **Resource Governance:** Orbital resources and radio spectrum, currently managed by the International Telecommunication Union (ITU) under a First Come, First Served principle [37], are finite. New governance mechanisms are required for the allocation and coordination of these resources to avoid conflict and harmful interference [38, 39].

7 Conclusion

The vision of Orbital Intelligence is technologically sound and strategically necessary. By exploiting the thermodynamic advantages of space—zero-carbon solar power and passive deep-space cooling—we can establish a sustainable platform that addresses the escalating environmental footprint of frontier AI. The strategic roadmap must prioritize the deployment of LEO constellations as an immediate, sustainable *inference* network (the 90% problem), while aggressively developing the radiation-hardened and thermally optimized GEO infrastructure for foundational *training* (the 10% problem). The ultimate feasibility of this transition rests on a non-negotiable strategic mandate: the development and utilization of launch vehicle technology that can achieve a ten-fold reduction in lifecycle emissions, thereby ensuring the orbital solution does not simply trade one carbon burden for another.

References

[1] Wikipedia contributors, "Large language model," *Wikipedia, The Free*

Encyclopedia, September 2025, accessed: [2025-09-28]. [Online]. Available: https://en.wikipedia.org/wiki/Large_language_model

[2] NASA LLIS, "Solar constant, albedo, and earth radiation values," NASA Lesson Learned 693, 1995. [Online]. Available: https://llis.nasa.gov/lesson/693

[3] Innovation News Network, "New study highlights the viability of space data centres," *Innovation News Network*, October 2024, citing Thales Alenia Space Ascend Study. [Online]. Available: https://www.innovationnewsnetwork.com/space-data-centres-viability-study/

[4] ADASci, "How much energy do llms consume? unveiling the power behind ai," *ADASci Blog*, August 2024. [Online]. Available: https://adasci.org/llm-energy-consumption/

[5] Shop Without Plastic, "The carbon cost of training large ai models," *Shop Without Plastic Blog*, March 2024. [Online]. Available: https://shopwithoutplastic.com/carbon-cost-ai-models/

[6] Reddit User /u/throwaway-3413, "The carbon footprint of gpt-4 is the equivalent of powering more than 1300 homes for one year!" Reddit r/LocalLLaMA, April 2024. [Online]. Available: https://www.reddit.com/r/LocalLLaMA/comments/carbon-footprint-gpt4/

[7] A. Luccioni, S. Varma, and A. Awe, "The environmental impact of large language models: Training and inference," *arXiv preprint arXiv:2505.09598*, May 2025. [Online]. Available: https://arxiv.org/abs/2505.09598

[8] S. Poddar, P. Koley, J. Misra, N. Ganguly, and S. Ghosh, "Towards sustainable nlp: Insights from benchmarking inference energy in large language models," *arXiv preprint arXiv:2502.05610*, February 2025. [Online]. Available: https://arxiv.org/abs/2502.05610

[9] S. Hugon and T. Giraud, "Llm inference methodology," Ecologits.ai Technical Note, November 2024. [Online]. Available: https://ecologits.ai/methodology

[10] Eutelsat OneWeb, "Edge computing solutions with leo connectivity," Eutelsat OneWeb Solutions, 2024. [Online]. Available: https://oneweb.net/solutions/edge-computing

[11] Wikipedia contributors, "Solar constant," *Wikipedia, The Free Encyclopedia*, September 2025, accessed: [2025-09-28]. [Online]. Available: https://en.wikipedia.org/wiki/Solar_constant

[12] Booz Allen Hamilton, "Deploying a large language model in space," *Booz Allen Hamilton Insights*, July 2024. [Online]. Available: https://www.boozallen.com/insights/deploying-llm-space.html

[13] Wikipedia contributors, "Geostationary orbit," *Wikipedia, The Free Encyclopedia*, September 2025, accessed: [2025-09-28]. [Online]. Available: https://en.wikipedia.org/wiki/Geostationary_orbit

[14] N. Al-Falahy, "Advanced communications systems for 4th class students," *University of Anbar Engineering College*, 2019.

[15] R. M. Mahendran, S. K. Narayanan, and T. V. Pradeep, "Computation of eclipse time for leo small satellites," *International Journal of Advances in Engineering and Architecture*, vol. 3, no. 2, pp. 45–52, 2019.

[16] B. Lin, Z. Xu, R. Li, Z. Wang, M. Chen, and Q. Zhang, "Dynamic velocity-aware resource allocation for leo edge computing networks," *arXiv preprint arXiv:2406.10856*, June 2024. [Online]. Available: https://arxiv.org/abs/2406.10856

[17] A. Pira, "Interconnect needs for llm inference to drive networking bandwidth," *650 Group Blog*, May 2024. [Online]. Available: https://650group.com/llm-interconnect-bandwidth/

[18] Dig.Watch and Space News, "China launches first ai satellites in orbital supercomputer network," *Digital Watch*, June 2024, based on reports from Space News and MojoAuth. [Online]. Available: https://dig.watch/updates/china-ai-satellites-orbital-supercomputer

[19] NVIDIA Developer, "Turbocharge llm training across long-haul data center networks with nvidia nemo framework," *NVIDIA Developer Blog*, March 2024. [Online]. Available: https://developer.nvidia.com/blog/turbocharge-llm-training-nemo-framework/

[20] Y. Min *et al.*, "Hpe spaceborne computer-3: Enabling real-time ai and edge computing for space exploration," *bioRxiv preprint bioRxiv:2025.01.14.633017*,

January 2025. [Online]. Available: https://www.biorxiv.org/content/10.1101/2025.01.14.633017

[21] NASA Jet Propulsion Laboratory, "New ai algorithms streamline data processing for space-based instruments," *NASA ESTO News*, October 2022. [Online]. Available: https://esto.nasa.gov/news/ai-algorithms-space-instruments

[22] S. Barmada, O. Koudelka, and M. Pacher, "Onboard processing for advanced ai/ml space applications: Challenges and prospects," *Aerospace*, vol. 10, no. 2, p. 101, February 2023.

[23] Curtiss-Wright, "Radiation mitigation in space: A cots approach," White Paper, 2025. [Online]. Available: https://www.curtiswright.com/products-solutions/technologies/radiation-mitigation

[24] NASA Engineering and Safety Center, "Radiation mitigation with good fault-tolerant design practices," NASA Technical Bulletin No. 19-01-1, 2021. [Online]. Available: https://nesc.nasa.gov/NESC$_T B_1 9-01-1.html$

[25] NASA Technical Reports Server, "Mission radiation modeling: Gto, leo, and beo environments," NASA, Tech. Rep. NTRS 20220011775, 2022, relevant to GTO TID exposure analysis. [Online]. Available: https://ntrs.nasa.gov/citations/20220011775

[26] AMD, "Radiation defense across all orbits and beyond," AMD Solutions Page, 2025. [Online]. Available: https://www.amd.com/en/products/embedded-and-semi-custom-solutions/radiation-defense

[27] D. Karsenty and J. Mather, "The cots-capsule csee mitigation scheme: A novel approach to shielding commercial-off-the-shelf electronics in space," *arXiv preprint arXiv:2502.04504*, February 2025. [Online]. Available: https://arxiv.org/abs/2502.04504

[28] S. Barmada, M. Zannoun, and O. Koudelka, "Ai at the edge on radiation-hardened hardware: Leveraging the high performance data processor," *arXiv preprint arXiv:2504.03680*, April 2025. [Online]. Available: https://arxiv.org/abs/2504.03680

[29] J. Ku, "Loop heat pipe overview," *SAE Technical Paper 1999-01-2007*, 1999.

[30] S. Semenov, "Introduction to loop heat pipes," NASA Technical Course - TFAWS 2023, 2023. [Online]. Available: https://tfaws.nasa.gov/TFAWS23/Proceedings/TFAWS-23-024-Semenov.pdf

[31] Calyos TM, "Loop heat pipes: How they work and applications," Calyos Technical Page, 2024. [Online]. Available: https://www.calyos-tm.com/loop-heat-pipes

[32] E. Kok, J. Rauer, P. Sachdeva, and P. Pikul, "Scaling bigger, faster, cheaper data centers with smarter designs," *McKinsey & Company*, September 2023. [Online]. Available: https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/scaling-bigger-faster-cheaper-data-centers

[33] NASA Office of Technology, Policy, and Strategy, "Space-based solar power report," NASA Technical Report, 2024. [Online]. Available: https://www.nasa.gov/wp-content/uploads/2024/01/otps-sbsp-report-final-tagged-approved-1-8-24.pdf

[34] Neuraspace, "Responsible rules in orbit: Federated governance for a crowded and congested space," *Neuraspace Blog*, January 2025. [Online]. Available: https://www.neuraspace.com/blog/responsible-rules-orbit-federated-governance

[35] S. Johnson, "Mitigation of orbital debris in low earth orbit with a focus on mega-constellations," Master's Thesis, University of Illinois, 2022. [Online]. Available: https://www.ideals.illinois.edu/items/124567

[36] European Space Agency (ESA), "Managing mega-constellations and space debris mitigation," ESA Clean Space Initiative, 2025. [Online]. Available: https://www.esa.int/Space$_S afety/Clean_S pace/Managing_m e$

[37] V. Strelets, "Sustainable space: Satellites need harmonized spectrum and more," *ITU Hub*, February 2025. [Online]. Available: https://www.itu.int/hub/2025/02/sustainable-space-satellites-spectrum/

[38] S. Hobe, "Orbital law: Governance of resources in space," *Oxford Research Encyclopedia of Planetary Science*, 2024.

[39] G. Rotola and A. Williams, "Regulatory context of conflicting uses of outer space: Astronomy and satellite constellations," *Air and Space Law*, vol. 46, no. 4, pp. 465–490, 2021.

Prescriptive and Contextual Technical Debt Management with LLM and SonarQube

M. Daffa Raygama*, Jati H. Husen, Villy S. Praditha, Yijun Lu, and Adam R. Faqih
Telkom University, Bandung, West Java, Indonesia
(daffraygama, adamrafif)@student.telkomuniversity.ac.id, (jatihusen, villysatria)@telkomuniversity.ac.id,
yijun@ruri.waseda.jp
*corresponding author

***Abstract*—Technical Debt (TD) is a short-term technical decision that compromises long-term quality and stability. Static analysis tools, such as SonarQube, can identify TD but only provide diagnostic analysis without prescriptive recommendations or project-specific context. However, prescriptive and contextual recommendations are essential because the conditions and priori- ties of each project are different. This research aims to develop an LLM-based TD management system integrated with SonarQube and GitHub to provide prescriptive and contextual analysis. Evaluation is carried out using the Rule2Text LLM-as-a-judge framework to measure the prescriptive quality of the system's output. We found that LLM shows a significant improvement from SonarQube and can reliably provide prescriptive analysis, but falls a bit short contextually as it tends to hallucinate.**

***Index Terms*—LLM, technical debt management, SonarQube, software maintenance, continuous integration**

I. Introduction

Technical debt (TD) arises when teams take suboptimal shortcuts that speed delivery but increase future costs, the "interest" paid through rework, schedule slippage, and reduced business value [1]–[3]. To manage TD, practitioners rely on static analysis tools, backlogs, coding standards, and documentation [4], [5]. These approaches are valid but essentially diagnostic. SonarQube, one of the leading static analysis tools, has proven to flag issues without explaining why they matter in this context or what concrete action best suits the project's needs [6]. SonarQube's findings typically stop at "what," leaving developers to infer the why and how, which slows remediation and weakens prioritisation [6]. Recent work in TD calls for analyses that are not only diagnostic but also contextual and prescriptive, grounded in project realities and accompanied by actionable guidance [7]–[9].

The term "contextual" means that TD management varies across different projects and companies [8]. While "prescriptive" means there is justification and the ability to explain the effects of the detected TD [6]. Prescriptive and contextual analyses would also help data-driven decision making, increasing trust and transparency with stakeholders [6]. To fulfill these prescriptive and contextual requirements of TD management, we leverage recent advancements in artificial intelligence, specifically Large Language Models (LLMs). LLM offers a promising avenue for exploration, especially in the Software Engineering field. Models such as GPT-4 can understand code context, explain code, and give adaptive refactoring suggestions based [10]. Furthermore, GPT-4 is one of the most recent LLMs from OpenAI that shows high performance in reasoning and mathematics [11].

We propose an LLM-centric approach using SonarQube, integrated with GitHub, to manage TD at the pull request (PR) level. Leveraging LLM's high reasoning and code understanding capabilities, we intend to integrate LLM's analysis directly into the software development cycle as a PR comment. When developers intend to push an update, the LLM would provide a precise analysis of potential TD(s) that would arise if the changes are committed. Highlighting the severity, short and long-term effects, and suggested action on each TD issue. We raise these research questions (RQs) to evaluate our approach:

- **RQ1:** To what extent do LLM-based systems provide contextually relevant explanations for SonarQube-detected TD?
- **RQ2:** How effective are LLM-based systems in providing prescriptive guidance for addressing TD?

RQ1 and RQ2 are answered using a separate LLM-as-Judge module that evaluates the agent's output Contextual and Prescriptive scores computed on PR-level analyses.

This study contributes: (1) an end-to-end, CI-friendly pipeline that augments SonarCloud findings with LLM-generated, PR-scoped explanations and remedies; and (2) implementation artefacts, including GitHub Action

workflow, prompts, and evaluation harnesses to enable replication.

The remainder of this paper is structured as follows: a survey of related works on TD management and the tools proposed by previous works. The proposed approach is described in detail, from the individual parts of the system to the evaluation design. The result would then be presented and analysed further, discussing the implications and answers for RQs. The discussion also follows the limitations we faced during our approach. Finally, a conclusion is drawn from our discussion, highlighting the results of our research and calling for further study in our area with specific directions.

II. Related Work

TD management is a primary concern at the PR level. Introducing TD in a PR increases the risk of that PR being rejected [12]. At the same time, Karmakar et al. found that nearly 40% of comments in PR are related to identifying or debating TD [13]. Calixto et al. revealed that TD is neglected in approximately 75% of cases, either remaining unchanged or increasing [14]. Developers struggle to identify and address TD effectively during the manual code review process because the discussions tend to be unstructured and lack context [12].

Recent works have produced several tools and solutions for managing code TD. Martini et al. developed Anacondebt, a tool that tracks and evaluates TD through systematic visualisation and refactoring plans [15]. Mendes et al. introduced VisMinerTD, which identifies TD, code smells, defects, and more within a project [16]. Dornauer et al. noticed that while SonarQube is effective in detecting TD and code smells, it cannot prioritise debt based on project impact or explain when and why certain debts must be addressed or deferred [17]. This limitation can lead to inefficiencies in managing TD. To address this, they proposed SoHist, a tool designed to overcome SonarQube's weaknesses. Tsoukalas et al. emphasised that while these tools are effective at identifying TD, they still fall short in providing explanations of causes, impacts, and strategies for contextual resolution [6]. Even efforts like SoHist do not fully address the need for reasoning and recommendations.

Unlike prior tools that emphasise detection, visualisation, or history tracking of TD, our contribution is a CI-centred framework that augments SonarQube results with LLM-driven, *contextual and prescriptive* explanations at the PR level. This closes the gap between purely diagnostic findings and actionable remediation steps within developers' existing workflows. Our work extends the previously mentioned approaches by implementing an LLM-based method for TD management. This is important because there is a need for a method to provide not only a diagnostic solution to TD but also a prescriptive solution to TD [17].

III. Proposed Approach

This section provides a detailed explanation of our LLM-centric system.

A. System Workflow Overview

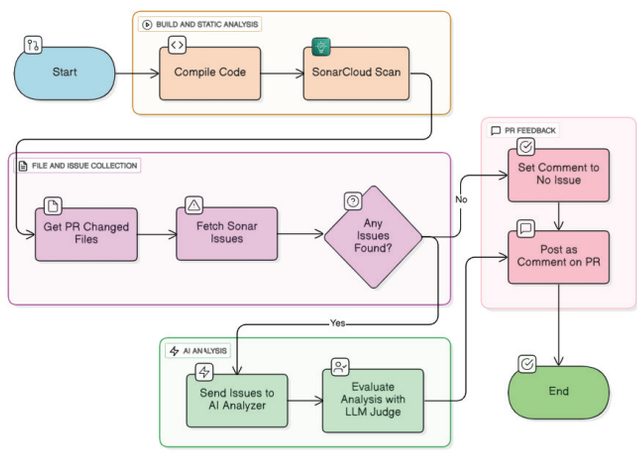


Fig. 1. The full integration flow with GitHub, SonarCloud, and AI Analyser agent.

Figure 1 details our workflow in detail. The workflow begins when a developer opens or updates a PR on GitHub. This event acts as the trigger for the automated analysis pipeline. After a successful build, the code undergoes static analysis using SonarCloud, a cloud-hosted version of SonarQube. Then, the system fetches changed files from the most recent commit, along with Sonar issues from SonarCloud. If no problems are present, the flow will end with a message. Otherwise, the issues and file changes are sent directly to the AI agent analyzer via input message. Another AI agent would then judge the analysis, and the final result will be sent as a PR comment.

B. Build and Static Analysis

The trigger will follow up on another process of compiling and scanning the code. Firstly, we set up the environment on our GitHub Action. Secondly, we prepare to scan with SonarCloud. Compiling the code is required before any SonarCloud scan, so we do so while skipping tests for simplicity. Finally, we initiate the scan to find sonar issues that would be analysed as potential TD.

C. File and Issue Collection

We intend to leverage GPT-4's code understanding capabilities. Therefore, we need to inject some of the troubled code snippets into the AI analyser agent. For this, we need to get the changed files within that PR. Firstly, we list the changed files for this PR. Then, we build a list of files mentioned by the Sonar issues. Secondly, we find the intersection of changed files in PR and in the Sonar issues. Finally, the intersect will list the changed files that would then be sent to another process. After we get the changed files, we set up another action to fetch Sonar issues and their code snippets in one block. We set up configurable caps to limit or extend the sonar issues fetched from the previous scan. Secondly, we iterate the Sonar issues through SonarCloud's API, appending them to a JSON file. We trim the issues to the previously mentioned caps. Finally, we build the compact snippets, then we output them. If no sonar issues are found, we skip the AI analyser agent and comment that there is no potential TD in the current PR. Otherwise, it would be sent to the AI analyser agent to address whether there are TD(s) and what their effects are.

D. AI Analysis

We send the previous snippet to the AI analyser agent through Flowise's prediction API. Firstly, we set up another configurable payload settings, such as max summary chars, max question chars, max payload bytes, etc. Secondly, we format the snippet into a more readable markdown summary and set up the question text, instructing the agent to analyse Sonar Cloud issues from the current PR number in this repository, targeting this PR title and message. Finally, we create the payload, which consists of snippets of rules, severity, location, and message for each Sonar issue. Then we send it through Flowise's API to the analyzer agent Agentflow.

This Agentflow consists of the start node and the AI Agent node. We set up the AI agent using a persona prompt pattern [18]. We let the Agent use GitHub MCP, which it can use to search for code, issues, etc., in a repository. The knowledge base contains Jira issues from the public Jira board. The judge agent also follows this same set-up, although using a Flowise chatflow instead. This setup aims to allow the AI agent to maximise the quality of the analysis generated.

After the analysis has finished generating, we also incorporated the LLM judge agent in this part. We load the analysis output from the previous step, reuse the sonar summary text for context, and build the judge prompt, which essentially instructs the judge agent to evaluate the LLM's output and return only JSON. The judge's output would be appended to the PR comment.

E. PR Feedback

This is the final part of the flow where the final payload is sent out as a PR comment. If the sonar scan doesn't find any issue, this comment will state that there are no potential TDs in the current PR. Otherwise, it will take the previous AI analyzer agent's analysis and append it with the judge's output as a comment. The final comment would include the TD issue number, issue description, including the sonar rule, affected code, is a technical debt field, justifications, rationale, TD severity, long and short-term effects, and suggested actions. The judge's evaluation details the final scores for both contextual and prescriptive aspects.The input consists of the standard SonarQube's message, while figure 3 shows the sample output of our system. It generates an analysis from the Sonar issue to explain TD's effect in great detail.

IV. Validation

This section will provide an explanation of our validation system as well as its results.

A. Validation Overview

Figure 4 shows the overview of our validation process. We first extract TD issues from the Technical Debt Dataset V2.0 [19]. We extract TD issues created between 2015 and 2022, querying the dataset. For each TD issue, we run the system that generates AI analysis. If the sonar scan did not detect any problem, or if the AI analysis deemed that there are no TD available, then further investigation is unnecessary. We marked the said issue as no TD and skipped it. Otherwise, the analysis generated by the AI analyser agent will be passed to the judge agent for evaluation. The judge agent will evaluate the analysis's prescriptive and contextual criteria and assign a final score ranging from one to five, one being very poor and five being very good. Finally, we record the scores of each TD item and visualise them in our paper.



Fig. 2. Sample of the PR feedback input from the SonarCloud that the AI analyzer agent would analyze.

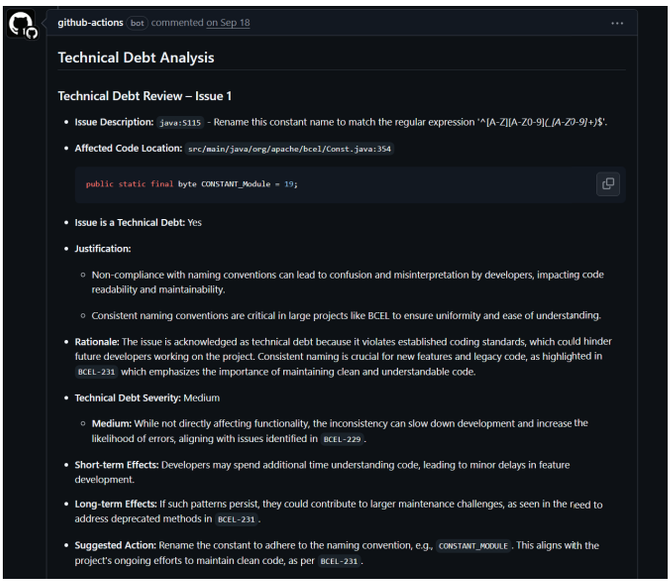


Fig. 3. Sample of the PR feedback, analysis of potential TD by the AI analyzer agent.

B. The Technical Debt Dataset

We use the Technical Debt Dataset, Version 2.0 for our study case and experiments [19]. This dataset contains TD issues that are linked with a Github commits and SonarQube. There are a total of 31 Apache projects, each with various numbers of TD issues found. We randomly picked two projects for our experimentation, which are Apache Commons Collections and Apache Commons BCEL. Each of them contains 55 and 53 TD issues respectively. We also assigned a time range of 2015 - 2022 from which TD issues were found. This generates a fault-inducing commit hash, its parent commit hash, the author's name, and the commit message for our validation process. These issues would then be used as the basis for the experiment, where the judge agent would evaluate the AI analysis.

We found that the three most common rules that were identified as TDs were squid: S1166, squid: S1213, and squid: S00117. Therefore, we selected the three issues in table I that align with the most common rules listed in this paper as a reference for the types of TD issues examined in this research. Interestingly enough, out of all 25 issues, only three matched the most common rules. This means that the most frequent TD is not necessarily the most impactful.

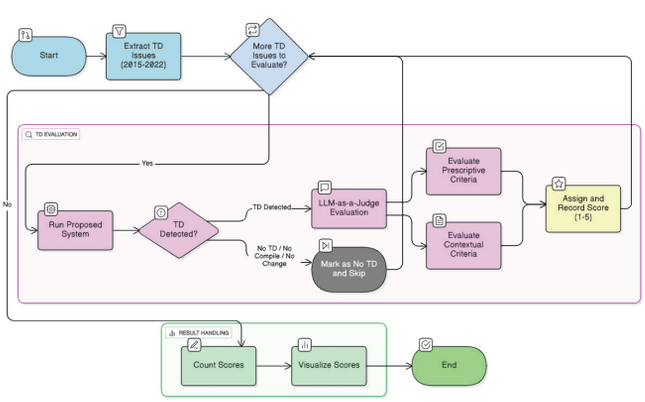


Fig. 4. The overview of how the evaluation is handled using the LLM-as-a-judge framework.

C. TD Analysis Evaluation with LLM-as-a-judge

We evaluate the analysis produced by the analyzer agent using an LLM-as-a-judge framework. This framework would allow scalable and human-free evaluation, which correlates with human assessment on reasoning and instruction-following tasks, whilst providing logical scores to the analysis [20]–[22]. We implemented the judge agent within Flowise's chatflow. The judge agent has the same project's context to evaluate its contextual score, but using a different LLM base model. As specified in Rule2Text, this evaluation method is considerably biased [20]. To minimise this bias, we used a different LLM model, specifically Gemini 2.5 Flash. To further reduce bias and overconfidence, we enforce JSON-only outputs and evaluate per criterion. We follow the Rule2Text framework to create our own LLM-as-a-judge, using Gemini as a base model, although it is an older Gemini 2.0 Flash. First, we make our criterion utilising a chain of thought prompt pattern for the judge agent.

Step 1: Verify Contextual Analysis

- (C1) Cites a relevant Jira key from the project context? (Yes/No)
- (C2) Provides a logical rationale linked to project goals? (Yes/No)
- (C3) Specifies a technical debt severity? (Yes/No)
- (C4) Describes contextual short- and long-term effects? (Yes/No)

Step 2: Verify Prescriptive Depth

- (P1) Explains short- and long-term effects of not fixing? (Yes/No)
- (P2) Are effects compelling and specific? (Yes/No)
- (P3) Suggests what to do if the issue is identified as technical debt? (Yes/No)

Step 3: Verify Quality of Generation

- (Q1) Is the analysis clear and well-written? (Yes/No)

TABLE I
EXAMPLE OF FAULT-INDUCING COMMITS AND THEIR PARENT COMMITS. NAMES ARE BLURRED OUT FOR PRIVACY

Fault Inducing Commit Hash	Fault Inducing Commit Parent Hash	Commit Message
7a2...7103	e05...51fe	Use final on local vars when possible (which already do on fields and parameters when possible.) git-svn-id: https://svn.apache.org/repos/asf/commons/proper/bcel/trunk@1749603 13f79535-47bb-0310-9956-ffa45
feb...8b40	f3f...4a17	BCEL-221 BCELifier is not working for Java8Example Workround for may need further work git-svn-id:https://svn.apache.org/repos/asf/comm /proper/bcel/trunk@1702447 13f79535-47bb-0310-9956-ffa450edef68
02f...28c9	bad...ccad	Gradually working towards restoring binary compatibility Revert Const back to interface git-svn-id: https://svn.apache.org/repos/asf/commons/p /bcel/trunk@1702355 13f79535-47bb-0310-9956-ffa450edef68

- (Q2) Includes all required structural components? (Yes/No)

Step 4: Final Scoring

- *Contextual Score:* 1–5 based on rules C1–C4.
- *Prescriptive Score:* 1–5 based on rules P1–P3.

Secondly, we integrate each judge agent of each project with its own knowledge base, similar to the analyzer agent's knowledge. Finally, for each issue we previously extracted, we run the judge agent after the analyzer agent has finished analyzing. This judge agent also follows previous studies, where the result of the LLM-as-a-judge evaluation is a Likert scale of 1 to 5 [20], [21].

Figure 5 shows a comparison in the contextual score of our LLM-centric system and SonarQube's default message against the judge agent. SonarQube's default messages show a mostly very poor performance, with 88% falling on the poor/very poor scores and 12% falling to the fair score. In contrast, our LLM-centric approach to contextual TD management generates somewhat good results. The results show 64% of scores were good / very good, 28% were fair, and 8% were poor. This shows a significant increase in performance from the default SonarQube messages to our LLM-based analysis. However, this score also indicates that LLM can hardly provide consistent, reliable contextual explanations, usually suffering from hallucination.

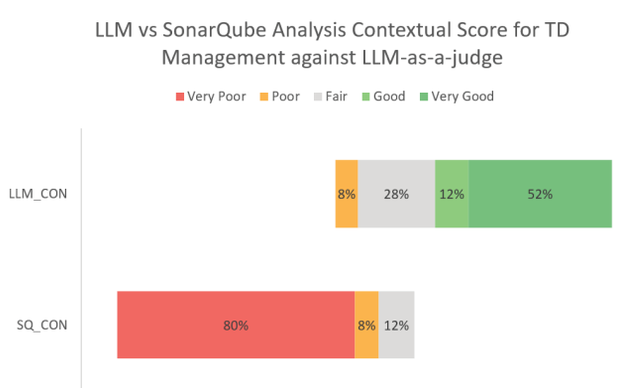


Fig. 5. AI analyzer agent VS SonarQube Analysis Contextual Score against LLM-as-a-judge. 64% of LLM outputs were rated good/very good; 0% of SonarQube outputs fell into good/very good.

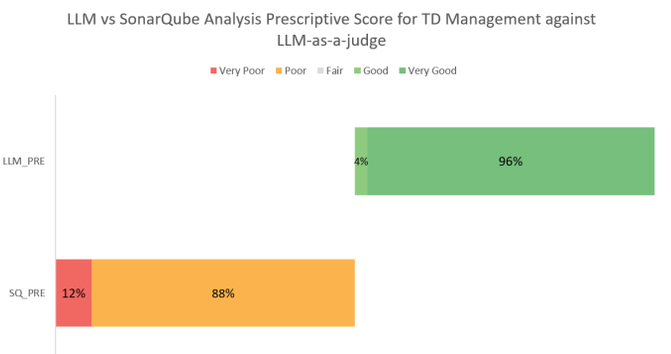


Fig. 6. AI analyzer agent VS SonarQube Analysis Prescriptive Score against LLM-as-a-judge. 100% of LLM outputs were good/very good; 0% of SonarQube outputs fell into good/very good.

Figure 6 details the prescriptive scores of our system's analysis and SonarQube's messages. We found a poor performance for SonarQube, with 100% of scores falling into poor/very poor categories. Our LLM-based analysis performed excellently, achieving 96% very good scores and 4% good scores. This highlights a significant improvement that our system generates for TD management from SonarQube's messages. LLM demonstrates its ability to create consistent prescriptive analysis in managing TD, guiding developers on how to handle TDs.

V. DISCUSSION

This section will discuss our findings and their correlation to our research questions.

RQ1. To what extent do LLM-based systems provide contextually relevant explanations for SonarQube-detected TD?

The contextual scores we found in figure 5 are generally high, but some lower values pull the mean down. Contextual explanations were considered good but slightly less consistent than prescriptive ones. Furthermore, LLM offers a substantial increase in contextually relevant TD management compared to the base SonarQube messages. Contextual scores also provide more variation, as some instances were identified where the judge agent saw missing context in the explanations. LLM can provide contextually relevant explanations to a great degree, but it is not perfect. This highlights potential areas where LLM contextual reasoning could be improved, such as project-specific requirements, deadlines, or more references to Jira issues.

RQ2. How effective are LLM-based systems in providing prescriptive guidance for addressing TD?

Almost all prescriptive scores we found in figure 6 are five, indicating that the judge agent consistently found the LLM's prescriptive analysis highly effective in guiding the addressing of TD. Compared to SonarQube, LLM provides a significant score improvement. Furthermore, we found that the judge agents mostly agree, with minimal variation in their judgments for all prescriptive scores. This means that the LLM is highly effective at generating prescriptive recommendations. It reliably informs developers on what to do, aligning with our goal of making TD handling actionable.

The results that we found are interesting. While the LLM-centric system works perfectly for prescriptive analysis, it shows a significant improvement in score against SonarQube's messages. Contextually, the AI analyzer system might produce mixed-quality analysis. Some analysis may correctly reference a logical Jira issue from the knowledge base. While other times, it hallucinates and cites its own non-existing Jira issue, or even doesn't cite one at all. At the same time, the scores show that the AI analyzer agent can provide good, relevant contextual analysis. Thus, this approach may have improved SonarQube's abilities to manage TD, but further work is needed to solidify our contextual management of TD.

We faced several hallucination issues when experimenting on the dataset. Thus, we also implemented several mitigation attempts to address this: **F1: Weak citation to project context.** Sometimes the analyzer mentions issues without explicit IDs. *Mitigation:* enforce a "must include BCEL-XXX/Collections-XXX" rule in prompts and filter responses lacking IDs. **F2: Over-general advice.** Fix suggestions too generic. *Mitigation:* require "file: line" references and one concrete diff-style suggestion when applicable.

We also face a challenge due to OpenAI's token limit, which restricts the number of Sonar issues, code snippets, and additional context that can be sent to the agents. This, in particular, could directly impact the outcome of the analysis in a contextual sense. Furthermore, not all extracted issues from the dataset can be detected as Sonar issues by SonarQube, and not all matters extracted from the dataset actually contain code changes; some detected fault-inducing commits only show changes.xml, and we only experimented with issues that have code changes in the codebase and are detected as an issue by SonarCloud.

A. Threats to Validity

We consider two main threats:

External Validity: Refers to the generalizability of our results. Experiments were conducted on only two Apache Java projects, Commons Collections and Commons BCEL; the results may not apply to other software ecosystems or projects. This is because we are limited in both time and cost. Broader validation across different projects, larger TD issues, and even different programming languages would improve confidence in the generalizability of our findings. Future work may expand the experiments to include a larger set of projects and issues to validate our approach to TD management further.

Conclusion Validity: Threats that affect the reliability of inferences. Our study did not include human evaluation and samples due to limited time and resources. The results we found may not fully represent the spectrum of issues or judgments. Future studies could address this issue by incorporating human evaluation as a base reference for the LLM-as-a-judge evaluation, solidifying the validation of our research.

VI. Data Availability

The artefact collected in this research, such as the dataset, Flowise's flow, prompt templates, etc., will be available at this link: https://doi.org/10.5281/zenodo.18033027.

VII. Conclusion

In this study, we propose an LLM-centric approach to managing technical debt at the PR level and time, utilising GitHub, SonarCloud, and Flowise for its implementation. By implementing an LLM-as-a-judge framework for our evaluation, we found that LLM proves to be very effective in generating prescriptive recommendations. LLM can also provide contextual explanations related to the project, but only to a limited degree. Hallucinations still prove to be a challenging aspect in generating contextually relevant explanations for TD management.

This research contributes to providing a prescriptive and contextual solution in TD management, albeit with some limitations in contextually relevant explanations.

There are still aspects of our research that can be expanded in the future. First, expanding the evaluation to include a larger set of open-source projects and multiple programming languages would enhance the generalizability of the LLM-based TD analysis. Secondly, integrating human expert validation alongside the LLM-as-a-Judge framework could provide a more robust benchmark for prescriptive and contextual scoring, improving the reliability of automated recommendations. Third, exploring real-time integration within CI/CD pipelines could enable the automated prioritisation and remediation of TD, thereby linking LLM recommendations directly to SonarQube quality gates or automated refactoring scripts. Furthermore, improving the RAG pipeline and fine tuning its indexing parameters may provide a more contextually relevant solutions for TD management. Finally, addressing limitations such as LLM hallucination, prompt sensitivity, and token efficiency through advanced prompt engineering or human-in-the-loop approaches with LLMs could further improve the precision and applicability of prescriptive guidance in software maintenance.

REFERENCES

[1] Z. Codabux, B. J. Williams, G. L. Bradshaw, and M. Cantor, "An empirical assessment of technical debt practices in industry," *Journal of software: Evolution and Process*, vol. 29, no. 10, p. e1894, 2017.

[2] H. Kleinwaks, A. Batchelor, and T. H. Bradley, "Predicting the dynamics of earned value creation in the presence of technical debt," *IEEE Access*, vol. 11, pp. 125381–125395, 2023.

[3] R. Banker, Y. Liang, and N. Ramasubbu, "Technical debt and firm performance," *Management Science*, vol. 67, no. 5, pp. 3174–3194, 2021.

[4] J. A. Ruiz, R. A. Aguilar, J. F. Garcilazo, and A. A. Aguileta, "A tertiary study on technical debt management over the last lustrum," *IJCOPI*, vol. 15, no. 5, p. 181, 2024.

[5] A. Martini, T. Besker, and J. Bosch, "Technical debt tracking: Current state of practice: A survey and multiple case study in 15 large organizations," *Science of Computer Programming*, vol. 163, pp. 42–61, 2018.

[6] D. Tsoukalas, N. Mittas, E.-M. Arvanitou, A. Ampatzoglou, A. Chatzigeorgiou, and D. Kechagias, "Local and global explainability for technical debt identification," *IEEE Transactions on Software Engineering*, 2024.

[7] P. Avgeriou, I. Ozkaya, A. Chatzigeorgiou, M. Ciolkowski, N. A. Ernst, R. J. Koontz, E. Poort, and F. Shull, "Technical debt management: The road ahead for successful software delivery," in *2023 IEEE/ACM ICSE-FoSE*, pp. 15–30, IEEE, 2023.

[8] Z. Codabux and B. J. Williams, "Technical debt prioritization using predictive analytics," in *Proceedings of the 38th ICSE-Companion*, pp. 704–706, 2016.

[9] W. N. Behutiye, P. Rodríguez, M. Oivo, and A. Tosun, "Analyzing the concept of technical debt in the context of agile software development: A systematic literature review," *Information and Software Technology*, vol. 82, pp. 139–158, 2017.

[10] L. Belzner, T. Gabor, and M. Wirsing, "Large language model assisted software engineering: prospects, challenges, and a case study," in *AISoLA*, pp. 355–374, Springer, 2023.

[11] A. Zhou, K. Wang, Z. Lu, W. Shi, S. Luo, Z. Qin, S. Lu, A. Jia, L. Song, M. Zhan, *et al.*, "Solving challenging math word problems using gpt-4 code interpreter with code-based self-verification," *arXiv preprint arXiv:2308.07921*, 2023.

[12] M. C. O. Silva, M. T. Valente, and R. Terra, "Does technical debt lead to the rejection of pull requests?," *arXiv preprint arXiv:1604.01450*, 2016.

[13] S. Karmakar, Z. Codabux, and M. Vidoni, "An experience report on technical debt in pull requests: Challenges and lessons learned," in *Proceedings of the 16th ACM/IEEE ESEM*, pp. 295–300, 2022.

[14] F. E. d. O. Calixto, E. C. Araújo, and E. L. Alves, "How does technical debt evolve within pull requests? an empirical study with apache projects," in *Simpósio Brasileiro de Engenharia de Software (SBES)*, pp. 212–223, SBC, 2024.

[15] A. Martini, "Anacondebt: A tool to assess and track technical debt," in *Proceedings of the 2018 International Conference on Technical Debt*, pp. 55–56, 2018.

[16] T. S. Mendes, D. A. Almeida, N. S. Alves, R. O. Spínola, R. Novais, and M. Mendonça, "Visminertd: an open source tool to support the monitoring of the technical debt evolution using software visualization," in *ICEIS*, 2015.

[17] B. Dornauer, M. Felderer, J. Weinzerl, M.-C. Racasan, and M. Hess, "Sohist: A tool for managing technical debt through retro perspective code analysis," in *Proceedings of the 27th EASE*, pp. 184–187, 2023.

[18] J. White, S. Hays, Q. Fu, J. Spencer-Smith, and D. C. Schmidt, "Chatgpt prompt patterns for improving code quality, refactoring, requirements elicitation, and software design," in *Generative ai for effective software development*, pp. 71–108, Springer, 2024.

[19] V. Lenarduzzi, N. Saarimäki, and D. Taibi, "The technical debt dataset," in *15th Conference on PROMISE*, January 2019.

[20] N. Shirvani-Mahdavi and C. Li, "Rule2text: A framework for generating and evaluating natural language explanations of knowledge graph rules," *arXiv preprint arXiv:2508.10971*, 2025.

[21] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, *et al.*, "Judging llm-as-a-judge with mt-bench and chatbot arena," *Advances in neural information processing systems*, vol. 36, pp. 46595–46623, 2023.

[22] A. Szymanski, N. Ziems, H. A. Eicher-Miller, T. J.-J. Li, M. Jiang, and R. A. Metoyer, "Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks," in *Proceedings of the 30th IUI*, pp. 952–966, 2025.

Comparative Analysis of LSTM and ARIMA Models for Stock Price Prediction in Sharpe Ratio-Based Portfolio Construction

1st Rifky Muhammad Haidar
Management Study Program, School of Economics and Business
Telkom University
Bandung, Indonesia
rifkyyhaidar@gmail.com

2nd Dwi Fitrizal Salim
Management Study Program, School of Economics and Business
Telkom University
Bandung, Indonesia
dwifitrizalslm@telkomuniversity.ac.id

***Abstract*— This study conducts a comparative analysis between the Long Short-Term Memory (LSTM) model and the Autoregressive Integrated Moving Average (ARIMA) model in predicting based on the Sharpe Ratio optimization framework. The dataset used consists of the daily closing price of 50 stocks on the Shanghai Stock Exchange, from which 10 uncorrelated stocks with the highest returns were selected thru a correlation-based selection process. The predicted return result from both models are used as input for portfolio optimization with the goal of maximizing the Sharpe Ratio under long-only constraints to achieve the best efficiency between risk and return. The experimental results show that the LSTM model has better prediction accuracy compared to ARIMA, with lower RMSE, MAE, and sMAPE values. The LSTM-based portfolio yielded a sharpe ratio of 3.2011, an annual return of 59,70% and a volatility of 18.65%, out performing the ARIMA portfolio. These findings demonstrate the effectiveness of deep learning models in improving prediction accuracy and portfolio efficiency within the framework of modern quantitative trading.**

Keywords— deep learning, long short term memory, autoregressive integrated moving average, sharpe ratio, shanghai stock exchange

I. INTRODUCTION

In the modern investment world, the strategy for optimal portfolio construction has become a crucial aspect of achieving a balance between risk and return. Since the introduction of Modern Portfolio Theory (MPT) [1], the mean-variance approach has become the cornerstone of investment decision-making. However, in practice, the assumptions of normal distribution and linear correlation between assets often do not reflect the dynamic and complex conditions of financial markets. Therefore, a more adaptive and intelligent approach is needed to capture nonlinear patterns between stock prices and optimize portfolios more accurately [2].

The development of Machine Learning (ML) and Deep Learning (DL) has opened new opportunities in the field of quantitative finance [3]. One of the models widely used in predicting stock prices is Long Short-Term Memory (LSTM), which can learn long-term dependencies in time series data [4]. LSTM has proven effective in capturing the nonlinear and volatile dynamic of the market, thus improving the accuracy of stock price predictions [5]. And Autoregressive Integrated Moving Average (ARIMA) as an established linear baseline for assessing the added value of non-linearity [6].

In the context of the Chinese stock market, particularly the Shanghai Stock Exchange (SSE), machine learning-based research is still rapidly developing. The SSE 50, which consists of 50 blue-chip stocks, reflects China's macroeconomic conditions and is one of the most influential indices in Asia. However, investors often face the challenge of high correlation between stocks, which can reduce the effect of portfolio diversification. Therefore, correlation analysis is needed to select a combination of stocks with low correlation, making the portfolio more efficient in managing risk [7].

This research can be summarized as follows: (1) proposing a systematic approach to stock selection and portfolio optimization by identifying stocks with high returns and low pairwise correlations, (2) using the Long Short-Term Memory (LSTM) model to predict the daily closing prices of the selected stocks, and comparing the prediction results with those obtained from the Autoregressive Integrated Moving Average (ARIMA) model, and (3) conducting a portfolio performance evaluation, which revealed that the constructed portfolio outperformed the benchmark in terms of active returns and risk control, with portfolio weights determined using accurate prediction results between LSTM and ARIMA.

II. LITERATURE REVIEW

This section will review the theoretical foundations and previous research that serve as the basis for the hybrid framework proposed in this study.

A. Stock selection using a correlation matrix

When analyzing stock movements, the relationship between stocks becomes very important to consider. The correlation coefficient is used to indicate the extent of the relationship between the movements of two variables relative to their respective deviations. [8]. This measurement serves as the basis for the portfolio diversification process, which is one way to manage investment risk. According to modern portfolio theory, first proposed by Harry Markowitz in 1952, investors can reduce portfolio risk without sacrificing their desired level of return by spreading or diversifying their investments across various asset classes with low correlation [7], [9].

B. Stock price prediction using LSTM

In this study, LSTM and ARIMA will be used to predict stock prices. LSTM is a deep learning method designed to address the problems of data sequence exploration and gradient vanishing [10]. LSTM, or Long Short-Term Memory, has been used in various time series studies, particularly for predicting stock market prices. [11] conducted research on the UK 100 stock exchange from March 1994 to March 2019, comparing LSTM with Support Vector Machine (SVM), Random Forest (RF), and Autoregressive Integrated Moving Average (ARIMA). The results show that LSTM outperformed other baseline models by an obvious margin. In contrast, (Kumar et al. 2022) found that Generative Adversarial Networks (GANs) performed better in terms of root mean squared error compared to LSTMs.

The research conducted by [12], [13] found promising results, showing that the LSTM model could detect price changes. [14], [15] also found that LSTM performed better in predicting stock prices (especially in expressing stock prices changes), while ARIMA was more practical to implement. [16] found that LSTM was quite effective in solving sequence prediction problems because it could store information from the past.

C. Stock price prediction using ARIMA

Stock Price Predicting using Autoregressive Integrated Moving Average (ARIMA). The ARIMA model is a linear framework for forecasting time series data that combines autoregressive (AR) components, integration/differencing (1) to achieve stationarity, and moving average (MA). This model is effective for relatively stable short-term patterns and serves as an interpretable statistical baseline for comparing the added value of nonlinear models.

The study [17] demonstrated the suitability of ARIMA and SARIMA for short-term stock price forecasting. Applying the ARIMA model to forecast the Bombay Stock Exchange index demonstrated its potential to accurately capture stock price movements. In a study of the energy company BP, a multinational oil gas company based in London, England, the ARIMA model provided accurate predictions, with a MAPE value of 1.47%, indicating a very low prediction error rate [18]. Similar results were also obtained from a study [19] comparing the performance of the ARIMA model with Support Vector Machine (SVM) in predicting the monthly stock prices of PT. Telkom Indonesia (TLKM). The results showed that ARIMA outperformed SVR, indicating better adaptability to the structured linear patterns found in stock price series.

D. Portofolio simulation

This research weights the portfolio by maximizing the Sharpe ratio. The Sharpe ratio, as measure of risk-adjusted performance, is a commonly used theory and formula when maximizing portfolio performance based on returns relative to risk [20] It is highly relevant when optimizing portfolio weights to maximize the Sharpe ratio. In [21] research, the Sharpe ratio proved to be a slightly stronger strategy because the model-free framework is not suitable for the domain of market asset allocation.

[22] showed that portfolio construction maximizing the Sharpe ratio yielded a cumulative return of 0.0909, significantly outperforming the CSI 300 index return of 0,0348. These findings indicate that the maximum Sharpe ratio strategy is effective in achieving a balance between returns and outperforming the market, offering valuable insights for investors seeking optimal portfolio allocation.

III. METHODOLOGY

A. Data Source and types

This research utilises 50 stocks included in the Shanghai Stock Exchange's 50 (SSE 50) index in 2025. The data used consists of the daily closing prices of each stock, covering a 10-year observation period from September 1, 2015, to September 1, 2025. Stock price data was obtained from investing.com website.

B. Data preprocessing

The initial stage of the research began with the data preprocessing process using historical return data from 50 stocks listed on the Shanghai Stock Exchange. The dataset is imported from an Excel file stored on Google Drive and processed using pandas and NumPy libraries. After the data is read, the correlation matrix between stocks is calculated using the Pearson coefficient to identify the level of interconnectedness between assets. For LSTM and ARIMA modelling, the data used consists of daily closing prices and utilizes 12 supporting variables representing technical and statistical factors: return, alpha, beta, variance, momentum, Simple Moving Average (SMA) 5, SMA 20, Exponential Moving Average (EMA) 12, EMA 26 MACD line, MACD signal, and log return. These variables were chosen because they can capture the relevant characteristics of market volatility, trends, and signal strength within the context of stock price movements.

C. Analysis of stock correlations

This stage aims to determine the relationship between stock returns using the Pearson correlation matrix:

$$\rho_{ij} = \frac{Cov(R_i, R_j)}{\sigma_i \sigma_j}$$

The results are visualized as a heatmap to identify stocks with low correlation. Stock selection was conducted in a data-driven manner with two main criteria: a pair correlation between stocks of less than 0 to maximize diversification and positive average returns throughout the observation period. Based on these criteria, 10 stocks with high return and low correlation characteristics were selected and subsequently used in the LSTM modelling stage.

D. Price predictions using LSTM

The LSTM model is built using TensorFlow or Keras with the following structure:

Layer	Layer Type	Main Parameter
Input Layer	Time Series Data (lag return)	window size = 60 days
Hidden Layer 1	LSTM Layer	50 neuron, activation = ‘tanh’
Hidden Layer 2	Dropout Layer	rate = 0.2
Hidden layer 3	LSTM Layer	50 neurons
Output layer	Dense Layer	1 neuron (Price prediction)

Table 1 Architecture model LSTM
source: data processed by the author

The model architecture consists of several layers. The input layer receives normalized time series data. Then, Hidden Layer 1 is an LSTM layer with 50 neurones and a tanh activation function, which is used to recognise nonlinear patterns in stock price changes. Hidden Layer 2 is used to prevent overfitting by ignoring some neurones during the

training process. Next, Hidden Layer 3 again uses an LSTM layer with 50 neurones to deepen the model's ability to understand the dynamics of historical data. Finally, the Output Layer uses a Dense Layer with one neurone that produces a prediction of the next stock price.

Parameter	Nilai
Optimizer	Adam
Learning Rate	0.001
Loss Function	Mean Squared Error (MSE)
Epochs	10
Batch Size	32
Validation Split	0.2

Table 2 Hyperparameter model LSTM
source: data processed by the author

This LSTM model is optimized using the Adam Optimizer algorithm, which is know for its efficiency in accelerating convergence during the neural network training process. The learning rate value is set to 0.001 so that the weight update process runs stably and is not too aggressive. The loss function used is Mean Squared Error (MSE), because MSE is suitable for measuring the difference between actual values and prediction results in stock data.

The training process was conducted over 10 epochs, with a batch size of 32, meaning the model updates its weight after every 32 data samples. Additionally, the training data was split with a validation split of 0.2, which means 20% of the training data was used as validation data to monitor the model's performance during training.

E. Price prediction using ARIMA

Model architecture

The Autoregressive Integrated Moving Average (ARIMA) models prices (or log-prices) as a combination of three components: AR(p), I(d), and MA(q). In summary:

- AR(p): the dependence of the current value on previous price lags.
- I(d): the d-order difference to achieve stationarity.
- MA(q): the current value depends on the previous error lags.

Formulation of the backward operator the backward operator is defined as By_t = y_(t-1), and the difference operator is represented as ∇^d = (1-B)^d.

$$\phi(B)\,\nabla^d y_t = c + \theta(B)\,\varepsilon_t, \phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p,\ \theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$$

where it is white noise and c is a constant/drift.

In this study, one-step-ahead forecasting was performed with a 60-day rolling window, so each prediction only used the last 60 observations to prevent time leakage.

Parameter	Value
ARIMA Order (p, d, q)	Simple grid : $p \in \{0,1,2,3\}$, $d \in \{0,1\}$, $q \in \{0,1,2,3\}$
Selection Criteria	Lowest AIC/BIC in the 60-day training window
Estimation	MLE (maximum likelihood)
Horizon	One-step-ahead
Scheme	60-day rolling window for each test date
Transformation	Optional log-price; reintegration to the level at the time of interpretation

Table 3 Hyperparameter model ARIMA
source: data processed by the author

In this study, the ARIMA model is optimized with parameters (p, d, q), which respectively represents autoregressive (p), differencing (d), and moving average (q) components. The values of the third parameter are determined through a grid search process with the following parameter ranges: p ∈ {0, 1, 2, 3}, d ∈ {0, 1}, and q ∈ {0, 1, 2, 3}. The best model selection process is based on the lowest Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values over a 60-day training window. This criterion is used to balance the accuracy and complexity of the model, resulting in an efficient one.

F. Prediction accuracy evaluation

The performance of the LSTM model was evaluated using three main metrics:

- **RMSE (Root Mean Squared Error)**

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n} (y_t - \hat{y}_t)^2}$$

where: y_t is the actual value at time t, $\hat{y}_t$ is the predicted value at time t, n is the number of observations, and t is the time index. RMSE measures the average magnitude of error in the original units and is more sensitive to outliers because the errors are squared.

- **MAE (Mean Absolute Error)**

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n} | y_t - \hat{y}_t |$$

where: y_t, $\hat{y}_t$, n, and t are defined the same as above. MAE measures the average absolute error, which is more robust to outliers than RMSE.

- **sMAPE (Symmetric Mean Absolute Percentage Error)**

$$\text{sMAPE} = \frac{100\%}{n}\sum_{t=1}^{n} \frac{2\ | y_t - \hat{y}_t |}{| y_t | + | \hat{y}_t |}$$

where: y_t, $\hat{y}_t$, n, and t are defined the same; use a small ε addition to the denominator when $|y_t| + |\hat{y}_t|$ approaches zero for numerical stability. sMAPE expresses the average error in percentage terms and is symmetrical with respect to the actual and predicted values, making it more stable when the price scale is small.

G. Portofolio simulation

The main goal of this optimization is to maximize the Sharpe Ratio while considering investment weighting constraints.

The formulation is written as follows:

$$\max_{w} \quad S = \frac{W^T R - R_f}{\sqrt{W^T \sum W}}$$

with the constraint:

$$\sum_{i=1}^{n} \quad w_i = 1, w_i \geq 0$$

These constraints indicate that the total investment weight proportion must sum to 1 (100%) and there are no short selling positions, making the portfolio long-only, as is common practice among real investors. This optimization is performed numerically using SciPy. Optimize method or a gradient-based optimizer algorithm, where the values of w_i will be adjusted to provide the highest Sharpe ratio.

The steps for constructing an optimal portfolio are as follows:

- Calculate the covariance matrix between stocks based on historical closing price data.
- Use the prediction results from the LSTM model as the expected return for each stock.
- Perform the optimization process to obtain the stock weights w_i that maximize the Sharpe ratio with the constraints $\sum w_i = 1$dan $w_i \geq 0$.
- Evaluate the portfolio results by calculating the annual return (R_p), annual volatility (σ_p), and Sharpe ratio of the optimization results.
- The results indicate taht the stock weight combination that yields the highest return per unit of risk, according to the best model's prediction (LSTM), is the optimal one.

IV. RESULTS

The result and analysis cover stock return prediction, stock risk modelling, and investment portfolio weights. Diversification, a key risk management tool, was achieved by selecting stock based on the correlation matrix

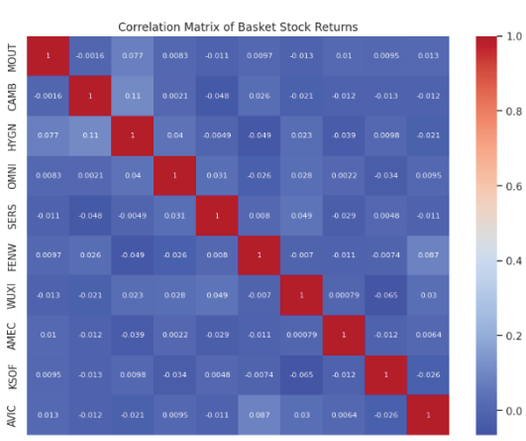


Figure 1 Correlation matrix of basket stock returns

Based on the results of the previous correlation matrix analysis, 10 stocks were selected that showed a combination of high returns and low correlation between stocks. The correlation matrix image illustrates the relationship between the price movements of the selected stocks, where most correlation values are below 0. This indicates that the price movements of these stocks are relatively independent of each other, so a decline in one stock does not significantly affect the others. Thus, a portfolio consisting of these stocks has higher potential for stability and an optimal level of diversification.

After selecting the stocks, the stock price was predicted using a machine learning model. The following table shows the 10 selected stocks along with their average returns and the correlation levels between them, which can be used as a reference for investment decisions based on portfolio diversification.

Ticker	Average Return	Average Correlation
MOUT	0.093033	0.020236
CAMB	0.088707	0.22116
HYGN	0.002552	0.031365
OMNI	0.001984	0.017584
SERS	0.001861	0.016438
FENW	0.001432	0.051127
WUXI	0.001391	0.019636
AMEC	0.001334	0.030642
KSOF	0.001309	0.017080
AVIC	0.001121	0.044856

Table 4 Average return and average correlation of selected stocks source: data processed by the author

Comparison of LSTM vs ARIMA:

Ticker	LSTM RMSE	LSTM MAE	LSTM sMAPE
AVIC	13.828	10.858	2.50
KSOF	15.03	12.73	4.82
AMEC	5.19	3.46	2.06
WUXI	3.22	2.72	4.81
FENW	8.92	7.87	3.54
SERS	10.86	7.29	7.10
OMNI	3.96	3.31	3.02
HYGN	27.44	25.84	19.97
CAMB	0.82	0.65	1.83
MOUT	3646.75	2843.07	1.75

Table 5 Evaluation metrics for LSTM source: data processed by the author

Ticker	ARIMA RMSE	ARIMA MAE	ARIMA sMAPE
AVIC	12.745	10.031	24.89
KSOF	49.82	37.12	13.81
AMEC	42.13	32.73	20.21
WUXI	16.14	13.19	21.54
FENW	72.89	64.41	26.10
SERS	77.96	63.84	106.83

OMNI	32.12	24.24	24.59
HYGN	63.08	60.95	54.96
CAMB	4.19	3.54	9.87
MOUT	50151.43	44039.92	23.94

Table 6 Evaluation metrics for ARIMA
source: data processed by the author

Based on the comparison table of LSTM and ARIMA predictions, LSTM is consistently more accurate, with lower RMSE, MAE, and sMAPE compared to ARIMA. For example, for AVIC stock, LSTM recorded an RMSE of 1.3828 and an sMAPE of 2.50%, while ARIMA recorded 12.745 and 24.89%. The difference is even more extreme for MOUT stock, where ARIMA struggled to capture high volatility (RMSE 50,151.43 vs. 3,646.75 for LSTM). These results indicate that LSTM excels at predicting stock prices, especially for data with high fluctuations or non-linear patterns.

Model	Mean RMSE	Mean MAE	Mean sMAPE
LSTM	372.76	290.23	**5.15%**
ARIMA	5091.50	4512.80	**32.58%**

Table 7 Average metric statistics
source: data processed by the author

LSTM has significantly lower error values across all metrics, with an average sMAPE of 5.15%, making it far more accurate than ARIMA (32.58%). It is evident that LSTM has lower error values across all metrics compared to ARIMA, which is because ARIMA struggles to capture non-linear patterns and high volatility in stock price time series.

After obtaining more accurate prediction results, weighting was then performed with the goal of the Maximum Sharpe Ratio Long Only strategy, which is to maximize returns per unit of risk taken. This exercise was done to achieve an optimal portfolio. The following figure display the annual portfolio weights obtained by maximizing the Sharpe Ratio.

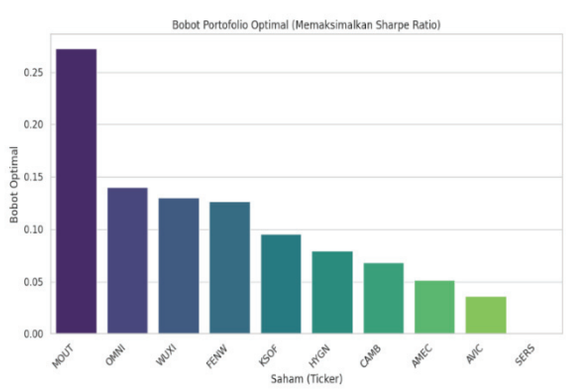


Figure 2 Optimal Portfolio Weights under Sharpe Ratio Maximization
source: data processed by the author

A portfolio with the highest Sharpe ratio is considered optimal because it generates the best risk-adjusted return. The LSTM prediction portfolio did well in this study, with a Sharpe ratio of 3.2011, an annual return of 59.70%, and an annual volatility of 18.65%. MOUT shares have the highest weight, which is 0.2725. Other significant weights are in OMNI (0.1402), WUXI (0.1302), and FENW (0.1263), indicating that these stocks are the main drivers of risk-adjusted portfolio performance. CAMB (0.0683), HYGN (0.0792), AMEC (0.0516), and KSOF (0.0955) hold smaller weights. SERS shares have a weight of 0.0000, indicating that they do not contribute to the portfolio's return-to-risk ratio. This finding strengthens the evidence that deep learning-based approaches are more adaptive to nonlinear patterns and stock market volatility, enabling them to form more risk-efficient portfolios compared to traditional statistical models.

V. CONCLUSIONS

This research proves the effectiveness of integrating machine learning-based prediction results and portfolio optimization in improving investment performance. Using daily closing price data for 50 stocks listed on the Shanghai Stock Exchange, 10 stocks were selected using a correlation matrix with the criteria of low correlation below 0 and high returns to form a diversified portfolio. The LSTM model was used to predict stock prices and compared to the ARIMA model to assess prediction accuracy and its impact on portfolio performance.

Empirical results show that the LSTM model consistently outperforms ARIMA, with lower RMSE, MAE, and sMAPE values, making it better at capturing nonlinear and dynamic price patterns. The predicted results from the LSTM are then used in the Sharpe Ratio optimization framework to form an optimal portfolio. The resulting portfolio achieves an annual return of 59.70%, a volatility of 18.65%, and a Sharpe Ratio of 3.2011. This finding confirms that deep learning models like LSTM can improve portfolio efficiency thru more accurate prediction results.

REFERENCES

[1] H. Markowitz, “Portfolio Selection,” 1952.

[2] R. Yan, J. Jin, and K. Han, “Reinforcement learning for deep portfolio optimization,” *Electronic Research Archive*, vol. 32, no. 9, pp. 5176–5200, 2024, doi: 10.3934/ERA.2024239.

[3] M. Obthong, N. Tantisantiwong, W. Jeamwatthanachai, and G. Wills, “A survey on machine learning for stock price prediction: Algorithms and techniques,” in *FEMIB 2020 - Proceedings of the 2nd International Conference on Finance, Economics, Management and IT Business*, SciTePress, 2020, pp. 63–71. doi: 10.5220/0009340700630071.

[4] I. H. Sahid and I. Budi, “Leveraging LSTM Predictions for Enhanced Portfolio Allocation with Markowitz Mean-Variance Optimization,” *The Indonesian Journal of Computer Science*, vol. 13, no. 6, Jan. 2025, doi: 10.33022/ijcs.v13i6.4573.

[5] H. Y. Kim and C. H. Won, “Forecasting the volatility of stock price index: A hybrid model integrating LSTM with multiple GARCH-type models,” *Expert Syst Appl*, vol. 103, pp. 25–37, Aug. 2018, doi: 10.1016/j.eswa.2018.03.002.

[6] J. Shah, D. Vaidya, and M. Shah, “A comprehensive review on multiple hybrid deep learning approaches for stock prediction,” Nov. 01, 2022, *Elsevier B.V.* doi: 10.1016/j.iswa.2022.200111.

[7] D. P. Manurung and I. Palupi, “Penglolaan Portoflio Investasi dengan Penerapan LSTM pada Sektor Strategis Indonesia dengan Evaluasi Sharpe Ratio dan Mean Variance,” 2025.

[8] I. Yunita, “Markowitz Model Dalam … Markowitz Model Dalam Pembentukan Portofolio Optimal (Studi Kasus Pada Jakarta Islamic Index) Markowitz Model In Optimal Portfolio Formation (Case Study In Jakarta Islamic Index),” 2018. [Online]. Available: Https://Www.Researchgate.Net/Publication/326298907

[9] Andreansyah's, Sidiq Saeful's, Putri Sharikha's, and Dasman Sunitas's, "Pengaruh Strategi Diversifikasi Portofolio Dalam Mengelola Risiko Dan Memaksimalkan Return Investasi," 2025, doi: 10.37817/jurnalmanajemen.v12i1.

[10] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[11] W. Wang, W. Li, N. Zhang, and K. Liu, "Portfolio formation with preselection using deep learning from long-term financial data," *Expert Syst Appl*, vol. 143, Apr. 2020, doi: 10.1016/j.eswa.2019.113042.

[12] A. Moghar and M. Hamiche, "Stock Market Prediction Using LSTM Recurrent Neural Network," in *Procedia Computer Science*, Elsevier B.V., 2020, pp. 1168–1173. doi: 10.1016/j.procs.2020.03.049.

[13] Z. Li, H. Yu, J. Xu, J. Liu, and Y. Mo, "Stock Market Analysis and Prediction Using LSTM: A Case Study on Technology Stocks," *Innovations in Applied Engineering and Technology*, pp. 1–6, Nov. 2023, doi: 10.62836/iaet.v2i1.162.

[14] R. Xiao, Y. Feng, L. Yan, and Y. Ma, "Predict stock prices with ARIMA and LSTM," Aug. 2022, [Online]. Available: http://arxiv.org/abs/2209.02407

[15] E. G. A. Osman and F. A. Otaibi, "Integrating deep learning and econometrics for stock price prediction: A comprehensive comparison of LSTM, transformers, and traditional time series models," *Machine Learning with Applications*, vol. 22, p. 100730, Dec. 2025, doi: 10.1016/j.mlwa.2025.100730.

[16] A. K. Singh, A. Gupta, F. Rabbani, A. Yadav, and Mr. A. P. Singh, "Stock Prediction using LSTM Technique," *Int J Res Appl Sci Eng Technol*, vol. 12, no. 5, pp. 1352–1355, May 2024, doi: 10.22214/ijraset.2024.61815.

[17] J. Patel, S. Shah, P. Thakkar, and K. Kotecha, "Predicting stock market index using fusion of machine learning techniques," *Expert Syst Appl*, vol. 42, no. 4, pp. 2162–2172, 2015, doi: 10.1016/j.eswa.2014.10.031.

[18] R. Dwi Karunia, N. Hidayati, J. Soekarno Hatta, T. Kulon, K. Pedurungan, and K. Semarang, "Implementasi Model ARIMA untuk Analisa Prediksi Harga Penutupan Saham Historis Harian Perusahaan BP Minyak dan Gas Multinasional," vol. 13, no. 3, 2025, doi: 10.26418/justin.v13i3.90747.

[19] A. Amalia, I. Hayati, A. Afandi, A. S. Lubis, and J. L. Marpaung, "Comparative Forecasting of Indonesian Stock Prices Using ARIMA and Support Vector Regression: A Statistical Learning Approach," *Mathematical Modelling of Engineering Problems*, vol. 12, no. 7, pp. 2307–2315, 2025, doi: 10.18280/mmep.120711.

[20] W. F. Sharpe, "Sharpe1966," 1966, Accessed: Nov. 04, 2025. [Online]. Available: Sharpe, W. F. (1964). Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk. The Journal of Finance, 19(3), 425–442. https://doi.org/10.2307/2977928

[21] Z. Liu, "Reinforcement Learning in Portfolio Management with Sharpe Ratio Rewarding Based Framework," European Alliance for Innovation n.o., Mar. 2023. doi: 10.4108/eai.18-11-2022.2327121.

[22] Y. Hu, "Portfolio Optimization Using Machine Learning Method and Monte Carlo Simulation," 2024. [Online]. Available: https://cn.investing.com

An Efficient Hybrid Transformer–XGBoost Model for Improved Accuracy in Thermal-Based Rice Seed Infection Detection

1st Ani Dijah Rahajoe *, 2nd I Gede Susrama Mas Diyasa †, 3rd Basuki Rahmat ‡, 4th Agussalim §
* *Department of Computer Science , Universitas Pembangunan Nasional Veteran Jawa Timur, Surabaya, Indonesia*
† *Department of Computer Science , Universitas Pembangunan Nasional Veteran Jawa Timur, Surabaya, Indonesia*
‡ *Department of Computer Science , Universitas Pembangunan Nasional Veteran Jawa Timur, Surabaya, Indonesia*
§ *Department of Computer Science , Universitas Pembangunan Nasional Veteran Jawa Timur, Surabaya, Indonesia*
Emails: *anidijah.if@upnjatim.ac.id, igsusrama.if@upnjatim.ac.id, basukirahmat.if@upnjatim.ac.id, agussalim.si@upnjatim.ac.id*

Abstract*—Rice seed infection by *Fusarium spp.* can significantly reduce crop productivity, making early detection a crucial step in precision agriculture systems. This research proposes a hybrid deep learning method combining the Transformer and XGBoost (Extreme Gradient Boosting) architectures for early detection of rice seed infection using daily temperature data from an infrared thermal camera. Temperature data from three rice varieties—Ciherang, IR64, and Mekongga—were collected for eight consecutive days and classified into two categories: healthy and infected. Transformer was used as a temporal feature extractor, while XGBoost served as the final classification model. Experimental results showed that the Transformer–XGBoost hybrid model achieved a peak accuracy of 98% on day 5, enabling infection identification three days faster than the conventional method (day 8). For comparison, the CNN (Convolutional Neural Network) model achieved 96% accuracy on day 7. The Transformer–XGBoost approach also showed stable performance with a minimum accuracy of 96 percent since the second day for all observed rice varieties. These results demonstrate that the proposed hybrid model provides an efficient, non-destructive, and reliable solution for temperature-based seed infection monitoring and has the potential to be applied in automated disease-detection systems in precision agriculture.

***Keywords*— Transformer; XGBoost; Convolutional Neural Network; early detection; rice seeds.**

I. Introduction

Rice (*Oryza sativa*) is a strategic commodity for national food security, and its productivity is heavily influenced by environmental factors and disease. One common pathogen that attacks rice seeds is a fungus in the genus Fusarium, which can cause reduced germination, systemic infections, and even crop failure. Early detection of seed infection is a crucial step in precision agriculture management systems to prevent further pathogen spread and minimize economic losses. This research used a thermal dataset obtained from infrared camera images of three rice seed varieties: *Ciherang*, *IR64*, and *Mekongga*. Observations were conducted daily from day 0 (before planting) to day 7. Each data point reflects the average seed temperature taken from high-resolution thermal camera captures. Based on physiological studies and laboratory practices, *Fusarium spp*. infections can generally be identified only visually or by conventional testing on day 8 or later. Therefore, this research focuses on using pre-day 8 thermal data to detect possible infections as early as possible, aiming to provide a rapid, non-destructive temperature-based classification system.

Machine learning approaches, particularly Convolutional Neural Networks (CNNs), have been widely used for image and sequential data classification, including for interpreting temperature patterns. However, CNN models have limitations in adapting to the complex temporal dynamics of daily temperature data [1]. Meanwhile, the Transformer architecture, which is more flexible in handling long-term time dependencies, when combined with a classification model like XGBoost, offers potential as a lightweight yet accurate classification system. Similar hybrid frameworks combining sequential neural networks and gradient boosting have also been successfully applied in other domains, such as indoor positioning using Wi-Fi RSSI based on an LSTM-XGBoost combined model [2]. Hybrid methods that combine boosting techniques and transformer models have proven effective in reconstructing infrastructure monitoring data, such as in the recovery of bridge monitoring data using the AdaBoost-Transformer model [3]. Several hybrid neural network approaches have proven effective for various classification tasks. For example, a combination of CNN-RNN and CNN-MLP has been successfully applied in road damage detection to capture both spatial and sequential features [4]. A similar approach inspired the use of Transformer–XGBoost in this study to more efficiently handle temporal temperature dynamics.

This research examines two main approaches CNN and a combination of Temporal Transformer with XGBoost in a binary classification task for detecting rice seed infections using daily temperature data. A comprehensive evaluation was conducted over the entire observation period (days 0 to 8) to identify the optimal early-detection window. Furthermore, model performance was compared across rice varieties to evaluate their generalizability and robustness to biological variation.

The main contributions of this research include development of a Transformer-XGBoost hybrid approach capable of performing daily temperature-based classification with low complexity, comparative analysis per day and per variety shows that CNN has the highest overall accuracy, but Transformer+XGBoost is able to provide stable detections starting from day 2, visualization of feature representation using PCA and interactive visuals provides in-depth insight into class separability.

This research is structured as follows: Section II describes the related work; Section III presents the proposed method; Section IV describes research results and discussion; and

Section V concludes with the findings and provides suggestions for the development of future thermal-based early detection systems.

II. Related Work

A. Convolutional Neural Network (CNN)

A Convolutional Neural Network (CNN) is a deep neural network architecture commonly used for image classification and spatial pattern recognition. However, CNNs can also be applied to numerical time series data, such as temperature, by treating the time series as a single spatial dimension. In this context, convolution filters are used to recognize patterns of local temperature changes across observation times.

CNNs are capable of detecting local trends and inter-day temperature variations, but their performance tends to be limited in capturing long-term temporal relationships, especially as the number of observation days increases. Therefore, CNNs are often used as baseline models in time series classification research.

B. Transformers

Transformer is a deep learning architecture based on self-attention mechanisms and has become the new standard for processing sequential data [5]. Although originally designed for natural language processing, the Transformer has been successfully adapted to a range of other domains, including numerical data classification and time series analysis. Transformers primarily use the Self-Attention mechanism, which is formulated as (1)[6]

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

Where Q is the query matrix generated from the input projection, K is the key matrix, and V is the value matrix. Meanwhile, d_k represents the dimension of the key used to normalize the dot-product result in the self-attention mechanism. The self-attention output will go into the feed-forward layer with residual connections and the normalization layer, and formulated as (2)[6]:

$$Output_{encoder} = LayerNorm(X + FFN(Attention(Q,K,V))) \quad (2)$$

In the equation above, X represents the input to the encoder layer. The Attention (Q,K,V) function is a self-attention mechanism that extracts relationships between elements in the input sequence. The results of the attention process are then processed by the *FFN(·)*, a feed-forward network consisting of two dense nonlinear layers that operate independently at each time point. The *FFN* output is added to the input X via a residual connection, and the final result is normalized using *LayerNorm(·)* to stabilize training and accelerate convergence [7]. The enhanced performance of hybrid Transformer architectures suggests that the synergistic integration of complementary neural network components may provide superior capability in capturing temporal variations and long-range dependencies [7].

In this research, the Transformer functions as a feature extractor from daily temperature data recorded over eight days. Transformer's self-attention allows the model to consider the entire range of temperature data globally, allowing for simultaneous modeling of inter-day relationships without relying on sequential position. The Transformer's feature representation is then used as a classification input.

C. Extreme Gradient Boosting (XGBoost)

XGBoost is a decision tree-based boosting algorithm designed for high efficiency and superior classification performance. It works by incrementally building a model, with each iteration improving the previous model's prediction error.

XGBoost builds a predictive model as a stepwise summation of decision trees, as formulated below (3)[8][9][10].

$$\hat{y}_i = \sum_{k=1}^{K} f_k\ (x_i),\ f_k \in \mathcal{F} \quad (3)$$

Where $\hat{y}_i$ is the model's prediction for the *i*-th data point, f_k represents the *k*-th decision tree generated during the boosting process, and $\mathcal{F}$ is the function space encompassing all possible decision trees generated by the model. Each final prediction is the sum of the contributions from all trained trees [11]

XGBoost is particularly well-suited for numerical data such as temperature due to its ability to handle non-linear relationships between features and its tolerance for outliers. In this approach, XGBoost is used as the final classification stage after daily temperature features are extracted by a Transformer. This combination allows for more accurate classification with relatively short training times. The XGBoost component of the hybrid framework has been widely recognized for its high predictive accuracy and robustness in nonlinear regression problems, such as housing price forecasting [12]

III. Proposed Method

A. Datasets

Temperature measurements were conducted on three agronomically significant rice varieties: *Ciherang*, *IR64-*, and *Mekongga*. Each seed sample was monitored daily for its surface temperature over an 8-day period, from Day 0 to Day 7. This temperature data was used as a sequential (temporal) input feature in model training. For each variety, 200 seed samples were observed at 4 measurement points per day, yielding a total of 4,800 data. At the end of the observation period, the infection status of each seed was determined and assigned a binary classification label: "Positive" for seeds infected with *Fusarium spp*. and "Negative" for healthy seeds. These labels were used as prediction targets in the time-based temperature classification task. The dataset, consisting of 4,800 temperature datasets, was divided into two parts: 80% for training (3,840 samples) and 20% for validation (960 samples). The division was performed at random, maintaining a balanced ratio of infected to uninfected seeds.

B. Proposed Method: Transformer + XGBoost

The main proposed method is a hybrid approach that combines a Transformer as a time-series feature extractor with XGBoost as the final classification model. The Transformer architecture is used to capture the temporal dynamics of daily temperatures, leveraging self-attention to form a latent representation of the temperature data [13][14]. The Transformer's final output is a fixed-dimensional vector that comprehensively represents each temperature sample. Hybrid or multi-scale model optimization strategies have also been applied in other domains, such as the OMS-CNN framework for lung nodule detection [15].

These feature vectors are then fed to the XGBoost algorithm for classification. XGBoost was chosen for its ability to handle tabular data with non-linear relationships and for its efficient training. This combination results in a model that is computationally lightweight yet competitive in performance. Fig.1 shows the proposed method.

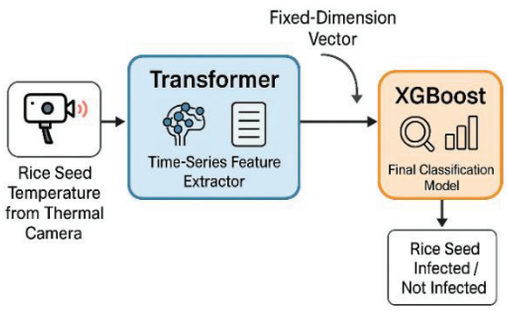


Fig. 1. Proposed Method.

C. *Comparative Model: Convolutional Neural Network (CNN)*

For comparison, a one-dimensional Convolutional Neural Network (CNN) model was used, applied directly to the daily temperature vector. The CNN architecture consists of: one 1D convolutional layer with a small kernel to capture day-to-day temperature changes, followed by a ReLU activation function and pooling, and then two fully connected layers to produce a two-class output.

The CNN was trained end-to-end using the cross-entropy loss function and the Adam optimization algorithm. The CNN was chosen as the baseline due to its effectiveness for simple time-series classification tasks, and its performance has been extensively demonstrated in previous studies [16].

D. *Experimental Parameters*

The following Table 1 summarizes the parameters used in both approaches.

TABLE I. PARAMETERS

Components	Parameter
Optimizer	Adam (for CNN dan Transformer)
Learning rate	0.001
Epochs	30
Batch size	32
Loss Function	Cross-Entropy (CNN & Transformer)
XGBoost Estimators	300 trees
XGBoost Learning Rate	0.05
Early Stopping	20 rounds (XGBoost)
Evaluation	Accuracy and Loss Evaluation (per day & per variety)

IV. RESEARCH RESULTS AND DISCUSSION

A. *Research Results*

Experiments were conducted on two models: a CNN and a Transformer-XGBoost combination, using temperature data from day 0 to day 8 for three rice seed varieties (*Ciherang*, *IR64*, *Mekongga*). The main objective was to evaluate each model's performance in detecting seed infection as early as possible, without waiting until day 8.

Fig. 2 shows the PCA (Principal Component Analysis) visualization of the latent features generated by the Transformer model on the validation data. The blue dots represent class 0 (uninfected), while the orange dots represent class 1 (infected). The distribution of the dots shows that the features of the two classes are beginning to separate clearly in two-dimensional space, particularly on the PC1 axis, indicating that Transformer successfully extracted discriminative feature representations between the two labels. Fig. 3 shows the PCA visualization of the latent features generated by the CNN model on the validation data. Blue dots represent class 0 (uninfected), while orange dots represent class 1 (infected). The data distribution in the PCA space shows a fairly clear separation between the two classes, especially along the principal component (PC1), though there is some overlap. This indicates that the CNN can extract representative features to differentiate infection status, albeit with slightly greater variation than the Transformer model.

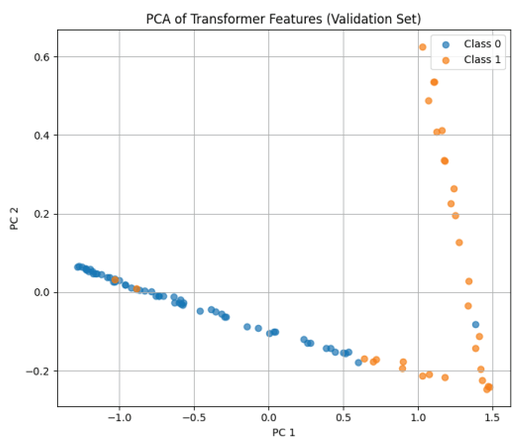


Fig. 2. PCA of Transformer Features (Validation Set)

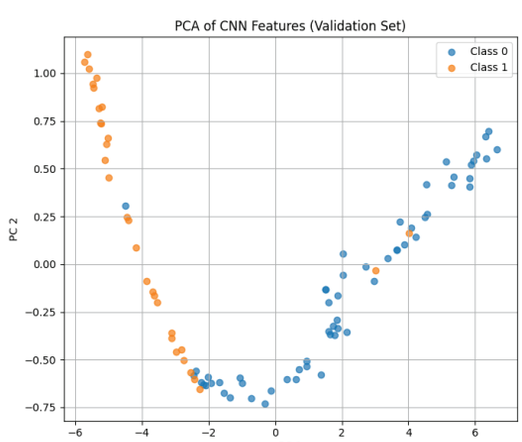


Fig. 3. PCA of CNN Features (Validation Set)

Fig. 4 shows the accuracy and loss performance of the CNN and Transformer+XGBoost models over 9 days of rice seed temperature observations. The horizontal axis shows the days on which temperature data were used as model input, from day 1 to day 9. The vertical axis shows the accuracy and loss values. The CNN model showed an increasing trend in accuracy from day 1 to day 8, with peak values on days 7 and 8. The CNN loss value gradually decreased over the same period. Conversely, the accuracy value decreased significantly on day 9, accompanied by a sharp increase in loss. Meanwhile,

the Transformer+XGBoost model demonstrated high, relatively stable accuracy from days 2 to 9. The loss value for this model was also low and showed little fluctuation over the observation period.

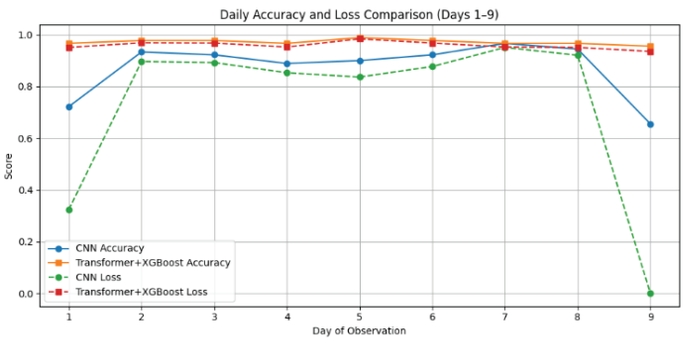


Fig. 4. Daily Accuracy and Loss Comparison (Days1-9)

These results provide an initial overview of each model's performance trends as a function of the length of the temperature data period used as input.

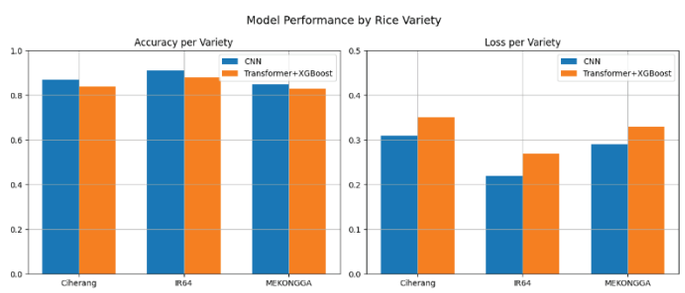


Fig. 5. Model Performance by Rice Variety

Fig. 5 shows the performance of the CNN and Transformer+XGBoost models based on the following rice varieties: *Ciherang*, *IR64*, and *Mekongga*. The left panel displays each model's accuracy, while the right panel shows the loss values.

For all three varieties, the CNN model produced higher accuracy than Transformer+XGBoost. Similarly, the CNN loss values tended to be lower across all varieties, especially for *IR64*. These results illustrate the differences in model performance across the characteristics of each rice variety.

B. Discussion

The CNN model achieved its highest accuracy of 96% on the 7th day of observation. Meanwhile, the combined Transformer and XGBoost model demonstrated its best performance earlier, on day 5, with a peak accuracy of 98%. These findings indicate that Transformer+XGBoost can identify infections earlier than CNN. able 2 shows the highest accuracy on day 5 for Transformer+XGBoost and day 7 for CNN.

TABLE 2. HIGHEST ACCURACY

Model	Day to	Highest Accuracy
CNN	7	0.96
Transformer + XGBoost	5	0.98

PCA analysis of the features extracted by each model (Fig. 2 and Fig. 3) shows that the Transformer features produce a fairly good class separation, albeit somewhat linear. Meanwhile, the CNN produces a more nonlinear feature distribution, which likely contributes to its better performance on complex data.

Experimental results show that the combined Transformer + XGBoost model achieves consistent, accurate performance in detecting rice seed infection using daily thermal temperature data. From Fig. 4, the daily accuracy and loss graph, the Transformer + XGBoost model maintains high accuracy ($\geqslant$0.96) and low loss from day 2 to day 9, with a peak on day 5. In contrast, CNN shows greater fluctuations, with a sharp decrease in accuracy on day 9, indicating sensitivity to data variations. Similar to the unified ARP–ViT–CNN framework proposed by [7], which integrates convolutional, transformer, and geometric feature extraction for improved diagnostic accuracy, our Transformer + XGBoost hybrid model leverages both temporal and contextual information to achieve early and robust infection detection.

In the performance graph (Fig. 5), CNN has a slight advantage in accuracy over Transformer + XGBoost across all varieties (*Ciherang*, *IR64*, *Mekongga*), but experiences slightly higher loss on Transformer + XGBoost, especially on *Mekongga*. This indicates that although Transformer + XGBoost is more stable, CNN has an advantage in generalization for certain types of varieties.

Based on the results of the early-detection accuracy evaluation (Figure 5), the Transformer model combined with XGBoost achieved a high level of accuracy from the first day of observation (around 0.96) and maintained stable performance through day 5. This indicates that the model can detect infections earlier, even before day 8, when physiological symptoms of the disease typically appear. In contrast, the CNN model requires a longer observation period to achieve consistently high accuracy, with the best performance only achieved after day 5.

Overall, Transformer + XGBoost excels in early detection and cross-condition performance stability, while CNN demonstrates strength in handling data variation complexity but requires more observation data to achieve maximum performance. The Transformer + XGBoost approach can be an efficient solution for rapid implementation in temperature-based disease-detection systems for rice seeds.

V. CONCLUSION

This research proposes a hybrid Transformer + XGBoost method for early detection of *Fusarium spp*. infection in rice seeds based on temporal temperature data from a thermal camera. Based on the evaluation results, the proposed method demonstrated superior performance compared to a CNN, especially for early detection. The Transformer + XGBoost model achieved the best accuracy of 0.99 on day 5, significantly earlier than conventional approaches, which generally identify infections on day 8 based on biological symptoms.

The main contribution of this research is demonstrating that the combination of Transformer and XGBoost can effectively capture the temporal dynamics of seed temperature and provide accurate predictions within just five days of observation. This provides significant opportunities to implement an automated, efficient seed selection system capable of early infection detection and disease spread prevention at an agronomic scale. Future research is

recommended to expand the dataset to include a wider range of varieties and environmental conditions, and to test the integration of this method into an automated temperature-monitoring system. Furthermore, developing model interpretability analysis and direct biological validation are needed to ensure classification accuracy in real-world applications in the seed industry.

REFERENCES

[1] S. Mehta, "A Comparative Analysis of CNN, LSTM, and CNN-LSTM Models for Kidney Tumor Diagnosis in Medical Imaging," in 2025 International Conference on Computing Technologies (ICOCT), Bengaluru, India, 2025. doi: https://doi.org/10.1109/ICOCT64433.2025.11118720.

[2] Jinlong Li; Jungyu Hwang; Joongoo Park, "Indoor Positioning Using Wi-Fi RSSI Based on LSTM-XGBoots Combined Mode," in 2025 Sixteenth International Conference on Ubiquitous and Future Networks (ICUFN), Lisbon, Portugal, 2025. doi: https://doi.org/10.1109/ICUFN65838.2025.11170030.

[3] Z. Liu;, F. Li;, X. Yu;, H. Yu;, and J. Guo, "Bridge Monitoring Data Recovery Based on Adaboost-Transformer Model," in 2024 43rd Chinese Control Conference (CCC), Kunming China, 2024. doi: https://doi.org/10.23919/CCC63176.2024.10661805.

[4] A. D. Rahajoe, M. Suriansyah, and A. A. B. Jr, "Hybrid Neural Network-Based Road Damage Detection Using CNN-RNN and CNN-MLP Models," Jurnal Teknik Informatika (Jutif), vol. 6, no. 3, pp. 1217–1228, 2025, doi: 10.52436/1.jutif.2025.6.3.4435.

[5] Y. Wen, M. A. Mahmoud, W. Wu, H. Chen, Y. Zhang, and X. Pan, "Predicting pathological subtypes of pure ground-glass nodules using Swin Transformer deep learning model," Springer Nature Link, 2025, doi: 10.1186/s13244-025-02113-3.

[6] K. Mohiuddin et al., "Attention Is All You Need," International Conference on Information and Knowledge Management, Proceedings, no. Nips, pp. 4752–4758, 2023, doi: 10.1145/3583780.3615497.

[7] J. Zhao et al., "Advancing calving time prediction using tail acceleration data : Comparative evaluation of transformer , hybrid CNN-transformer and LSTM-transformer models," Smart Agricultural Technology, vol. 12, no. October, 2025.

[8] T. Chen and C. Guestrin, "XGBoost : A Scalable Tree Boosting System," in KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[9] A. R. Z, S. Bakhtiari, C. Aldrich, V. M. Calo, and M. Iannuzzi, "XGBoost model for the quantitative assessment of stress corrosion cracking," npj Materials Degradation, 2024, doi: 10.1038/s41529-024-00538-7.

[10] G. Grekousis, "Geographical - XGBoost : a new ensemble model for spatially local regression based on gradient - boosted trees," Journal of Geographical Systems, vol. 27, no. 2, pp. 169–195, 2025, doi: 10.1007/s10109-025-00465-4.

[11] L. Sluijterman, F. Kreuwel, E. Cator, and T. Heskes, "Composite quantile regression with XGBoost using the novel arctan pinball loss," International Journal of Machine Learning and Cybernetics, vol. 16, no. 10, pp. 7575–7589, 2025, doi: 10.1007/s13042-025-02671-4.

[12] K. Lin, "Model Research on Forecast of Second-Hand House Price in Shenyang Based on XGBoost Algorithm," in 2024 Asia-Pacific Conference on Software Engineering, Social Network Analysis and Intelligent Computing (SSAIC), Dehli, India, 2024. doi: https://doi.org/10.1109/SSAIC61213.2024.00092.

[13] A. Danandeh, M. Amir, A. G. Elman, G. Mir, J. Sadegh, and C. Q. Ke, "S-Transformer : a new deep learning model enhanced by sequential transformer encoders for drought forecasting," Earth Science Informatics, 2025, doi: 10.1007/s12145-025-01845-6.

[14] O. F. Hassan, A. F. Ibrahim, A. Gomaa, M. A. Makhlouf, and B. Hafiz, "Real-time driver drowsiness detection using transformer architectures : a novel deep learning approach," pp. 1–33, 2025.

[15] Y. Zamanidoost;, T. Ould-Bachir;, and S. Martel, "OMS-CNN: Optimized Multi-Scale CNN for Lung Nodule Detection Based on Faster R-CNN," in IEEE Journal of Biomedical and Health Informatics, 2025. doi: https://doi.org/10.1109/JBHI.2024.3507360.

[16] M. Telmem, N. Laaidi, Y. Ghanou, S. Hamiane, and H. Satori, "Comparative study of CNN , LSTM and hybrid CNN - LSTM model in amazigh speech recognition using spectrogram feature extraction and different gender and age dataset," International Journal of Speech Technology, vol. 27, no. 4, pp. 1121–1133, 2024, doi: 10.1007/s10772-024-10154-0.

Multi-Task IndoBERT for Depression Detection: Intent, Emotion, and PHQ-9 Classification in Indonesian Chatbot

1st Rabiatun Nisa *, 2nd Bimo Sunarfri Hantono †, 3rd Indriana Hidayah ‡, 4th Fadhilla Najmi Qinthara §
* *dept. of Electrical Engineering and Information Technology, Universitas Gadjah Mada, Yogyakarta, Indonesia*
† *dept. of Electrical Engineering and Information Technology, Universitas Gadjah Mada, Yogyakarta, Indonesia*
‡ *dept. of Electrical Engineering and Information Technology, Universitas Gadjah Mada, Yogyakarta, Indonesia*
§ *Faculty of Psychology, Universitas Padjajaran, Bandung, Indonesia*
Emails: *rabiatunnisa@mail.ugm.ac.id, bhe@ugm.ac.id, indriana.h@ugm.ac.id, fadhillaqinthara@gmail.com*

***Abstract*—Indonesia faces a substantial mental health treatment gap, with nearly 70% of individuals with depression not receiving adequate care. To support early detection, this study develops a multi-task IndoBERT model for automated depression screening from Indonesian conversational text. The model jointly performs intent classification (F1 74.75%), emotion detection (F1 78.36%), severity assessment (F1 91.05%), and PHQ-9 symptom prediction (F1 73.21%, macro-recall 93.24%). High recall on safety-critical indicators such as 90% recall for suicidal ideation (Q9) and 84% for crisis intent demonstrates strong potential for triage-oriented mental health applications. Trained on 4,004 synthetic and augmented samples, with a subset reviewed by a clinician-trained psychologist, the model shows stable performance across tasks despite limited real-world data. These findings indicate that a unified IndoBERT framework offers a promising early step toward scalable, language-aware depression-screening tools for Indonesian users, although broader expert validation and evaluation on natural conversational data remain essential before clinical deployment.**

Keywords—IndoBERT, multi-task learning, depression screening, PHQ-9, mental health NLP, Indonesian language modeling, emotion detection, severity classification, clinical NLP.

I. INTRODUCTION

Depression affects more than 280 million people worldwide, and in Indonesia, an estimated 70% of individuals with depression do not receive adequate mental health services[1], [2]. Many cases remain undetected or treated only after significant delays due to limited service availability, persistent stigma, and a shortage of trained mental health professionals barriers that are particularly evident among adolescents and university students [3]. In primary care, the Patient Health Questionnaire-9 (PHQ-9) is widely used as a brief, reliable instrument for identifying depressive symptoms and assessing their severity across nine diagnostic items [4].

Alongside these clinical tools, advances in Natural Language Processing have enabled the use of everyday language as a non-invasive signal of emotional distress. Transformer-based models are increasingly employed to detect depression, anxiety, and related conditions from social media posts, clinical notes, and conversational text, including approaches that map language directly to PHQ-9 symptom categories [5], [6], [7], [8]. AI-based mental-health chatbots have also gained attention as accessible and scalable support tools. While systematic reviews show that such systems may help reduce mild to moderate distress when appropriately designed [9]. they remain complementary tools rather than replacements for clinical judgment [10].

In Indonesia, early research has explored NLP-based depression screening and chatbot systems, yet many implementations still rely on rule-based methods or simple scoring, limiting their ability to handle natural, emotionally nuanced conversations [11]. Moreover, there is currently no publicly available Indonesian mental health dialogue corpus and no model capable of jointly learning conversational intent, emotional state, severity level, and PHQ-9 symptom indicators within a single framework. This combination of limited linguistic resources and the absence of multi-task clinical modeling defines the key research gap.

To address this challenge, our study investigates whether a multi-task IndoBERT model can effectively learn four core aspects of mental health text analysis: intent, emotion, severity, and PHQ-9 symptom indicators within a unified architecture. The proposed model achieves an F1 of 0.91 for severity classification and attains 90% recall for suicidal ideation detection (Q9), a safety-critical indicator. Validation by a clinically trained psychologist shows substantial agreement (Cohen's kappa = 0.81). To our knowledge, this is the first multi-task IndoBERT framework developed for PHQ-9–based depression screening in Indonesian conversational text. This study is exploratory in nature, providing an initial multi-task modeling framework rather than a clinically deployable system.

II. RELATED WORK

A. Depression Detection Using NLP and Transformers

Research on depression detection has shifted from traditional machine learning toward transformer-based models that capture linguistic context with greater depth[6]. Studies on large social media datasets show high accuracy [12], though many rely on sentiment-based labels that miss clinical nuance. Later works introduced severity prediction with F1 scores in the mid-0.80s [13], and comparative evaluations of transformer variants reported accuracies in the mid-80s to low-90s [14] [15]. Despite this progress, most approaches remain single-task and binary, leaving clinical multidimensionality unmodeled.

B. PHQ-9 Based Assessment and Clinical Grounding

The PHQ-9 continues to serve as a reliable foundation for depression screening [4]. Automated assessments using traditional machine learning have shown strong performance for specific items such as suicidal ideation [16]. and analyses of PHQ item pairings have identified predictors outperforming the PHQ-2 [17]. Extensions to behavioral data have produced stable symptom-specific scores [16]. Still, these methods rely on structured responses rather than conversational text, and transformer-based PHQ-9 grounding remains largely unexplored outside English.

C. Indonesian NLP for Mental Health

IndoBERT provides the strongest monolingual baseline for Indonesian NLP, achieving state-of-the-art results across multiple benchmark tasks [18]. Because of its proven capability in modeling Indonesian morphology, syntax, and colloquial usage, this research adopts IndoBERT as the shared encoder in our multi-task mental-health classification framework. Early studies on Indonesian mental-health text primarily relied on keyword-based approaches for depression detection on Twitter [19], or LSTM-based sentiment models [20], while later work applied IndoBERT to healthcare sentiment analysis [21], and general emotion classification benchmarks [22],[23]. However, none of these efforts incorporate clinically grounded tasks, multi-label PHQ-9 symptom modeling, or expert-validated annotations. This gap motivates the development of a unified multi-task IndoBERT framework tailored for mental-health screening.

D. Multi-Task Learning for Mental Health

Multi-task learning has shown potential for integrating related psychological signals. Shared encoders have been used to jointly predict sentiment and mental health status [24]. and transformer-based models have managed multiple outputs such as emotion and sentiment simultaneously [25]. Yet these frameworks typically cover only two or three tasks and remain disconnected from clinically validated dimensions such as PHQ-9 symptom detection and Indonesian conversational context.

E. Treatment Gap and Mental Health Access

Globally, more than seventy-five percent of individuals with mental health conditions in low- and middle-income countries do not receive treatment [26] . Indonesia reflects this pattern, with only 9.3 percent accessing care, constrained by economic barriers and limited insurance coverage [27], Stigma remains pervasive [28]. and the shortage of mental health professionals leaves nearly seventy percent of those in need without support [29]. AI-based chatbots offer a potential bridge [19]. Though their usefulness depends on clinically aligned language understanding, which current Indonesian systems have yet to achieve.

F. Positioning of the Present Study

This study offers a more clinically grounded contribution by proposing an IndoBERT-based model that jointly learns conversational intent, emotional expression, severity, and PHQ-9 symptom indicators. Its alignment with the PHQ-9 framework allows everyday expressions of distress to be interpreted through clinically meaningful categories. Validation by a psychologist with clinical training yields strong agreement ($\kappa = 0.81$), adding needed credibility. The model achieves an F1 of 0.91 for severity and high recall for Q9 suicidal ideation, surpassing past baselines. A single shared encoder maintains efficiency while supporting robust performance across tasks, suggesting a promising path for scalable and clinically aligned Indonesian mental health screening. Table 1 summarizes comparisons with related studies. Table 1 presents a detailed comparison of our work with key related studies.

TABLE I. SUMMARY OF REFERENCE PAPER

Paper	Method	Lang	Tasks	Perform	Clinic
[14], 2023	RoBERTa	EN	Binary	98%	No
[13], 2025	Trans	EN	Severity	F1~0.80	Partial
[16], 2021	RF	KO	Q9	AUC=0.84	Yes
[19], 2024	IndoBERT	ID	Binary	-	No
[23], 2023	IndoBERT	ID	Emotion	F1=0.79	No
[22], 2024	IndoBERT	ID	E + S	0.79/0.92	No
[7], 2025	BERT	EN	S + S	87.5%	Partial
Proposed, 2025	IndoBERT	ID	4 tasks	F1=0.91	κ=0.81

III. METHODOLOGY

A. Overview

This study develops a multitask learning model based on IndoBERT to understand several dimensions of Indonesian mental health conversations simultaneously. The model is designed to perform four classification tasks in a single flow: identifying conversational intent, detecting emotional states, assessing severity, and predicting the presence of the nine PHQ-9 symptom indicators using a multi-label approach.The methodology unfolds through several stages: construction of the dataset, text preprocessing, model

architecture design, multi-task training, and adaptive threshold tuning to refine PHQ-9 predictions.

B. Dataset

1) Dataset Sources

All data were synthetically generated using prompt engineering and NLP-based augmentation, without using public forums, real conversations, or personal health information. The dataset consists of three main components: (1) intent labels (complaint, crisis, greeting, farewell, question, neutral_statement, phq9_response), (2) six emotion categories with three severity levels, and (3) nine PHQ-9 symptom indicators in a multi-label format. Linguistic variation was introduced through paraphrasing, synonym substitution, and light adjustments in formality. Each sentence was manually curated to ensure natural language quality, correct labeling, and alignment with PHQ-9 definitions. After cleaning, the final dataset contained 4,004 unique samples with automatically verified label distributions.

2) Data Preprocessing

Preprocessing included removing unnecessary symbols, normalizing repeated letters and mild slang, standardizing labels, converting text to lowercase, and filtering out sentences shorter than five words. Intent, emotion, and severity labels were encoded using a LabelEncoder, while PHQ-9 indicators were converted into a nine-dimensional binary vector for multi-label prediction. These steps ensured that all inputs followed a consistent format before being processed by the model.

C. Model Architecture

The proposed model adopts IndoBERT-base uncased [18] as the shared encoder. IndoBERT consists of approximately 125 million parameters and is pretrained on large-scale Indonesian corpora, making it well suited for capturing Indonesian morphological and semantic patterns. The model uses the 768-dimensional [CLS] embedding as the shared representation that feeds into all downstream prediction tasks in the multi-task learning framework.

On top of the shared encoder, we add four task-specific classification heads, each following an identical architectural design. First, the 768-dimensional encoder output is passed through a fully connected layer that reduces the representation to 256 dimensions, followed by a ReLU activation and a dropout layer (p = 0.3) to prevent overfitting. A final linear layer then maps the hidden representation to the appropriate output size for each task. The intent head predicts seven conversational intent classes, the emotion head predicts six emotional states, the severity head outputs three intensity levels (LOW, MEDIUM, HIGH), and the PHQ-9 head produces nine sigmoid-activated scores corresponding to the nine PHQ-9 depressive symptom indicators.

Reserve for IEEE Copyright

This architecture enables the shared encoder to learn general linguistic features while the task-specific heads learn specialized decision boundaries for each clinical or conversational task. The overall model structure is illustrated in Figure 1.

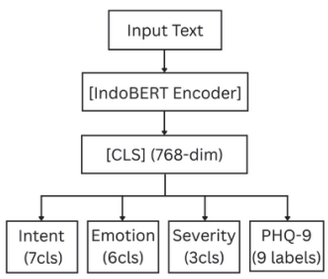


Figure I. Architecture

D. Loss Function

To help the model learn these four tasks together, a combined loss is used:

$$\mathcal{L}=1.0\cdot\mathcal{L}_intent1.0\cdot\mathcal{L}_emotion+1.5\cdot\mathcal{L}_severity+2.0\cdot\mathcal{L}_PHQ9 \quad (1)$$

where:

$\mathcal{L}_intent, \mathcal{L}_emotion, \mathcal{L}_severity$: CrossEntropyLoss

$\mathcal{L}_PHQ9$: BCEWithLogitsLoss

The higher weights for severity and PHQ 9 are chosen intentionally. They help address class imbalance and give priority to tasks that carry deeper clinical importance. These weights are selected through a grid search on the validation set, ensuring that each task receives the attention it needs.

E. Adaptive Threshold Tuning

The PHQ-9 outputs are sigmoid probabilities. A fixed threshold of 0.5 often falls short when indicators have different frequencies. For this reason, each indicator (Q1-Q9) has its own threshold. A grid search tests thresholds from 0.1 to 0.9. For each indicator, the threshold that yields the best F1 score on the validation set is chosen. This threshold, written as θ^*i is then used during inference. This careful tuning allows the model to be more sensitive to rare symptoms such as suicidal ideation in Q9 while still maintaining stability for common indicators like depressed mood in Q2. The result is an improvement in macro F1 and overall sensitivity.

F. Training Configuration

Training follows a compact and stable setup: AdamW optimizer with learning rate 2×10^{-5} and weight decay 0.01, batch size 16, and a maximum sequence length of 128 tokens. The model is trained for 10 epochs with 10 percent warmup, linear decay scheduling, early stopping after 3 stagnant epochs on severity F1, gradient clipping at 1.0, and a fixed random seed of 42. Each setting supports a balance between steady learning and model robustness.

G. Evaluation Metrics

Evaluation is conducted with care across all tasks. For the multi-class tasks of intent, emotion, and severity, the primary measure is macro F1, supported by accuracy and per-class precision, recall, and F1. For the multi-label PHQ-9 task, the evaluation includes macro-F1, micro-F1, and individual-F1 scores for each of the nine indicators. A threshold-adjusted evaluation is used so that the chosen values θ^*i of are reflected in the results. All metrics are computed on the test set. Statistical significance is assessed through bootstrap sampling with one thousand iterations and a threshold of $p < 0.05$.

IV. RESULT

This section presents the performance of the multi-task IndoBERT model across its four core objectives: intent classification, emotion classification, severity prediction, and PHQ 9 multi label symptom detection.

A. Overall Multi Task Performance

Table 2 summarizes the overall results for each task. The model demonstrates strong performance on severity prediction and maintains stable performance on both intent and emotion classification.

TABLE II. OVERALL MULTI-TASK EVALUATION RESULTS

Task	Accuracy	F1 Macro	F1 Weighted
Intent	0.8735	0.7475	0.8628
Emotion	0.8552	0.7836	0.8491
Severity	0.9118	0.9105	0.9120
PHQ 9 (flattened)	0.8813	0.7321	-

B. Intent Classification

Table 3 presents the per class performance for the intent classification task. The model performs well on majority classes such as *complaint* and *crisis*, while performance decreases on minority classes, particularly *greeting*, which reflects the expected effect of data imbalance.

TABLE III. PER-LABEL PERFORMANCE FOR INTENT CLASSIFICATION

Label	Precision	Recall	F1	Support
complaint	0.9141	0.9526	0.9330	380
crisis	0.8889	0.8421	0.8649	76
farewell	0.6071	0.9714	0.7473	35
greeting	1.0000	0.2069	0.3429	29
neutral statement	1.0000	0.5556	0.7143	27
phq9 response	0.7812	0.7143	0.7463	35
question	0.7917	1.0000	0.8837	19

C. Emotion Classification

Table 4 presents the per class performance for the emotion classification task. The model performs consistently well on majority categories such as *neutral*, *sadness*, and *hopelessness*. The *fear* label shows notably lower performance, likely due to its limited sample size and its semantic proximity to *anxiety*, which makes the two emotions harder for the model to distinguish.

TABLE IV. PER-LABEL PERFORMANCE FOR EMOTION CLASSIFICATION

Emotion	Precision	Recall	F1	Support
anger	0.9583	0.6571	0.7797	35
anxiety	0.7985	0.9469	0.8664	113
fear	0.8750	0.3043	0.4516	23
hopelessness	0.8729	0.7984	0.8340	129
neutral	0.9198	0.9313	0.9255	160
sadness	0.8065	0.8865	0.8446	141

D. Severity Classification

Severity prediction achieves the highest performance among all tasks, with a macro F1 score of 0.9105. Table 5 shows the per class results, indicating strong performance across all severity levels.

TABLE V. PER LABEL PERFORMANCE FOR SEVERITY

Severity	Precision	Recall	F1	Support
HIGH	0.8452	0.8792	0.8618	149
LOW	0.9506	0.9625	0.9565	160
MEDIUM	0.9261	0.9007	0.9132	292

E. PHQ-9 Multi-Label Symptom Detection

Table 6 presents the performance for each of the nine PHQ 9 indicators. The model achieves strong F1 scores for Q1, Q5, and the clinically critical Q9. In contrast, performance is lower on Q3, Q6, and Q7, which reflects the challenge of detecting symptoms that appear less frequently and share overlapping linguistic patterns.

TABLE VI. PER-INDICATOR PHQ-9 PERFORMANCE

Q	Precision	Recall	F1	Support
Q1	0.6098	0.8929	0.7246	28
Q2	0.4204	0.8407	0.5605	113
Q3	0.2596	1.0000	0.4122	27
Q4	0.3165	0.8333	0.4587	30
Q5	0.5652	1.0000	0.7222	26
Q6	0.2256	0.9250	0.3627	40
Q7	0.3051	1.0000	0.4675	36
Q8	0.2925	1.0000	0.4526	31
Q9	0.7241	0.9000	0.8025	70

V. DISCUSSION

The findings of this study show that the multitask IndoBERT model can learn four aspects of mental-health text analysis simultaneously: intent, emotion, severity, and PHQ-9 symptoms with stable performance across most tasks. This suggests that a shared encoder helps the model capture subtle linguistic patterns in Indonesian expressions of emotion and distress.

In intent classification, the model achieves 87.35% accuracy and a macro-F1 of 74.75%, while emotion classification reaches 85.52% accuracy and a macro-F1 of 78.36%. These results are encouraging given the variability of informal Indonesian language and the small sample sizes of several minority classes (e.g., fear:

23 samples, greeting: 29 samples, neutral_statement: 27 samples). Misclassification patterns such as fear frequently being interpreted as anxiety (69.6%) reflect semantic overlap commonly observed in emotion modeling literature [22], [30].

Severity classification is the strongest task, yielding a macro-F1 of 91.05% and accuracy of 91.18%. This indicates that IndoBERT is effective at recognizing Indonesian intensity markers (sangat, amat, lumayan, sedikit) and explicit emotional cues, enabling clear separation of LOW, MEDIUM, and HIGH severity levels an important property for triage-oriented applications.

For PHQ-9 multi-label prediction, the model achieves a macro-F1 of 73.21% and a macro-recall of 93.24%. The high recall shows that the model prioritizes capturing possible symptoms, which is appropriate for early-stage screening. Q9 (suicidal ideation) performs particularly well (F1 = 80.25%, recall = 90%), while indicators with broader linguistic usage such as Q3 (sleep problems) show lower discriminability (F1 = 41.22%) due to overlap with everyday conversation.

When compared to IndoBERT's original single-task performance on IndoLEM benchmarks [18], our results indicate that the pretrained representations generalize effectively to clinically aligned tasks. The multi-task variation used in this study improves preliminary single-task IndoBERT baselines by approximately 2–3 macro-F1 points, suggesting that shared representations are helpful for low-resource psychological text modelling.

Taken together, these findings suggest that multi-task learning is a promising direction for Indonesian mental-health screening. Although the dataset is primarily synthetic (4,004 samples, including 2,578 augmented instances), the model's stable performance indicates that it has begun to capture meaningful linguistic markers of emotional distress and depressive symptoms. These results provide an early but promising foundation for scalable, context-aware mental-health support tools in Indonesian. Further validation on expert-annotated and real-world conversational data remains necessary.

VI. CONCLUSSION

This study introduces a multi-task IndoBERT model designed to jointly predict conversational intent, emotional state, severity level, and PHQ-9 symptom indicators from Indonesian mental-health text. The model demonstrates stable performance across tasks, achieving a macro-F1 of 91.05% for severity and a macro-recall of 93.24% for PHQ-9 symptoms, indicating strong sensitivity to clinically relevant signals. Intent and emotion classification also show consistent performance, although smaller classes exhibit the expected challenges of data imbalance.

These findings suggest that a shared IndoBERT encoder can generalize effectively to domain-specific mental-health tasks, even when trained primarily on synthetic data. The model captures linguistic cues associated with distress, intensity, and depressive symptoms, providing preliminary evidence that multi-task learning is a viable direction for Indonesian mental-health screening.

However, the study has several limitations. The synthetic nature of the dataset and the reliance on validation from a single clinically trained annotator limit the generalizability of the findings. Broader expert review and real-world conversational data will be necessary to more accurately assess clinical reliability.

Overall, this work represents an early but meaningful step toward scalable, language-based mental-health screening tools for Indonesian speakers. Future work will focus on expanding expert-annotated datasets, improving class balance, and evaluating model behaviour in real user interactions.

VII. LIMITATION AND FUTURE WORK

The findings of this study show encouraging potential, but several limitations remain. First, the dataset relies heavily on synthetic generation, with only one clinically trained annotator providing initial validation; broader expert review and more naturalistic data are still needed. Second, several minority classes such as *fear*, *greeting*, and *neutral statement* contain very few samples, making them difficult for the model to learn and contributing to imbalance-related errors. Third, the model has not yet been evaluated in real-world settings, so its behaviour outside curated research data remains uncertain. These limitations will be addressed in the thesis through expanded expert validation, improved data balancing, and pilot testing with real users.

ACKNOWLEDGMENT

The authors gratefully acknowledge the assistance of an AI language model in refining the grammar, clarity, and cohesion of several written sections in this manuscript. All conceptual contributions, analyses, and interpretations remain the responsibility of the authors.

REFERENCES

[1] "Identifying Depressive Symptoms from Tweets: Figurative Language Enabled Multitask Learning Framework - ACL Anthology." Accessed: Nov. 16, 2025. [Online]. Available: https://aclanthology.org/2020.coling-main.61

[2] "News: Depression affects about 280 million... (The Washington Post) - Behind the headlines - NLM." Accessed: Nov. 16, 2025. [Online]. Available: https://www.ncbi.nlm.nih.gov/search/research-news/15672/

[3] K. Arulsamy *et al.*, "The economic burden of anxiety and depression in Indonesia: evidence from a cross-sectional web panel survey,"

Front. Public Health, vol. 13, p. 1667726, Oct. 2025, doi: 10.3389/fpubh.2025.1667726.

[4] K. Kroenke, R. L. Spitzer, and J. B. W. Williams, "The PHQ-9," *J. Gen. Intern. Med.*, vol. 16, no. 9, pp. 606–613, Sept. 2001, doi: 10.1046/j.1525-1497.2001.016009606.x.

[5] T. Zhang, K. Yang, H. Alhuzali, B. Liu, and S. Ananiadou, "PHQ-aware depressive symptoms identification with similarity contrastive learning on social media," *Inf. Process. Manag.*, vol. 60, no. 5, p. 103417, Sept. 2023, doi: 10.1016/j.ipm.2023.103417.

[6] T. Zhang, A. M. Schoene, S. Ji, and S. Ananiadou, "Natural language processing applied to mental illness detection: a narrative review," *Npj Digit. Med.*, vol. 5, no. 1, p. 46, Apr. 2022, doi: 10.1038/s41746-022-00589-7.

[7] S. Yadav, J. Chauhan, J. P. Sain, K. Thirunarayan, A. Sheth, and J. Schumm, "Identifying Depressive Symptoms from Tweets: Figurative Language Enabled Multitask Learning Framework," in *Proceedings of the 28th International Conference on Computational Linguistics*, D. Scott, N. Bel, and C. Zong, Eds., Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 696–709. doi: 10.18653/v1/2020.coling-main.61.

[8] N. H. Kim, J. M. Kim, D. M. Park, S. R. Ji, and J. W. Kim, "Analysis of depression in social media texts through the Patient Health Questionnaire-9 and natural language processing," *Digit. Health*, vol. 8, p. 20552076221114204, July 2022, doi: 10.1177/20552076221114204.

[9] H. Li, R. Zhang, Y.-C. Lee, R. E. Kraut, and D. C. Mohr, "Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being," *Npj Digit. Med.*, vol. 6, no. 1, p. 236, Dec. 2023, doi: 10.1038/s41746-023-00979-5.

[10] B. Omarov, S. Narynov, and Z. Zhumanov, "Artificial Intelligence-Enabled Chatbots in Mental Health: A Systematic Review," *Comput. Mater. Contin.*, vol. 74, no. 3, pp. 5105–5122, 2022, doi: 10.32604/cmc.2023.034655.

[11] F. Nuraini, M. N. D. Miharja, and A. H. Anshor, "Mental Health Chatbot for Detecting Depression Symptoms Using Natural Language Processing and DASS-21," *J. Teknol.*, vol. 17, no. 2, pp. 133–142, July 2025, doi: 10.24853/jurtek.17.2.133-142.

[12] "Depression detection in social media posts using transformer-based models and auxiliary features | Social Network Analysis and Mining." Accessed: Nov. 16, 2025. [Online]. Available: https://link.springer.com/article/10.1007/s13278-024-01360-4

[13] A. Qasim, G. Mehak, N. Hussain, A. Gelbukh, and G. Sidorov, "Detection of Depression Severity in Social Media Text Using Transformer-Based Models," *Information*, vol. 16, no. 2, p. 114, Feb. 2025, doi: 10.3390/info16020114.

[14] B. G. Bokolo and Q. Liu, "Advanced Comparative Analysis of Machine Learning and Transformer Models for Depression and Suicide Detection in Social Media Texts," *Electronics*, vol. 13, no. 20, p. 3980, Jan. 2024, doi: 10.3390/electronics13203980.

[15] İ. Baydili, B. Tasci, and G. Tasci, "Deep Learning-Based Detection of Depression and Suicidal Tendencies in Social Media Data with Feature Selection," *Behav. Sci.*, vol. 15, no. 3, p. 352, Mar. 2025, doi: 10.3390/bs15030352.

[16] S. Kim, H.-K. Lee, and K. Lee, "Which PHQ-9 Items Can Effectively Screen for Suicide? Machine Learning Approaches," *Int. J. Environ. Res. Public. Health*, vol. 18, no. 7, p. 3339, Jan. 2021, doi: 10.3390/ijerph18073339.

[17] D. Glavin *et al.*, "Patient Health Questionnaire-9 Item Pairing Predictiveness for Prescreening Depressive Symptomatology: Machine Learning Analysis," *JMIR Ment. Health*, vol. 10, p. e48444, Oct. 2023, doi: 10.2196/48444.

[18] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," in *Proceedings of the 28th International Conference on Computational Linguistics*, D. Scott, N. Bel, and C. Zong, Eds., Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 757–770. doi: 10.18653/v1/2020.coling-main.66.

[19] G. F. Situmorang and R. Purba, "Deteksi Potensi Depresi dari Unggahan Media Sosial X Menggunakan IndoBERT," *Build. Inform. Technol. Sci. BITS*, vol. 6, no. 2, pp. 649–661, Sept. 2024, doi: 10.47065/bits.v6i2.5496.

[20] I. D. Nugraha and Y. Azhar, "Deteksi Depresi Pengguna Twitter Indonesia Menggunakan LSTM-RNN," *J. Nas. Pendidik. Tek. Inform. JANAPATI*, vol. 11, no. 3, pp. 320–329, Dec. 2022, doi: 10.23887/janapati.v11i3.50674.

[21] H. Imaduddin, F. Y. A'la, and Y. S. Nugroho, "Sentiment Analysis in Indonesian Healthcare Applications using IndoBERT Approach," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 8, 2023, doi: 10.14569/IJACSA.2023.0140813.

[22] H. Ahmadian, T. F. Abidin, H. Riza, and K. Muchtar, "Hybrid Models for Emotion Classification and Sentiment Analysis in Indonesian Language," *Appl. Comput. Intell. Soft Comput.*, vol. 2024, no. 1, p. 2826773, 2024, doi: 10.1155/2024/2826773.

[23] C. F. Shaw, "A Transformer Based Architecture for Indonesian Sentiment Analysis - Exploring IndoBERT Variations, Training Size, and Self-Supervised Model Training".

[24] "Multi task opinion enhanced hybrid BERT model for mental health analysis | Scientific Reports." Accessed: Nov. 16, 2025. [Online]. Available: https://www.nature.com/articles/s41598-025-86124-6

[25] L. J. R&D, "Multiclass and Multilabel Text Classification in One BERT Model," Medium. Accessed: Nov. 16, 2025. [Online]. Available: https://lajavaness.medium.com/multiclass-and-multilabel-text-classification-in-one-bert-model-95c54aab59dc

[26] "Prevention." Accessed: Nov. 16, 2025. [Online]. Available: https://www.who.int/teams/mental-health-and-substance-use/treatment-care/mental-health-gap-action-programme

[27] M. F. Rizal, D. W. Johnston, N. Black, and R. Sweeney, "Mental Healthcare Access and the Treatment Gap in Indonesia".

[28] N. Hartini, N. A. Fardana, A. D. Ariana, and N. D. Wardana, "Stigma toward people with mental health problems in Indonesia," *Psychol. Res. Behav. Manag.*, vol. 11, pp. 535–541, Oct. 2018, doi: 10.2147/PRBM.S175251.

[29] "Mental Health in Asia: The Numbers." Accessed: Nov. 17, 2025. [Online]. Available: https://www.ourbetterworld.org/series/mental-health/support-toolkit/mental-health-asia-numbers

[30] C. Shaw, "A Transformer Based Architecture for Indonesian Sentiment Analysis - Exploring IndoBERT Variations, Training Size, and Self-Supervised Model Training," *Theses Diss.*, Mar. 2023, [Online]. Available: https://scholar.afit.edu/etd/7013

Real-Time Protection Scheme for Embedded Systems Using a *Slack Manager*

1st Kunwoo Park*, 2nd Ingeol Chun†, 3rd Daehee Seo‡

*Dept. of Computer Science, Sangmyung University, Seoul, Republic of Korea
†Electronics and Telecommunications Research Institute (ETRI), Daejeon, Republic of Korea
‡Dept. of Intelligent Data Convergence, Sangmyung University, Seoul, Republic of Korea

Emails: 202532012@sangmyung.kr, igchun@etri.re.kr, daehseo@smu.ac.kr

***Abstract*— Recently, Edge–cloud embedded systems must satisfy both real-time guarantees and security requirements. Violating timing constraints can delay security operations, causing missed detections or deadline overruns in critical domains such as autonomous driving and aircraft control. Previously, studies rely on a local Slack model in which each task can use only the Slack it generates, preventing unused Slack from being redistributed to tasks that require more runtime. To overcome this limitation, this work introduces a *Slack Manager*–based real-time protection scheme that globally aggregates Slack into a unified *Slack Pool* and redistributes it according to threat level, security contribution, and deadline tightness. This global Slack utilization enables more flexible activation of security mechanisms and improves resilience against both known and unknown attacks compared to traditional local Slack models.**

I. INTRODUCTION

Edge-cloud based embedded systems such as autonomous vehicles or aircraft controllers—operate under strict real-time constraints, where even small processing delays can cause system failures or safety risks. If timing guarantees are violated, attackers may exploit delayed responses or timing fluctuations to disrupt control flows, potentially leading to physical damage or threats to human life.

Thus, embedded systems must satisfy both real-time and security requirements simultaneously. To achieve this, prior studies applied Mixed-Integer Linear Programming (MILP) to assign security mechanisms using the execution Slack that remains after task completion. However, although each task may retain unused Slack, it cannot be shared across tasks, resulting in inefficient system-wide utilization.

To overcome this limitation, this paper introduces a real-time protection scheme employing a *Slack Manager* that globally collects and redistributes Slack. By allocating Slack based on threat level, security contribution, and deadline tightness, the proposed scheme enables dynamic activation of security mechanisms during runtime.

The remainder of this paper is structured as follows. Section II reviews existing research on MILP-based real-time protection and identifies the limitations of local-only Slack utilization. Section III presents the proposed *Slack Manager* Scheme, describing the global Slack collection, integration, and redistribution mechanisms. Section IV provides an analysis of the proposed scheme, including simulation setup, comparison with baseline schemes, and evaluation results. Section V concludes the paper by summarizing key findings and outlining future research directions [1].

II. RELATED WORK

This section reviews prior protection schemes proposed for embedded system security, focusing on MILP-based static protection scheme and MILP-based dynamic protection scheme that utilize Slack.

A. *MILP-Based Static Protection Schemes*

Static MILP-based methods assign security mechanisms to Basic Blocks by analyzing each task's CFG and incorporating their overhead into WCET [2]. Although this guarantees stable real-time behavior with a fixed configuration, it cannot adapt to runtime workload changes and cannot

utilize Slack generated during execution, limiting its ability to respond to new attacks and varying computational demands [3].

B. MILP-Based Dynamic Protection Using Slack

Dynamic MILP-based schemes improve flexibility by allowing each task to use its own unused Slack for additional security operations. However, Slack remains isolated per task, preventing surplus Slack in one task from assisting others, which leads to inefficient system-wide utilization. To overcome this limitation, this study introduces a *Slack Manager* that aggregates Slack across tasks into a global *Slack Pool* and redistributes it to tasks requiring additional runtime [4].

To address this limitation, this study proposes a *Slack Manager*–based real-time protection scheme that aggregates Slack generated by multiple tasks into a global *Slack Pool* and redistributes it to tasks that require additional runtime.

III. PROPOSED *Slack Manager*–BASED REAL-TIME PROTECTION SCHEME

Previous studies treated Slack as a task-local resource, limiting system-wide utilization and reducing responsiveness to high-priority security needs. To address this, we propose a real-time protection scheme using a *Slack Manager* that centrally collects and redistributes Slack while preserving existing MILP and dynamic Slack schemes. The scheme integrates Slack from all tasks into a unified structure and reallocates it according to threat level and security contribution, improving security effectiveness without violating real-time constraints. TABLE I summarizes the symbols used in the scheme.

TABLE I. Symbols used in the system model

Symbol	Description
D_i	Deadline of Task i
RT_i	Response time of Task i
y_i	Protection unit selection variable
z_i	Contribution reflection variable
$Slack_i$	Slack of Task i (Deadline – Response time)
$Slack_{min}$	Minimum Slack required for activation
$Slack_{pool}$	Total *Slack Pool* maintained by *Slack Manager*
R_i	Protection activation variable
A_i	Redistributed Slack
SM_i	Indicator of actual protection activation
U_i	Security contribution (utilization) score
U_{th}	Contribution threshold
w_i	Known-Attack defense contribution
q_i	Unknown-Attack detection contribution
o_i	Protection cost (overhead)
α	Weight for Known-Attack defense
θ	Weight for Unknown-Attack detection
η	Weight for overhead cost
β	Weight for Slack stability
γ	Weight for Slack redistribution activation
Δ_i	Laxity (Deadline Slack margin)

A. Contribution-Based Protection Activation Condition

In the proposed scheme, the *Slack Manager* evaluates Slack not merely as a residual timing resource but as a *security contribution resource*. The decision to utilize Slack is determined by the contribution scoring function presented in Equation (1), which quantifies the security value of activating protection for each task.

$$U_i = \alpha w_i + \theta q_i - \eta o_i + \beta\, Slack_i \qquad (1)$$

The utility score consists of four weighting terms, Each term reflects a security contribution or time stability factor:

- **Known-Attack Defense Term** (αw_i): Represents the task's contribution to mitigating previously identified or well-classified attack behaviors. The weight α controls the importance assigned to defending Known Attacks.
- **Unknown-Attack Detection Term** (θq_i): Quantifies the task's ability to detect zero-day or mutated attacks through generalized integrity checks or anomaly signals. The weight θ reflects how strongly the system prioritizes Unknown Attack detection.
- **Overhead Penalty Term** (ηo_i): Penalizes the execution-time overhead and resource cost incurred when a protection mechanism is applied. The weight η regulates the trade-off between security enhancement and real-time execution cost.
- **Slack Stability Reward Term** ($\beta\, Slack_i$): Rewards tasks that maintain sufficient timing slack, thereby contributing to overall deadline stability. The weight β determines how strongly the system favors temporal robustness during scheduling.

The *Slack Manager* evaluates Slack not only by its remaining time but also by how much this Slack can contribute to additional security

functions, such as blocking or detecting attacks. In other words, Slack is treated not merely as unused time, but as a quantifiable security resource that reflects the expected reduction in risk achieved by performing extra security operations (e.g., integrity checks, control-flow validation).

Based on this value—which represents how much each Task can contribute to the overall system security when additional security mechanisms are applied—the *Slack Manager* redistributes Slack accordingly. The activation of protection for each Task must satisfy the following condition, shown in Equation (2).

$$R_i = \begin{cases} 1, & \text{if } U_i \geq U_{\text{th}} \\ & \text{and } D_i - RT_i > Slack_{\min}, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

- $R_i = 1$: This indicates that the Task has sufficient Slack and its security utility exceeds the predefined threshold. In this case, the security mechanism is activated.
- $R_i = 0$: This indicates that the Task lacks sufficient Slack or its utility value is below the threshold, resulting in the deactivation of the security mechanism.

In summary, protection is activated only when the Slack security contribution of a Task exceeds the threshold U_{th} and its remaining deadline margin is greater than the minimum required Slack $Slack_{\min}$. This ensures that Slack is not treated as a mere leftover time but instead functions as a trigger for enabling real-time security scheme.

B. Optimization with Slack Manager Integration

The *Slack Manager*–based scheme extends the original Slack-based objective function by adding a redistribution efficiency term, resulting in the expanded formulation shown in Equation (3).

$$\begin{aligned} \max \quad & \sum_i w_i z_i + \theta \sum_i q_i z_i - \eta \sum_i o_i y_i \\ & + \beta \sum_i Slack_i + \gamma \sum_i R_i \end{aligned} \tag{3}$$

The extended Mixed-Integer Linear Programming (MILP) objective function consists of five key components designed to balance security, real-time guarantees, and resource efficiency [5].

1) **First term** ($\sum_i w_i z_i$): First term quantifies the system's defense capability against Known Attacks by reflecting the defensive value of each selected protection unit. Here, w_i represents the defense score of protection unit i, and z_i is a binary variable indicating whether the unit is selected. Selecting protection units with higher w_i increases this term, thereby strengthening the baseline protection through mechanisms such as control-flow integrity checks, execution-integrity validation, or input filtering [6].
2) **Second term** ($\theta \sum_i q_i z_i$): Second term represents the system's ability to detect Unknown Attacks. The value q_i indicates each protection unit's contribution to anomaly or integrity checking, while θ reflects how strongly the system prioritizes zero-day threats. Units with higher q_i increase this term, but stronger anomaly detection may also introduce false positives and additional overhead, requiring careful balance.
3) **Third term** ($\eta \sum_i o_i y_i$): Third term represents the cost-minimization component that penalizes execution-time and resource overhead. Here, o_i denotes the standardized cost of activating protection unit i (e.g., runtime, memory, or energy overhead), and y_i indicates whether the unit is applied. By discouraging high-cost selections, this term helps maintain real-time constraints while still allowing essential security mechanisms.
4) **Fourth term** ($\beta \sum_i Slack_i$): Fourth term promotes maintaining adequate Slack for each task to ensure timing stability and Deadline satisfaction. Slack is defined as the difference between a task's deadline and its response time, which includes both its own execution time and interference from higher-priority tasks. A larger β places greater emphasis on preserving real-time guarantees, allowing tasks to safely perform additional security operations. However, assigning an excessively high weight may reduce overall security efficiency by prioritizing timing margins too strongly. Thus, β must be tuned to balance real-time stability and security performance.

5) **Fifth term** ($\gamma \sum_i R_i$): Fifth term measures how effectively protection mechanisms are activated through global Slack redistribution. The *Slack Manager* allocates collected Slack to tasks with higher security needs, and R_i indicates whether a protection unit is actually enabled for task i. A larger γ increases the impact of redistribution within the objective function, emphasizing the conversion of unused Slack into meaningful security actions that enhance both security and real-time performance.

C. Slack Pool Definition and Redistribution Constraints

Each task reports its remaining Slack upon completion, and the *Slack Manager* aggregates it into a global shared *Slack Pool*. This pool represents the total available Slack for system-wide redistribution, allowing unused Slack to be reassigned to tasks needing more runtime, especially those with higher security or tighter timing constraints. To achieve this, the *Slack Manager* must satisfy the following constraints in Equations (4)–(6).

$$\sum_i A_i \leq \sum_i Slack_{pool} \tag{4}$$

$$A_i \leq Slack_{pool} \tag{5}$$

$$SM_i \leq y_i \tag{6}$$

- $\sum_i A_i \leq \sum_i Slack_{pool}$: Ensures that the total redistributed Slack never exceeds the amount accumulated in the *Slack Pool*, preventing infeasible allocations.
- $A_i \leq Slack_{pool}$: Limits each task from receiving more Slack than it contributed, maintaining fairness and preventing excessive concentration on a single task.
- $SM_i \leq y_i$: Ensures a protection mechanism can activate only if it was selected by the MILP model, preserving consistency between selection and actual activation.

As described above, the MILP optimization built on the *Slack Pool* allows the *Slack Manager* to jointly satisfy security, efficiency, and real-time guarantees [7], [8].

D. Slack-Based Protection Mode Switching

The *Slack Manager* supports three dynamic protection modes depending on the current system Slack state.

Through this mode-switching mechanism, the system can adapt in real time to changing threat conditions while maintaining stable protection even in constrained embedded environments.

E. Slack Integration and Redistribution Scheme

Conventional research computes and consumes Slack locally within each task, resulting in per-task optimization that cannot exploit system-wide Slack availability. In contrast, the proposed scheme treats Slack as a global resource by introducing a centralized Slack Management layer.

The *Slack Manager* collects the Slack reported by each task after completion and stores it in a global *Slack Pool*. It then redistributes Slack by considering task threat level, priority, and security contribution value.

Slack redistribution proceeds through the following steps:

1) The *Slack Manager* gathers the remaining Slack of each task and aggregates them into the *Slack Pool*.
2) The *Slack Manager* evaluates the available Slack using weighted policies defined in the objective function.
3) Based on this evaluation, the *Slack Manager* identifies tasks that lack sufficient Slack.
4) Slack is then reassigned to tasks by considering their threat level, deadline Slack margin, and security contribution value.
5) The reassigned Slack enables additional security mechanisms (e.g., control-flow integrity checks, execution integrity validation) to execute.

Through this process, the *Slack Manager* reduces Slack imbalance among tasks and prevents overload concentration on specific execution segments. Furthermore, the redistribution scheme strictly adheres to the MILP constraints, ensuring that security enhancement is achieved without harming real-time guarantees. The overall workflow of the redistribution scheme is illustrated in Fig 1 [9].

TABLE II. Task Details

Task	Deadline	WCET	Overhead	Base Slack
T1	20 ms	5 ms	4 ms	15 ms
T2	50 ms	12 ms	5 ms	38 ms
T3	100 ms	25 ms	6 ms	75 ms

IV. Analysis of the Proposed Scheme

This section outlines the experimental environment used to validate the proposed *Slack Manager*–based protection scheme. We describe the Python–MILP setup, compare it with the traditional local-Slack model, and analyze the resulting protection configurations and Slack utilization to highlight the benefits of the proposed approach.

A. *Experimental Environment for the Proposed Scheme*

To evaluate the performance of the proposed scheme, we implemented a simulation environment using Python 3.10 and the PuLP MILP solver. The experiment used a set of three periodic real-time tasks (T1, T2, T3), each defined with a deadline, base WCET, overhead, and Slack time. The initial Slack values were: T1 = 15 ms, T2 = 38 ms, and T3 = 75 ms, forming a total system *Slack Pool* of 128 ms.

In the baseline MILP model, Slack could only be consumed within each task (local-Slack model). In contrast, the proposed scheme introduces a *Slack Pool*, enabling global redistribution of Slack across all tasks. Both models were evaluated under identical weight parameters. The detailed task configuration is presented in TABLE II [10].

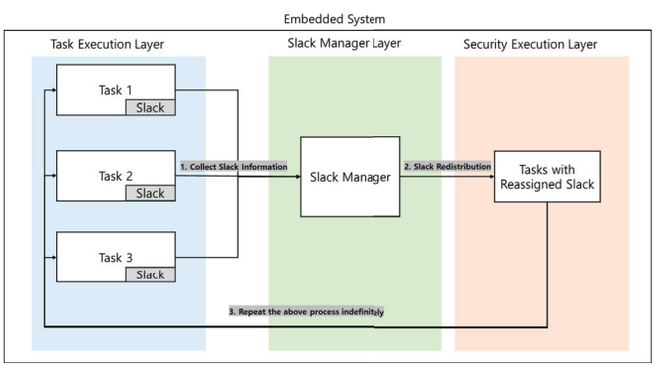


Fig. 1. Workflow of the *Slack Manager* and Slack Redistribution Scheme

B. *Comparison Between the Baseline and Proposed Scheme*

This subsection compares the performance of the traditional MILP-based local Slack scheme and the proposed *Slack Manager*–based global Slack redistribution scheme using tasks T1, T2, T3. The comparison focuses on the following four criteria.

1) **Protection Mechanism Activation**: Both schemes successfully activated all protection mechanisms because each task had enough local Slack. Thus, in terms of basic applicability, both schemes produced identical results.
2) **Slack Utilization Strategy**: Significant differences appeared between the two models in terms of Slack utilization.
   - **Baseline MILP with Local Slack**: Each task uses only its own Slack, consuming only what is needed for overhead. In the experiment, although 128 ms of Slack existed in total, only 15 ms were used and 113 ms (88.3%) were wasted.
   - **Proposed Scheme with Global Slack Redistribution**: All Slack is collected into a global *Slack Pool* and redistributed. Since total Slack exceeded the required overhead, the full 128 ms remained usable as shared Slack.
3) **Protection Activation**: Differences also appear in how protection mechanisms are actually activated.
   - **Baseline MILP with Local Slack**: The baseline selects protection mechanisms but cannot ensure they can run at runtime. As a result, selected mechanisms may fail to activate if Slack is insufficient.
   - **Proposed Scheme**: The proposed scheme separates *selection* from *activation*. Only tasks receiving enough Slack from the pool set $R_i = 1$, guaranteeing actual execution. In the experiment, all tasks secured enough Slack and achieved full activation.
4) **Objective Function Optimization Results**: The objective function results show a clear improvement when using the proposed

TABLE III. Comparison Between Baseline MILP and Proposed *Slack Manager*

Category	Baseline MILP Scheme	Proposed *Slack Manager* Scheme	Improvement Effect
Slack Utilization Structure	Local Slack only	Global *Slack Pool* sharing	Eliminates Slack wastage
Slack Utilization Ratio	11.7%	Up to 100%	+88.3% improvement
Protection Activation	Not guaranteed	Always $R_i = 1$ for all tasks	Actual security execution enabled
Objective Function Value	8.5	17.3	+103% improvement
Global Optimization	Not possible	Fully supported	System-wide optimization achieved

scheme compared to the baseline.

- **Baseline MILP with Local Slack**: The baseline achieved an objective value of 8.5, limited by its inability to use unused Slack.
- **Proposed Scheme with Global Slack Redistribution**: Adding Slack reward and activation terms raised the objective value to 17.3, an improvement of about 103%. This demonstrates that global Slack redistribution enhances security efficiency and supports proactive activation against high-risk threats.

Detailed numerical results comparing the two models are summarized in TABLE III.

V. Conclusion

This paper presented a *Slack Manager*–based real-time protection scheme that overcomes the limitations of traditional dynamic schemes restricted to local Slack. By aggregating Slack into a unified *Slack Pool* and redistributing it based on threat level and security contribution, the proposed scheme achieves more flexible and security-aware scheduling.

The *Slack Manager* strengthens defense against both known and Unknown Attacks while preserving real-time guarantees, effectively converting unused Slack into meaningful security actions. Future work will extend the Slack redistribution strategy to multicore and heterogeneous platforms and incorporate lightweight anomaly-detection techniques to further enhance system security.

Acknowledgment

This work was supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (RS-2025-25453740, Development of a large-scale mixed device control and management platform for edge AI server systems)

References

[1] S. Di Leonardi, F. Aromolo, P. Fara, G. Serra, D. Casini, A. Biondi, and G. Buttazzo, "Maximizing the Security Level of Real-Time Software While Preserving Temporal Constraints," *IEEE Access*, vol. 11, pp. 35591–35607, 2023.

[2] Z. Guo, S. Vaidhun, A. A. Arafat, N. Guan, and K. Yang, "Stealing Static Slack via WCRT and Sporadic P-Servers in Deadline-Driven Scheduling," in *Proc. IEEE Real-Time Systems Symposium (RTSS)*, Taipei, Taiwan, pp. 40–52, 2023.

[3] Y. Wang, A. Li, J. Wang, S. Baruah, and N. Zhang, "Opportunistic Data Flow Integrity for Real-Time Cyber-Physical Systems Using Worst Case Execution Time Reservation," in *USENIX Security Symp.*, Philadelphia, USA, pp. 1–16, 2024.

[4] M. F. Babar and M. Hasan, "DeepTrustR̂T: Confidential Deep Neural Inference Meets Real-Time!" in *Proc. Euromicro Conf. Real-Time Systems (ECRTS)*, Lille, France, 2024.

[5] B. Standaert, P. Axer, *et al.*, "Limited-Preemption EDF Scheduling for Multi-Phase Secure Tasks," *Leibniz Transactions on Embedded Systems (LITES)*, vol. 11, no. 1, pp. 1–25, 2025.

[6] D. Casini, "To MILP or Not to MILP? On AI Techniques for the Design and Optimization of Real-Time Systems," *Real-Time Systems*, vol. 61, no. 2, pp. 294–299, 2025.

[7] D. Senapati, B. Bhagat, and S. Karfa, "SHIELD: Security-Aware Scheduling for Real-Time DAGs on Heterogeneous Systems," *ACM Trans. Cyber-Physical Systems*, 2024.

[8] A. Sain, A. Al-Junaid, E. Smullen, *et al.*, "MAARS: Multi-Rate Attack-Aware Randomized Scheduling for Securing Real-Time Systems," *arXiv preprint arXiv:2408.00341*, 2024.

[9] A. J. Neto and J. M. da Silva, "Provable Execution in Real-Time Embedded Systems," *arXiv preprint arXiv:2505.13842*, 2025.

[10] Y. Zheng, Q. Guo, S. Guo, L. Ren, and Y. Jiang, "A Joint Delay-Energy-Security Aware Framework for Intelligent Task Scheduling in Satellite–Terrestrial Edge Computing Networks," *IEEE Open J. Commun. Soc.*, vol. 5, pp. 2720–2732, 2024.

Hippocampal-Inspired Memory-Enhanced YOLO for Context-Aware Urban Crowd Density Analysis

1st Ming Shou An *, 2nd Hye-Youn Lim †, 3rd Dae-Seong Kang †

* *School of Computer Science and Engineering, Xi'an Technological University, No.2 Xuefu zhonglu, Weiyang District, Xi'an, Shaanxi Province, China*

† *dept. Electronics Engineering of Dong-A University, 37, Nakdong-daero 550beon-gil, Saha-gu, Busan, Korea*

Emails: *1st anmingshou@xatu.edu.cn, 2nd hylim@dau.ac.kr, 3rd dskang@dau.ac.kr*

***Abstract*—Recent large-scale crowd disasters such as the Itaewon tragedy underscore the need for proactive crowd behavior analysis and advanced surveillance in urban environments. While YOLO (You Only Look Once) provides accurate real-time object detection, its frame-by-frame focus limits understanding of temporal changes in crowd density and transient risk patterns. This paper proposes Hippocampal-Inspired Memory-Enhanced YOLO (HiMeYOLO), which augments a YOLO-based crowd density pipeline with an episodic memory module inspired by hippocampal function. YOLO first detects individuals and estimates zone-wise crowd counts per frame, and the Episodic Memory Unit (EMU) then models temporal sequences of zone densities to capture evolving risk trends. Experiments on public crowd datasets and simulated sequences show that HiMeYOLO improves anomaly detection performance by about 5.2% in F1-score and anticipates rapid density escalation 2–3 frames earlier, demonstrating its potential for predictive urban crowd safety management.**

Keywords—crowd analysis, population density estimation, YOLO, hippocampal-inspired memory, episodic memory, anomaly detection

I. Introduction

Crowd disasters in urban settings, such as the Itaewon tragedy, highlight a global imperative for intelligent surveillance that moves beyond mere headcounts. The root cause often lies not in a single moment of high density, but in the rate and pattern of density change over time. Therefore, effective crowd safety management requires the ability to analyze population density trends in real-time within their temporal context and to identify dangerous patterns proactively[1].

In computer vision-based crowd analysis, YOLO-series object detectors have become a cornerstone due to their high accuracy and real-time performance[2]. For density estimation, YOLO provides an intuitive approach by detecting persons and counting them within predefined zones. However, conventional YOLO-centric methods are inherently frame-independent. They accurately measure instantaneous density but fail to preserve and utilize long-term temporal context—such as how density has evolved over the past minute or whether the current increasing trend resembles previously observed hazardous patterns. This limitation restricts their predictive power and early warning capability for dynamic risks like sudden inflows or slowly building congestion.

Meanwhile, advancements in artificial intelligence have seen the development of memory-augmented neural networks inspired by the human hippocampus, such as Episodic Memory Networks[3] and Neural Turing Machines[4]. These models mimic the brain's ability to store experiences as episodes and recall relevant past contexts to inform current decisions. They have proven effective in tasks requiring long-term dependency modeling. However, the integration of such bio-inspired memory architectures into real-time, object detection-based crowd analysis pipelines remains largely unexplored.

This paper bridges this gap by proposing HiMeYOLO, a unified framework that integrates the efficient real-time person detection capability of YOLO with the contextual reasoning principles of a hippocampal-inspired Episodic Memory Unit (EMU). The core idea is to treat the time series of zone-based population counts generated by YOLO as "episodes." The EMU stores these episodes and performs similarity-based recall, allowing the system to understand crowd dynamics contextually through historical patterns. This enables more accurate anomaly judgment and provides predictive insights beyond simple thresholding.

The main contributions of this work are:

- A Novel Hybrid Architecture: We design the first end-to-end framework that combines a YOLOv8-based detection/counting pipeline with a bio-inspired episodic memory module for real-time crowd analysis.
- Context-Aware Density Analysis Algorithm: We propose an algorithm that goes beyond per-frame area-based population calculation. It patterns density change trajectories over time and detects risk patterns via memory-based similarity matching.
- Enhanced Performance & Predictability: Experiments show our model outperforms conventional YOLO-based thresholding in F1-Score (0.901 vs. 0.808) with lower false alarm rates and demonstrates potential for trend-based early warning.
- Practicality & Deployment Feasibility: The system is compatible with existing CCTV infrastructure without requiring additional expensive sensors. The EMU is designed to be lightweight, satisfying real-time processing requirements.

II. Related Work

A. YOLO-based Object Detection and Crowd Counting

YOLO revolutionized object detection with its single-shot, real-time paradigm[2]. The latest version, YOLOv8, offers an excellent speed-accuracy balance, making it a de

facto standard for surveillance. For crowd analysis, YOLO is used to detect individuals, and the bounding box count serves as an indirect density estimate[5]. This method allows for intuitive zone-based quantification but struggles with heavy occlusion and, critically, ignores temporal context.

B. Crowed Density Estimation Methods

Beyond YOLO, dedicated crowd density estimation methods like Density Map Estimation(DM-CNN[6], CSRNet[7]) regress a continuous density map over the entire image. They are robust in high-density scenes but computationally heavy and less precise for exact zone-wise headcounts. Transformer-based methods[8] have also been proposed but demand substantial data and resources. Our work prioritizes real-time capability, zone-wise quantification, and the integration of temporal pattern analysis, hence we build upon and extend the YOLO-based approach.

C. Neuroscience-Inspired Memory Architectures

Neural Turing Machines (NTMs)[4] and Memory-Augmented Neural Networks (MANNs)[9] introduced external memory matrices to extend the memory capacity of neural networks. Episodic Memory Networks[3], in particular, emphasize an attention-based mechanism to retrieve past experiences (memory slots) similar to a current query. These have proven effective in tasks like video QA. We present a novel application by defining zone-wise population count sequences as 'episodes' and employing an EMU to store and reconstruct these numerical episodes to capture long-term patterns in crowd dynamics.

TABLE I. CAMPARATIVES SUMMARY OF METHODS

Model Type	***Key Characteristics***	***Strengths***	***Limitations***
YOLO	Real-time detection, zone-based counting	Fast, interpretable, precise headcounts	Ignores temporal context, performance drops under occlusion
Density Map Regression	Pixel-wise density prediction	Robust in very dense scenes	Difficult for exact zone counts, high computation, less real-time
Memory-Augmented Networks	Long-term dependency learning via external memory	Captures contextual patterns, leverages similar past experiences	Complex training, scarce application to real-time vision tasks
HiMeYOLO (Proposed)	YOLO+Hippocampal-inspired EMU	Real-time counting + temporal pattern analysis, predictive insight	Requires tuning of EMU hyperparameters, no need for massive historical data

III. PROPOSED METHOD

The overall architecture of HiMeYOLO is illustrated in Fig.1 and consists of the following stages.

The process begins with an Input Video stream (e.g., from a CITY System at 30 FPS). The video frames are fed into a YOLOv8 DETECTION module, which performs person detection and outputs bounding boxes for each individual.

The detected persons are then processed by the ZONE MAPPER. This module maps each person to a pre-defined spatial zone within the scene and calculates the population count for each zone, generating a zone-wise count vector.

This sequence of count vectors is passed to an LSTM ENCODER, which models the temporal dependencies in the crowd dynamics. The encoder outputs a hidden state (ht) that encapsulates the temporal context of the scene.

A core innovation of the system is the Episodic Memory Unit(Hippocampal-inspired). This unit simulates the memory functions of the brain's hippocampus. It consists of a Memory Matrix with 256 memory slots (represented in the diagram as m1, m2, m3...), each storing a past "episode" (a combination of a hidden state ht and a zone vector zt). The unit has two key controllers:

- Write Controller: Determines when to store a new episode into memory, based on a significant change in the zone vector or a Least Frequently Used (LFU) replacement policy.
- Read Controller: Retrieves relevant past episodes by computing attention-based similarity scores between the current hidden state and all stored memory slots.

The output of this memory retrieval process is a Context Vector, which is a weighted summary of the most relevant past experiences.

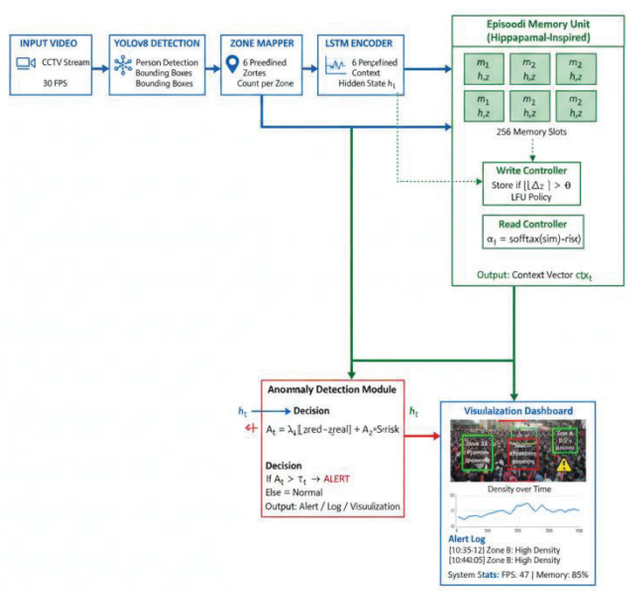


Fig. 1. HiMeYOLO architecture and data processing pipeline.

Finally, the system's outputs are displayed on a Visualization Dashboard. This dashboard provides:

- A real-time video display with overlaid zone boundaries and person counts.
- A Density over Time graph showing historical trends.
- An Alert Log listing past anomalous events.
- A System Status panel showing key metrics such as processing speed (FPS: 47), memory usage (85%), and alert status.

The architecture effectively combines the real-time object detection capability of YOLO with the contextual, long-term pattern recognition enabled by a brain-inspired episodic memory, facilitating proactive crowd safety management.

A. Input Data and Zone Definition

- Visual Data: Real-time video streams from CCTV cameras.
- Zone Definition: The surveillance scene is partitioned into pre-defined spatial/functional zones such ad entrance, plaza center, staircase, waiting line, using polygonal masks, reflecting domain knowledge.

B. YOLO-based Real-Time Detection and Zone Mapping

A YOLOv8 model detects person objects in each fram I_t .

$$\mathcal{B}_t = (bbox_1, bbox_2, \ldots, bbox_n) = YOLO(I_t) \quad (1)$$

Where, $\mathcal{B}_t$ is the set of all detected person bounding boxes at time t.

- The Zone Mapper determines which pre-defined zone Z_k the center point of each bounding box falls into, aggregating counts per zone.

$$c_{t,k} = \sum_{bbox \in \mathcal{B}_t} 1(Center(bbox) \in Z_k) \quad (2)$$

$$z_t = \left[c_{t,1}, c_{t,2}, \ldots, c_{t,K}\right]^T$$

- Here, z_t is the zone-wise population count vector at time t, forming the primary time-series input for our system.

C. Sequence Encoding(LSTM)

To learn temporal dependencies in the zone count vectors, we employ an LSTM layer. The LSTM's hidden state serves as a temporal representation summarizing crowd density changes up to time, which is passed to the EMU.

D. Hippocampal-inspired Episodic Memory Unit(EMU)

The EMU mimics how the human hippocampus forms, stores, and recalls episodic memories. This allows the system to learn from past crowd patterns and use that knowledge to better understand current situations.

Memory Storage Structure: The EMU contains a set number of memory slots, each capable of storing a complete "episode" of crowd behavior. Each episode combines both the temporal pattern of crowd movement and the spatial distribution of people across different zones.

When Memories are Stored: New episodes are saved to memory in two main cases:

- When significant changes occur: If the crowd distribution changes substantially between consecutive time frames, this indicates an important event worth remembering.
- To manage limited memory: Older, less-used memories are periodically replaced with new ones, ensuring the memory remains relevant and doesn't get filled with outdated information.

Recalling Relevant Past Experiences: When analyzing current crowd behavior, the system searches through all stored memories to find past situations that resemble the current one. It uses a similarity scoring mechanism to identify which past episodes are most relevant.

Creating Contextual Understanding: The system then combines information from these similar past episodes to form a "context vector." This vector summarizes what the system has learned from previous similar situations, providing deeper insight than just analyzing the current moment alone.

E. Context-Aware Anomaly Detection and Alerting

The system detects anomalies by comparing current crowd behavior against both general historical patterns and specific known risks.

- Pattern Deviation Detection: Using the contextual understanding from the EMU, the system predicts what should happen next in normal conditions. Large differences between these predictions and actual observations indicate unusual crowd behavior.
- Known Risk Pattern Matching: The system maintains a collection of previously identified dangerous crowd patterns. If current behavior closely matches any of these known risk patterns, this triggers immediate concern regardless of other factors.
- Combined Alert Decision: These two approaches are weighted and combined into a single anomaly score. This score is compared against a threshold that can adapt based on factors like time of day or event type. When the score exceeds this threshold, the system generates detailed alerts specifying which zone is affected, the nature of the risk, and recommended actions.
- Practical Output: Alerts are displayed on the visualization dashboard with clear visual indicators on the video feed, accompanied by text descriptions of the detected risk and suggestions for intervention if needed.

IV. Results and Discussion

A. Experimental Setup

1) To evaluate the proposed HiMeYOLO framework, we utilized the following datasets:

a) Video sequences from the PETS2009 S2.L1 dataset (crowded street scene)[10] and the Mall Dataset (indoor shopping mall surveillance)[11]. For both datasets, we manually annotated five to six functional zones per scene (e.g., entrance, exit, corridor, open area). Zone-wise ground truth population counts were generated by applying the YOLOv8 detector across all frames and aggregating counts within each defined polygon.

b) Custom Synthetic Dataset: To systematically test the model's sensitivity to temporal dynamics, we created a synthetic video sequence dataset using simulation tools. This dataset includes controlled scenarios such as normal pedestrian flow, gradual congestion buildup, and sudden, rapid influx, providing a benchmark for anomaly trend detection.

2) Implementation Details

All experiments were conducted on a workstation with the following specifications:

- Hardware: NVIDIA GeForce RTX 4090 GPU (24GB VRAM), Intel Core i9-13900K CPU, and 64GB DDR5 RAM.
- Software: The framework was implemented in Python 3.9 using PyTorch 2.0. The object detection backbone utilized the official YOLOv8n (nano) model from Ultralytics for its balance of speed and accuracy.
- Training & Inference: The models were trained using the AdamW optimizer[12] with an initial learning rate of 1e-3 and a batch size of 32. The EMU was configured with 256 memory slots, each of 128 dimensions. The dynamic threshold τττ was initialized based on the validation set's anomaly score distribution and was updated online using a moving average of recent scores.

3) Comparison Models:

We compared HiMeYOLO against three baseline models to isolate the contribution of its memory module:

a) YOLO-Only: A baseline that applies a simple, static threshold to the real-time zone-wise person count. An anomaly is flagged if the count in any zone exceeds its predefined limit.

b) YOLO+LSTM: This model extends YOLO-Only by adding an LSTM layer to model the temporal sequence of zone counts, but it lacks the external episodic memory mechanism.

c) YOLO+Transformer: A more advanced baseline where the LSTM is replaced with a standard Transformer encoder to model long-range dependencies in the count sequence.

4) Evaluation Metrics:

Performance was evaluated using standard classification and timeliness metrics: Accuracy, Precision, Recall, F1-Score, Detection Latency (in frames).

a) Accuracy: The proportion of total frames correctly classified as normal or anomalous.

b) Precision: The ratio of correctly detected anomalies to all frames flagged as anomalous.

c) Recall (Detection Rate): The ratio of correctly detected anomalies to all actual anomalous frames.

d) F1-Score: The harmonic mean of Precision and Recall, providing a single balanced metric.

e) Detection Latency: Measured in frames, this indicates the delay between the actual onset of an anomaly and its detection by the system. A lower latency is critical for timely intervention.

B. Crowd Anomaly Detection Performance

As shown in Table II, the proposed HiMeYOLO framework achieves superior performance compared to all baseline models. The model achieves the highest accuracy (92.1%), precision (89.8%), recall (90.5%), and a leading F1-Score of 0.901 on the simulation dataset. This represents a substantial 9.3% improvement in the critical F1-Score over the basic YOLO-Only baseline, which relies solely on static thresholding.

The performance gain is attributed to the integrated episodic memory, which enables contextual analysis beyond simple sequential learning. While the YOLO+LSTM model shows better sequential understanding than YOLO-Only, it lacks a mechanism for long-term pattern retention. The YOLO+Transformer model, though powerful, introduces higher detection latency (3-5 frames), making it less suitable for real-time response. HiMeYOLO effectively bridges this gap, offering enhanced recall of subtle risks and improved precision by reducing false positives through memory-based pattern matching. It maintains a practical detection latency of 1-3 frames, providing a balanced solution for accurate and timely crowd anomaly detection in dynamic environments.

TABLE II. CROWD ANOMALY DETECTION PERFORMANCE COMPARISON(SIMULATION DATASET)

Model Type	*Accuracy (%)*	*Precision (%)*	*Recall (%)*	*F1-Score*	*Detection Latency(Freames)*
YOLO	85.2	78.6	83.1	0.80	0
YOLO+ LSTM	88.7	84.3	86.5	0.85	1-2
YOLO+ Transformer	89.5	86.1	87.2	0.86	3-5
HiMeYOLO (Proposed)	92.1	89.8	90.5	0.90	1-3

C. Computational Efficiency Analysis

Table III presents a comparison of computational efficiency among the evaluated models. The proposed HiMeYOLO achieves an inference time of 21 ms per frame, translating to a processing speed of approximately 47 FPS. This performance comfortably exceeds the standard real-time surveillance requirement of 25 FPS.

TABLE III. MODEL EFFICIENCY COMPARISON(FRAMES PER SECOND)

Model	*Inference Time*	*FPS*	*Model Size (MB)*
YOLOv8 (Detection only)	15	~66	22
YOLO+ LSTM	18	~55	25
YOLO+ Transformer	35	~28	58
HiMeYOLO (Proposed)	21	~47	28

While the vanilla YOLOv8 detection model is the fastest at 66 FPS, it lacks temporal analysis capabilities. The addition of an LSTM layer introduces a moderate overhead, reducing speed to 55 FPS. In contrast, the Transformer-based model suffers from significant computational burden,

slowing down to 28 FPS due to its quadratic attention complexity, which hinders real-time deployment.

Although HiMeYOLO's Episodic Memory Unit adds some computational cost compared to the YOLO+LSTM baseline, it remains over 65% faster than the Transformer model. With a compact model size of 28 MB, HiMeYOLO demonstrates an effective trade-off, delivering advanced memory-augmented analysis while maintaining practical efficiency for resource-constrained edge or server-based monitoring systems.

D. Ablation Study

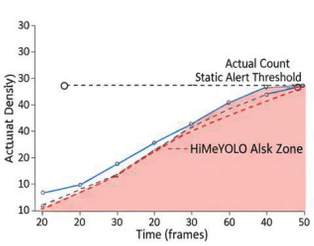
The contribution of each component was analyzed.

- w/o EMU: Removing the EMU dropped the F1-Score from 0.901 to 0.854, confirming the importance of episodic context.
- w/o LSTM: Removing the LSTM and feeding directly to the EMU caused a severe drop in F1-Score to 0.821 due to the loss of sequential dependency learning.
- w/o Zone Mapping: Using only the total frame count instead of zone-wise analysis severely degraded recall, highlighting the loss of localized risk detection capability.

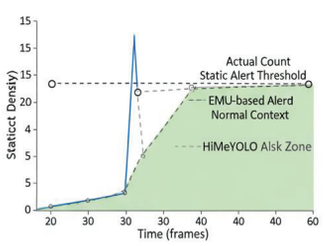
E. Case Analysis and Visualization

The contribution of each component was analyzed.

The system demonstrated its ability to provide an early warning (approx. 3 frames lead time) in a gradual congestion scenario (Figure 2a). While a simple threshold method would not yet trigger an alert, the EMU recalled similar historical increase patterns, prompting an early warning. Conversely, a temporary spike caused by a group moving across the scene (Figure 2b) was filtered out as a non-persistent anomaly, as the EMU context did not support a sustained risk pattern, demonstrating effective false positive suppression.



(a) Gradual Congestion: Early Warning



(b) Temporary Spike: False Positive Suppression



(c)Zone A Density



(d) Temporary Spike: False Positive Suppression

Fig. 2. Examples of early warning of crowd congestion risk and suppression of false positives for temporary spikes by HiMeYOLO

In Fig,2, This figure visually illustrates how HiMeYOLO leverages temporal context to detect crowd risks earlier while simultaneously reducing unnecessary false alarms.

- (a) In a gradual congestion scenario, a static alert threshold triggers an alarm only at a late stage, whereas the HiMeYOLO risk zone (red area) learns the increasing density trend and provides an earlier warning. The x-axis denotes time (frames) and the y-axis represents actual density. The blue solid line shows the change in the number of people, the black dashed line indicates the static alert threshold, and the red shaded region with a dashed line denotes the HiMeYOLO risk zone. In a situation where the crowd density slowly increases, the static threshold raises an alert only near the end, whereas HiMeYOLO captures the upward trend earlier and issues a timely early warning.
- (b) Temporary Spike: False Positive Suppression.This graph shows a case where density briefly spikes and then quickly drops. The blue line represents the momentary spike, and the black dashed line is the static threshold. With static thresholding alone, such a spike would trigger an alarm, but the EMU-based HiMeYOLO considers the normal patterns before and after the spike and treats it as a temporary fluctuation, suppressing the false alarm (green area).
- (c) In the plaza scene, HiMeYOLO issues an "EARLY WARNING" label for Zone A based on the current count (28 people) and the increasing trend (+8 people/min). In a large plaza video, Zone A is highlighted with a yellow box, and the overlay displays the number of people (28) and the increase rate (+8 people/min). HiMeYOLO considers not only the current count but also the rate of increase, producing an "EARLY WARNING" icon so that operators can be alerted before the situation reaches a critical threshold.
- Temporary Spike: False Positive Suppression (indoor corridor scene). In the corridor environment, Zone B is highlighted with a green box, showing the count (5 people), a "YOLO-Only Alert" label, and the final status ("Status: Normal"). Although a YOLO-only criterion might treat this as an alert candidate, the EMU recognizes that the overall density trend is stable and, therefore, does not classify it as a risk, labeling the trend as "Stable – EMU Suppressed."

These examples demonstrate that HiMeYOLO can provide trend-based early warnings and context-aware false positive suppression, enabling more reliable decision-making for real-time crowd safety monitoring.

V. Conclusion

This paper proposed HiMeYOLO, a novel framework that fuses the real-time object detection capability of YOLO with a neuroscience-inspired episodic memory module for context-aware urban crowd density analysis and early anomaly detection. Experimental results demonstrated that the proposed method outperforms conventional static thresholding and sequential models, providing not only superior detection performance but also predictive insights into crowd trend evolution while maintaining computational efficiency. This work contributes a step towards transitioning

crowd safety management from reactive monitoring to proactive prevention. Future work will focus on validation with more diverse real-world data, enhancing the EMU's autonomous learning of risk patterns, and extending the framework to multi-camera scenarios for holistic scene understanding.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(No.RS-2023-00247045)

REFERENCES

[1] SON, Changwon, et al. 158 deaths at Halloween night: An AcciMap analysis of 2022 Itaewon crowd crush in South Korea. Safety Science, 2025, 184: 106741.

[2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779-788.

[3] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-learning with memory-augmented neural networks," in International Conference on Machine Learning, 2016, pp. 1842-1850.

[4] A. Graves, G. Wayne, and I. Danihelka, "Neural turing machines," arXiv preprint arXiv:1410.5401, 2014.

[5] G. Jocher, A. Chaurasia, and J. Qiu, "YOLO by Ultralytics," Jan. 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[6] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 589-597.

[7] Y. Li, X. Zhang, and D. Chen, "CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1091-1100.

[8] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in International Conference on Learning Representations, 2021.

[9] IATROPOULOS, Georgios; BREA, Johanni; GERSTNER, Wulfram. Kernel memory networks: A unifying framework for memory modeling. Advances in neural information processing systems, 2022, vol.35, pp. 35326-35338.

[10] J. Ferryman and A. Shahrokni, "PETS2009: Dataset and challenge," in IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, 2009, pp. 1-6.

[11] K. Chen, C. C. Loy, S. Gong, and T. Xiang, "Feature mining for localised crowd counting," in British Machine Vision Conference, 2012, pp. 1-11.

[12] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in International Conference on Learning Representations, 2019.

Design and Implementation of Multi-Level Voice Phishing Detection Model Based on LLM

SiHoon Kim*, Kwangil Jeon*, SoHyun Beak*, Syed Shehryar Ali Naqvi†, DoHyeun Kim*
* *Department of Computer Engineering, Jeju National University, Jeju-do, Republic of Korea*
† *Department of Electronics Engineering, Jeju National University, Jeju-do, Republic of Korea*
Emails: fpem617@gmail.com, goodjki11@gmail.com, qorthgus1112@gmail.com, syedshehyar@stu.jejunu.ac.kr, kimdh@jejunu.ac.kr

***Abstract*—As voice phishing evolves to exploit conversational contexts, traditional binary classification has become insufficient. This study proposes a hierarchical detection model that segregates the roles of NLP, LSTM, and LLM modules. The KoBERT-based NLP module generates token embeddings for an LSTM module, which captures temporal dependencies to classify block-level labels (intent, conversational patterns, and tactics). Simultaneously, an LLM module provides supplementary inference on speaker roles and emotional cues using a self-consistency scheme. Evaluated on the KorCCVI_v2 dataset, the model achieved a perfect F1-score (1.0) for conversation-level detection and an average of 0.6 for block-level classification. While performance varied across specific categories such as direct-action requests and behavioral control, this hierarchical framework provides a robust foundation for detecting sophisticated phishing strategies beyond simple binary discrimination.**

Keywords— Voice phishing, KoBERT, LSTM, LLM, multi-layer labeling, KorCCVI_v2

I. Introduction

The increasing sophistication of voice phishing has rendered traditional binary classification approaches insufficient for detecting modern, context-driven attack patterns [1]. National statistics highlight this shift: while the frequency of incidents has declined, total financial damage continues to rise, underscoring the need for detection systems capable of analyzing increasingly elaborate methods [2]-[4]. To address these limitations, this study proposes a multi-level labeling framework that identifies not only the presence of phishing but also specific intents, conversational patterns, and tactics. The system integrates three key components:

The proposed system utilizes a KoBERT-based NLP module to transform dialogue into structured sequential embeddings, which are then processed by an LSTM module that captures temporal dependencies to perform block-level classification for intent (Label-2), conversational patterns (Label-3), and tactics (Label-4). Additionally, an LLM module enhances the system's analytical depth by providing supplementary inference on speaker roles and emotional inducement, thereby complementing the LSTM's predictions. By integrating these modules, the proposed framework enables a hierarchical analysis (Labels 1-4) of phishing dialogues. This approach moves beyond simple binary discrimination to establish a foundation for a practical, fine-grained detection system capable of mitigating complex real-world threats.

II. Related work

A. Existing Studies on Voice Phishing Detection

Existing studies on voice phishing detection have predominantly relied on binary classification methods that distinguish legitimate conversations from fraudulent ones. Early text-based approaches commonly employed BERT-derived Korean language models to extract contextual representations from dialogue. At the same time, audio-based research focused on identifying abnormal prosodic or emotional patterns characteristic of phishing calls. Some studies explored multimodal approaches by combining KoBERT text embeddings with acoustic feature extraction, reporting improved detection performance compared to unimodal models.

In natural language processing research, BERT-based models have demonstrated strong performance in capturing semantic dependencies in Korean text, and LSTM architectures have been widely used to incorporate temporal structure into phishing detection. However, most prior work applied these models solely to enhance binary classification accuracy and did not address finer-grained behavioral components such as intent, conversational patterns, or tactical strategies [5]. Lightweight classification models designed through knowledge distillation have also been proposed, but these approaches similarly remain limited to phishing-versus-non-phishing discrimination without hierarchical label structures [6].

Large language model–based approaches have also been examined recently. Studies employing GPT-based models attempted to classify phishing conversations using semantic reasoning [7], though their performance strongly depended on prompt formulation and lacked consistent multi-layer outputs. In parallel, audio-based voice fraud detection research has investigated robustness against adversarial perturbations and synthetic voice manipulation, utilizing architectures such as RawNet3 to analyze fraudulent or manipulated speech signals [8]. These findings highlight the increasing sophistication of real-world voice phishing and the need for more comprehensive analytical frameworks.

Despite these developments, most existing research remains limited to binary decision-making. It does not incorporate detailed labeling schemes capable of distinguishing the diverse behavioral characteristics observed in actual phishing scenarios [9]. Phishing conversations often involve multiple intertwined strategies—such as personal information requests, financial transfer inducement, and psychological pressure but these are rarely modeled as separate analytical dimensions. Consequently, binary models struggle to detect newly emerging or intentionally disguised phishing attempts. A multi-level labeling framework that distinguishes intent, conversational patterns, tactics, and speaker-related cues is therefore required to overcome these limitations and enhance applicability in real-world environments [10].

III. DESIGN OF THE MULTI-LEVEL VOICE PHISHING CLASSIFICATION MODEL

A. Overall System Architecture

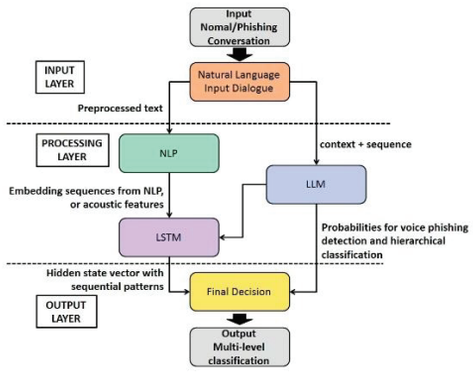


Fig. 1. Overall architecture of the proposed multi-level voice phishing classification model

The overall architecture of the proposed multi-level voice phishing classification model is illustrated in Fig. 1. The system preprocesses conversational text, converts it into a machine-readable format, performs multi-level label classification, and finally determines whether a conversation constitutes voice phishing. The model is composed of four principal components:

- An NLP module for text preprocessing
- An LSTM module for sequential learning and block-level classification
- An LLM module that provides supplementary inference on speaker roles and emotional inducement
- A decision module that integrates outputs from the preceding components to generate final phishing predictions

As shown in Fig. 2, the NLP module employs a KoBERT-based tokenizer to transform conversational text into token embeddings and attention masks. These embeddings serve as input to the LSTM module, which predicts block-level labels: intent (Label-2), conversational patterns (Label-3), and tactics (Label-4). In parallel, the LLM module analyzes the same conversation blocks to infer speaker-role characteristics and emotional-inducement cues. A self-consistency strategy is applied to stabilize the LLM outputs and reduce variability.

The decision module aggregates the outputs of both the LSTM and LLM modules to produce integrated predictions at both the block level (Label-2 through Label-4) and the conversation level (Label-1). To accomplish this, the decision module employs a combination of voting, weighted-score fusion, and rule-based strategies, enabling the system to reduce reliance on any single model and enhance robustness in the final phishing assessment.

This study primarily focuses on the design and implementation of the NLP and LSTM components as the core structure of the multi-level classifier. At the same time, the LLM module serves as a supplementary component whose outputs are fused with those of the LSTM to produce the final prediction.

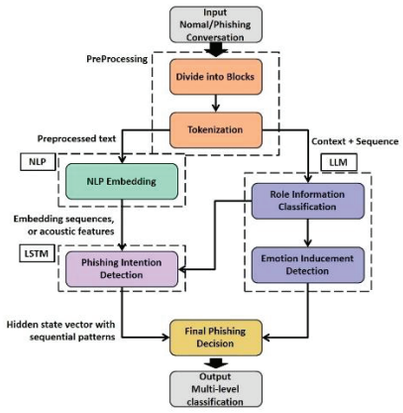


Fig. 2. Multi-level processing flow of the proposed phishing detection system

B. NLP-Based Preprocessing Design

The structure of the NLP preprocessing pipeline is depicted in Fig. 3. Conversational text is first tokenized using a KoBERT-based tokenizer and subsequently embedded into tensors suitable for LSTM processing [8]. The input text is segmented into tokens, including a CLS token to capture the overall semantic summary. Right-padding is applied to maintain sequence order, while sequences exceeding the maximum length are truncated. This approach ensures consistent input length for the LSTM while accommodating limitations inherent to transformer-based models [9]. During embedding, token vectors are generated with shape [batch_size, seq_len, embed_dim]. An attention mask is also produced to distinguish between actual tokens and padded positions, preventing unnecessary loss computation during LSTM training. KoBERT's pretrained model is used to generate embeddings, and the entire preprocessing workflow is automated through the PyTorch framework. The resulting embeddings and masks are forwarded to the LSTM module for sequential learning.

C. LSTM-Based Multi-Level Classification Design

The structure of the LSTM-based multi-level classifier is presented in Fig. 4. This module receives token embeddings and attention masks generated by the NLP preprocessing pipeline and learns temporal dependencies within conversational sequences. Embedded token sequences are processed sequentially through the LSTM network, where contextual information is accumulated and encoded via hidden states. At the same time, the attention mask ensures that padding tokens are not included in the learning process. The final hidden state vector is then passed into a fully connected layer, which produces block-level predictions for the following three labels:

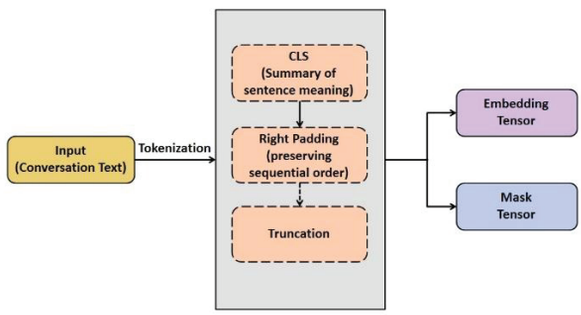


Fig. 3. NLP-based embedding and preprocessing pipeline.

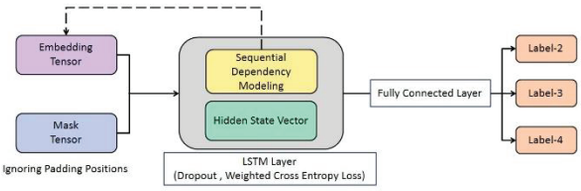


Fig. 4. LSTM-based multi-level classification module.

- Label-2 (Intent): classification of the speaker's objective, such as personal information solicitation, direct action requests, and psychological pressure
- Label-3 (Conversational Pattern): identification of conversational progression types, including normal responses, crisis induction, fabricated scenarios, trust-building mediation, and direct or composite demands
- Label-4 (Tactic): detection of tactical strategies used within the dialogue, such as psychological manipulation or behavioral control

Through these functions, the LSTM-based classifier serves as the core component that directly identifies phishing-related attributes at the block level across intent, conversational patterns, and tactics.

In addition, block-level analysis is supplemented by auxiliary inference from the LLM module, which predicts speaker roles (e.g., impersonated institutions, acquaintances, or service agents) and emotional-inducement cues (e.g., signals of pressure, anxiety induction, or coercion). These LLM-derived cues are generated from the same conversational blocks and provide contextual and affective information that the LSTM cannot capture directly. The auxiliary signals are subsequently integrated with the LSTM outputs to enhance the interpretability and reliability of block-level phishing assessments.

Accordingly, the LSTM-based classifier performs the primary task of multi-level block classification. At the same time, the LLM module contributes complementary contextual and psychological indicators, enabling a more robust and comprehensive detection framework.

D. LLM-Based Speaker Role and Emotional Inducement Analysis

Fig. 5 illustrates the LLM-based module used for supplementary inference. This component processes block-level conversational inputs to determine speaker roles and emotional inducement.

The LLM receives preprocessed embeddings and contextual information and classifies the speaker's role into one of seven categories: public institution, financial institution, family or acquaintance, customer service agent, business partner, delegated representative, or general individual. It also assesses whether an utterance induces psychological pressure, such as anxiety or emotional manipulation.

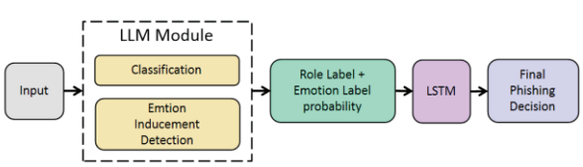


Fig. 5. LLM-based emotional inducement detection process.

The Mistral-7B model deployed in the Ollama environment is used for this module. For reliability, multiple outputs are generated per input, and the final prediction is selected via a self-consistency (majority voting) strategy. This reduces inference variation and enhances output stability. The LLM outputs probability scores for each label, which are later fused with LSTM outputs in the final decision stage.

E. Final Phishing Decision Integration

The final determination of whether a message is phishing is made by integrating the outputs of the LSTM and LLM modules. Fig. 6 illustrates the architecture of the decision module, which combines block-level and conversation-level information to generate the final prediction (Label-1). The module receives heterogeneous outputs from both components and fuses them through a multi-strategy integration process designed to enhance reliability and robustness.

The LSTM module generates block-level probability outputs for intent (Label-2), conversational pattern (Label-3), and tactic (Label-4) based on linguistic and structural cues extracted from the conversational text. In parallel, the LLM module provides auxiliary probability signals for speaker-role classification and emotional-inducement detection derived from the same conversation blocks. Because these outputs capture different dimensions of phishing behavior linguistic structure via LSTM and contextual-psychological cues via LLM, the decision module integrates them using three complementary strategies.

- Voting-based integration compares the directional consistency between LSTM-derived risk indicators and LLM-derived cues. When multiple outputs support the same risk interpretation, the corresponding label is prioritized, reflecting a consensus between models
- Weighted-score fusion aggregates the probability outputs from both modules with assigned weights; the LSTM predictions are given higher weights as the primary classifier, while the LLM outputs function as auxiliary contextual signals. This fusion produces a refined block-level phishing score
- Rule-based refinement applies additional domain-informed conditions observed in voice phishing scenarios.

For instance, when the LSTM predicts high-risk intent and tactic labels simultaneously and the LLM detects strong emotional inducement, the rule-based component reinforces the phishing likelihood of the block, thereby increasing the reliability of the prediction.

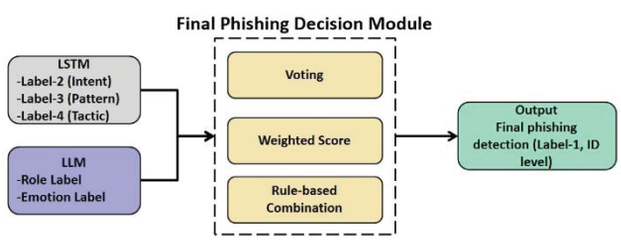


Fig. 6. Final Phishing Decision Module Integrating LSTM and LLM Outputs.

This multi-stage fusion approach reduces reliance on any single model and enables the system to leverage both linguistic structure and contextual and emotional cues. The

fused block-level results are then aggregated across all blocks belonging to the same conversation ID. During this aggregation, the decision module considers the distribution of block-level predictions and the consistency of LLM-derived auxiliary cues throughout the conversation. The final output, Label-1, represents the conversation-level determination of phishing presence.

IV. IMPLEMENTATION OF THE MULTI-LEVEL VOICE PHISHING CLASSIFICATION MODEL

A. Experimental Environment

TABLE I. EXPERIMENTAL IMPLEMENTATION ENVIRONMENT

Component	Specification
Operation System	Window 11
Processor(CPU)	AMD Ryzen5 5600 6-Core Processor
Graphic Processing Unit	NVIDIA GEFORCE RTX 4060
RAM	16 GB DDR4
Storage	1 TB NVMe SSD
Python Version	3.12.9
Deep Learning Framework	PyTorch 2.5.1

The KoBERT tokenizer and pretrained embedding model were used for natural language processing, and the LSTM classifier and the LLM-based inference module were integrated into a single system pipeline, as summarized in Fig. 7 (Algorithm 1). Fig.7 describes the end-to-end workflow from conversation segmentation and KoBERT-based preprocessing to LSTM block-level classification, LLM-based auxiliary inference, and the final fusion step for phishing determination. The Mistral-7B model, executed via the Ollama runtime, communicated with the system via a REST API to provide the LLM module. All experiments were carried out in the implementation environment summarized in Table I, and the KorCCVI_v2 dataset was collected, annotated, and used for training and evaluation.

B. Implementation of the Multi-Level Detection System

Conversational text data were preprocessed using the KoBERT tokenizer. A CLS token was inserted at the beginning of each sequence, and Right-padding or Truncation was applied to maintain a fixed sequence length. An attention mask was generated to distinguish meaningful tokens from padded positions. The output of this preprocessing stage consisted of:

- A token embedding tensor of size $[batch, seq_len, embed_dim]$
- An attention mask tensor of size $[batch, seq_len]$.

These tensors served as the primary inputs to the LSTM classifier.

The LSTM-based classifier received the embedding and attention mask as inputs and performed sequential modeling. Hidden state vectors obtained from the LSTM layers were passed into separate fully connected layers corresponding to:

- Label-2 (intent)
- Label-3 (conversational pattern)
- Label-4 (tactic)

Algorithm 1: Fully Detailed Multi-Level Workflow for Voice Phishing Detection

Input: Conversational dataset D with conversation IDs and timestamps; KoBERT tokenizer T; LSTM classifier f_{LSTM}; LLM module f_{LLM}; fusion module F_{dec}
Output: Block-level labels (Label2, Label3, Label4), conversation-level label (Label1)
Step 0: Conversation segmentation and normalization;
for *each raw conversation* r *in* D **do**
  extract conversationID from metadata;
  sort utterances in r by timestamp;
  remove noise tokens (extra spaces, system markers);
  normalize punctuation and unify Korean spacing;
  merge very short utterances to avoid fragmentation;
  define segmentation boundaries based on change of speaker, long silence, intent shift;
  split r into blocks $b_1, b_2, \ldots, b_K$;
  store (conversationID, blockIndex, blockText);
Step 1: KoBERT-based NLP preprocessing;
for *each block (ID, i, text)* **do**
  tokens $\leftarrow T(text)$;
  insert special tokens [CLS], [SEP];
  apply right-padding to max length L;
  create attention mask where real tokens are 1 and padding is 0;
  emb $\leftarrow$ KoBERTEmbedding(tokens);
  store (ID, i, emb, mask, text);
Step 2: LSTM-based multi-label block classification;
for *each stored entry (ID, i, emb, mask, text)* **do**
  $h \leftarrow f_{LSTM}(emb, mask)$;
  $y_2 \leftarrow FCIntent(h)$;
  $y_3 \leftarrow FCPattern(h)$;
  $y_4 \leftarrow FCTactic(h)$;
  store LSTMOut[i] = (ID, y2, y3, y4);
Step 3: LLM-based auxiliary inference (speaker role and emotional inducement);
for *each stored entry (ID, i, emb, mask, text)* **do**
  roleScoresSet = empty list;
  emoScoresSet = empty list;
  for $t = 1$ **to** N **do**
    rs $\leftarrow f_{LLM}(text, role)$;
    es $\leftarrow f_{LLM}(text, emotion)$;
    append rs to roleScoresSet;
    append es to emoScoresSet;
  finalRole $\leftarrow$ SelfConsistency(roleScoresSet);
  finalEmotion $\leftarrow$ SelfConsistency(emoScoresSet);
  store LLMOut[i] = (ID, finalRole, finalEmotion);
Step 4: Multi-strategy fusion for block-level decision;
for *each block index* i **do**
  read (y2, y3, y4) from LSTMOut[i];
  read (finalRole, finalEmotion) from LLMOut[i];
  voteResult $\leftarrow$ Voting(y2, y3, y4, finalRole, finalEmotion);
  weightedScore $\leftarrow$ WeightedFusion(y2, y3, y4, finalRole, finalEmotion, wLSTM_intent, wLSTM_pattern, wLSTM_tactic, wRole, wEmotion);
  refinedLabel $\leftarrow$ RuleBasedRefine using finalEmotion and finalRole together with y2, y3, y4;
  fused $\leftarrow F_{\text{dec}}(voteResult, weightedScore, refinedLabel)$;
  store BlockLabel[i] = (ID, fused);
Step 5: Aggregation to conversation-level decision;
for *each conversation ID* c **do**
  collect all fused block labels for blocks belonging to c;
  compute distribution over fused labels;
  compute risk trajectory by tracking label changes over time;
  compute aggregated risk score using highRiskCount and maxConsecutiveRisk and tacticalRiskBoost;
  Label1[c] $\leftarrow$ Aggregate(highRiskCount, maxConsecutiveRisk, tacticalRiskBoost);
  output final Label1 and all per-block labels for conversation c;

Fig. 7. Final Phishing Decision Module Integrating LSTM and LLM Outputs

Weighted Cross-Entropy Loss was applied to mitigate the class imbalance present in the dataset. Dropout regularization was added during training to reduce overfitting and improve generalization. Block-level prediction results from this classifier were later used in the final decision process.

The LLM-based module processed block-level conversational text to infer speaker roles and emotional inducement patterns. The Mistral-7B model running in the Ollama environment produced multiple outputs per block. A self-consistency strategy implemented through majority voting was used to determine the final Role Label and Emotion Label for each block. These outputs were subsequently passed to the system's final decision module along with the LSTM predictions. The final phishing classification for each conversation (Label-1) was determined by integrating outputs from both the LSTM and LLM modules. Three fusion strategies were applied:

- Voting,
- Weighted score aggregation, and
- Rule-based integration that incorporated role- and emotion-related cues.

This multi-strategy fusion ensured that block-level attributes (intent, conversational pattern, tactic) and LLM-derived inferences (speaker role and emotional inducement) were jointly considered when producing the final conversation-level phishing or non-phishing label.

V. Performance Analysis

A. Best Hyperparameter Config of Model

TABLE II. Best-Performing Hyperparameter Settings

Hyperparameters	best-performing configuration
Learning Rate	1e-3
Weight Decay	1e-3
Hidden Dimension	128
Dropout	0.3

Experiments were conducted on the KorCCVI_v2 dataset using text-only inputs at both ID and block levels (4 utterances/block). The model, implemented in PyTorch, utilized KoBERT for embedding and an LSTM for sequence classification, applying weighted Cross Entropy Loss and dropout to optimize performance. For auxiliary inference, Mistral-7B generated speaker-role and emotional-inducement predictions via a self-consistency strategy. Finally, a decision module integrated LSTM and LLM outputs using weighted fusion or voting to produce the final phishing predictions.

B. ID-Level Detection Performance (Label-1)

Label-1 represents the binary classification performed at the conversation (ID) level. The proposed model achieved perfect detection performance, correctly identifying all voice phishing instances and legitimate conversations with zero false positives and zero false negatives. As shown in the matrix, predictions are densely concentrated along the diagonal, confirming an F1-score of 1.0. This result demonstrates that the uncertainties observed in individual block-level predictions (Label-2 through Label-4) were effectively mitigated by the multi-strategy fusion mechanism, which integrates LSTM-based outputs with auxiliary inference from the LLM module. Consequently, the system delivers robust, reliable identification at the global ID level, indicating strong potential for practical deployment.

C. Intent Classification Performance (Label-2)

Label-2 classifies the intent of each utterance into three categories: information request, direct action request, and psychological pressure. As illustrated in the confusion matrix of Fig. 8, a distinct imbalance in class-wise performance is observed. While the information request and psychological pressure categories exhibit a dense concentration of correct predictions along the diagonal, the direct action request class demonstrates significantly more off-diagonal misclassifications. This discrepancy is primarily due to severe class imbalance and the limited number of training samples in the direct action category. These findings, evidenced by the dispersed prediction results in Fig. 9, highlight the critical need for dataset expansion and targeted imbalance mitigation strategies to enhance robustness in intent-level classification.

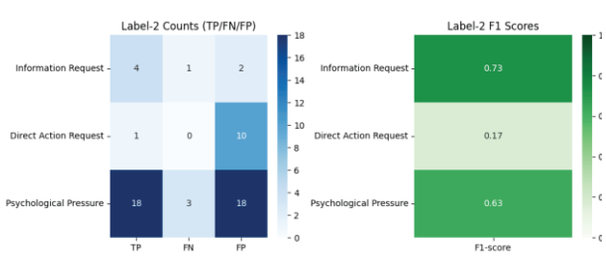


Fig. 8. Intent Classification Performance (Label-2)

D. Conversational Pattern Classification Performance (Label-3)

Label-3 categorizes conversational patterns into five distinct classes: normal response, anxiety/threat induction, trust mediation, fabricated incident manipulation, and direct request/complex patterns. As presented in the confusion matrix of Fig. 9, the normal response category is classified with near-perfect precision, with predictions densely concentrated along the diagonal. This high performance is attributed to the distinct neutral tone and absence of coercive keywords, which clearly differentiate it from malicious patterns. Similarly, fabricated incident manipulation exhibits a strong diagonal presence, likely due to specific, high-frequency keywords (e.g., "accident," "investigation") that facilitate effective feature extraction. In contrast, trust mediation and direct request/complex patterns show a higher incidence of off-diagonal misclassifications. The matrix reveals that trust mediation is frequently confused with normal responses. This is likely because this category mimics polite, rapport-building behavior to lower the victim's guard, lacking explicit aggressive markers. Consequently, these observations indicate that while the model effectively classifies clearly structured or keyword-driven patterns, it struggles with context-dependent scenarios where malicious intent is camouflaged by polite phrasing.

E. Tactical Behavior Classification Performance (Label-4)

Label-4 differentiates utterances into two tactical classes: psychological manipulation tactics and behavior-inducement/control tactics. As shown in the confusion matrix in Fig. 10, the psychological manipulation category shows a dense concentration of correct predictions along the diagonal. This confirms that the model effectively captures the explicit linguistic markers of coercion and urgency, maintaining high detection reliability for these tactics.



Fig. 9. Conversational Pattern Classification (Label-3)

In contrast, the behavior-inducement/control category exhibits a higher incidence of off-diagonal misclassifications, indicating a struggle to distinguish these tactical patterns from standard conversational sequences. This visual discrepancy is primarily driven by severe class imbalance and the scarcity of training samples in the behavior-control category, which hinders the model's robust generalization of specific features. Consequently, these observations underscore the necessity for data augmentation to mitigate the distributional gap and enhance overall tactical classification performance.

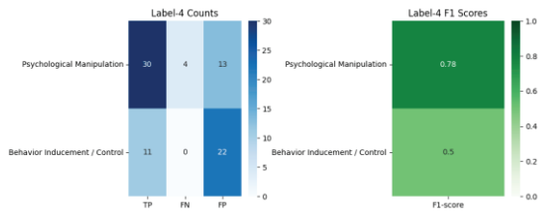


Fig. 10. Tactical Behavior Classification (Label-4)

F. Overall Summary of Experimental Results

TABLE III. OVERALL PERFORMANCE OF THE PROPOSED MULTI-LEVEL VOICE PHISHING DETECTION MODEL

	Label-1	Label2	Label3	Label4
Description	**Detection**	**Intent**	**Pattern**	**Tactic**
Accuracy	1.000	0.999	0.999	0.999
Precision	1.000	0.419	0.557	0.515
Recall	1.000	0.886	0.562	0.652
Mcc	1.000	0.562	0.722	0.681
Macro-F1	1.000	0.508	0.652	0.639

Table III presents a consolidated summary of the experimental results obtained across all classification levels (Label-1 to Label-4). The proposed multi-level detection model demonstrated highly stable performance at the conversation level (Label-1), achieving perfect accuracy and an F1-score of 1.0, with no false positives or false negatives. This indicates that the uncertainty present in block-level predictions was effectively mitigated through the multi-strategy fusion mechanism, which combines LSTM outputs with LLM-based auxiliary inference.

At the block level (Label-2 to Label-4), performance varied depending on class distribution and linguistic complexity. For Label-2 (Intent Classification), the information request and psychological pressure categories showed moderate-to-strong performance, whereas the direct action request category exhibited substantially lower accuracy due to severe class imbalance. For Label-3 (Conversational Pattern Classification), well-defined categories such as normal response and fabricated incident manipulation achieved strong performance. In contrast, context-heavy classes, such as trust mediation and direct request/complex patterns, demonstrated lower F1 scores. For Label-4 (Tactical Behavior Classification), the model reliably detected psychological manipulation tactics but performed more weakly on behavior-inducement/control tactics, primarily due to high false-positive rates stemming from limited training samples and overlapping linguistic cues. Overall, these results confirm that the proposed hierarchical architecture effectively captures multi-dimensional phishing cues and provides robust ID-level predictions suitable for real-world deployment.

VI. CONCLUSION

Voice phishing has recently emerged as a significant social problem, yet most existing detection methods rely solely on binary classification, limiting their effectiveness against newly evolving attack strategies. To address these limitations, a multi-level analysis framework is required, one that incorporates conversational intent, linguistic patterns, tactical behaviors, and speaker roles. In this study, we designed and implemented a multi-level detection model that integrates KoBERT embeddings with an LSTM classifier using the KorCCVI_v2 dataset. In addition, a supplementary LLM module based on Mistral-7B, running in the Ollama environment, was employed to infer speaker roles and emotional inducement, and its outputs were combined with those of the LSTM via a final decision module to produce ID-level predictions. Experimental results demonstrated perfect performance at the conversation (ID) level, achieving an F1-score of 1.0, while the average block-level performance yielded an F1-score of approximately 0.6. Specific categories, specifically direct-action requests in Label-2 (intent) and behavioral control in Label-4 (tactics), performed worse, primarily due to data imbalance in the dataset. Furthermore, the LLM module did not achieve the expected performance in emotional inducement detection, which we attribute to inconsistencies in labeling schemes and constraints in input formatting.

By applying a multi-level labeling mechanism rather than a simple binary classifier, this study demonstrates the feasibility of detecting not only the presence of phishing but also its underlying intent and tactical characteristics. This enables a model capable of providing explanations—namely, “why the conversation was classified as phishing”—which is essential for practical deployment. The findings of this study provide a foundation for developing next-generation intelligent security systems that can adapt to emerging voice phishing strategies through future dataset expansion and reinforcement learning–based enhancements.

ACKNOWLEDGMENT

The following are the results of a study on the " Convergence and Open Sharing System" Project, supported by the Ministry of Education and the National Research Foundation of Korea. Corresponding Author (kimdh@jejunu.ac.kr).

REFERENCES

[1] J. Yoo, H. Han, and W. Lee, “Modeling and real-time alert system for detecting new AI voice phishing,” in Proc. Korea Computer and Information Science Society Winter Conf., vol. 33, no. 1, 2025.

[2] E. Kim, “Analysis of the process of face-to-face voice phishing crimes: Focusing on crime script analysis,” Journal of Criminal Investigation Studies, vol. 8, no. 2, 2022.

[3] S. Oh and H. Song, “A study on characteristics and processes of new voice phishing based on psychological mechanisms,” Journal of the Society of Disaster Information, vol. 19, no. 3, pp. 510–518, 2023.

[4] S. Bae, M. Choi, and H. Lee, “A study on conversational prototypes in voice phishing dialogues: Focusing on impersonation of investigative agencies,” The Sociolinguistic Journal of Korea, vol. 31, no. 4, pp. 203–233, 2023.

[5] S. Choi, J. Jang, and B. Ahn, “Real-time and post-call voice phishing detection and classification system combining KLUE/BERT and LSTM,” in Proc. The Institute of Electronics and Information Engineers Autumn Conf., pp. 1085–1089, 2024.

[6] J. Shim and S. Kim, “Voice phishing detection method using GPT-3.5-based large language models,” Journal of KIISE, vol. 51, no. 1, pp. 67–77, 2024.

[7] X. Wang, J. Wei, and D. Schuurmans, “Self-consistency improves chain of thought reasoning in language models,” in Proc. Int. Conf. Learning Representations (ICLR), 2023.

[8] S. Kim and S. Roh, “Detection of suspicious voice phishing utterances using deep learning–based NLP and knowledge distillation,” Journal of the Korea Institute of Electronic Commerce, vol. 29, no. 4, 2024.

[9] K. Clark, M.-T. Luong, and Q. V. Le, “ELECTRA: Pre-training text encoders as discriminators rather than generators,” in Proc. Int. Conf. Learning Representations (ICLR), 2020.

Dental Condition Diagnosis Using U-Mamba

MinGyu Son, Hyunsik Ahn
Department of Artificial Intelligence
Tongmyong University, Korea
E-mail: thsalsrb615@naver.com, hsahn@tu.ac.kr

Abstract-Recently, the importance of early diagnosis and prevention in dental treatment has been increasingly emphasized. In particular, research interest in utilizing AI technologies for dental condition diagnosis has been growing. This paper proposes a dental condition diagnosis methodology optimized by comparing models from the YOLO and U-Net families. From the tooth image, the dental condition was diagnosed by classifying it into four classes: 'caries', 'cavity', 'crack', and 'tooth'. In the experimnets, U-Mamba showed the best performance among the models for dental condition diagnosis with the COCO teeth image dataset. It was able to precisely divide the lesion and structure in the teeth images showing high performance which resulted in 0.91 of mAP and 0.98 of F1-score.

Keywords - Dental Condition Diagnosis, U-Mamba, teeth dataset, SSM

I. Introduction

Recent dental advancements focus heavily on early detection and preventive health of teeth. Specifically, there is an intensifying research effort to integrate AI technology for more accurate and automated dental assessments. For instance, studies on Dental AI are being conducted using dental medical imaging from specialized equipment, such as X-rays and CT scans, to diagnose dental health [1, 2]. However, such specialized equipment has limited accessibility for individuals since it can only be utilized within dental clinics. On the other hand, research on diagnosing dental conditions based on RGB imagery is also underway [3]; examples include dental detection in smartphone photography using YOLO v5 and dental caries classification using CNN models [4]. Nevertheless, these novel attempts face challenges in practical application due to performance limitations in classification. These studies require further improvement, as they exhibit restricted performance in instance segmentation and are often limited to identifying only a single tooth within a dental image.

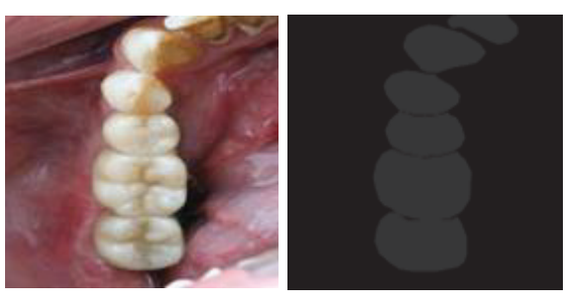
Among instance segmentation algorithms, the YOLO series (e.g., YOLOv8, YOLOv12) is widely utilized; however, the U-Net series, including U-Net [5], U-Net Transformer [6], and U-Mamba [7], has recently gained attention for demonstrating superior performance.

In this paper, we evaluate YOLO-based and U-Net-based models to select the optimal model for a dental condition diagnosis system and propose a mobile-compatible diagnostic system based on these findings. This approach enables individuals to input dental images via their mobile phones and apply the proposed diagnostic framework, allowing anyone to easily monitor their dental health and maintain oral hygiene.

II. Dental Condition Diagnosis

A. Teeth Image Dataset

In this paper, 'Dental Computer Vision Dataset' [8] provided by Robotflow is used to construct a deep learning dataset for tooth object segmentation. The dataset is divided into four classes: 'tooth', 'crack', 'cavity', and 'caries'. The data set is trained with deep learning models by adjusting it from a format in which the existing COCO format to the size of 640 × 640.



(a) (b)

Figure 1. A sample of the teeth image dataset in the COCO format. (a) RGB tooth image (b) tooth region segmented image

B. U-Mamba Model and Learning

In this section, we describe U-Mamba model which produced the best performance in the experiments. As shown in Figure 2, U-Mamba is the model that combines U-NET and SSM (State Space Model) and shows excellent performance in object division. U-Mamba has the advantage of being able to obtain the efficiency of CNN and the global context of SSM through the combination of CNN's regional feature extraction ability and SSM's long-range dependence modeling function. U-Mamba first uses the existing U-NET, and skip connection is constructed in the encoder decoder model of the U-NET. Then we go into the SSM-based Mamba architecture, which has a feature called optional SSM by filtering out unnecessary information. After that, it enters the decoder area and

doubles the dimension of the map in a symmetrical left and right form, reducing the number of channels in half. It is used to reduce the loss of existing data and alleviate the problem of gradient loss through skip connection.

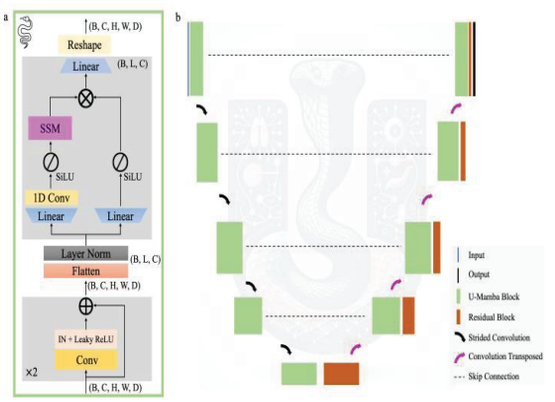


Figure 2. U-Mamba block diagram.

III. Experimental Results

For experiments of the proposed approach, U-Mamba was applied to dental condition diagnosis with the dataset 'Dental Computer Vision Dataset'. It adjusted learning rate for each epoch for learning and 1000 times of epoches. Table 1 shows the results of dental condition diagnosis by applying U-Mamba. It shows the accuracy of 98% on the f1-score and 0.98 of F1 score. Figure 3 is an image obtained by synthesizing with the results of instance segmentation of an original input image. The area of classes were marked by various colors shown in the image. Table 2 shows the comparision results of the multiple models, Yolo v12, DETR (Detection with Transformer), and U-Mamba. U-Mamba produced exellent performance compared to other models.

From the experimental results, it seems that a specialized diagnosis of teeth can be possible when this dental condition diagnosis system is applied to a tooth image when it is personally inputted by a smartphone. If this dental condition diagnosis approach is applied as a smartphone app, it will be able to establish a self dental care service practically.



Figure 3. Dental diagnosis segmentation results.(orange : tooth, green : caries, puple : cavity)

Table 1. The performance of dental condition diagnosis by U-Mamba

Teeth Condition	mAP	f1-score
Tooth	0.98	0.98
Crack	0.88	0.97
Cavity	0.91	0.99
Caries	0.88	0.98
Total avg	0.91	0.98

Table 2. The performance results of models

Models	mAP	f1-score
Yolo v12	0.51	0.53
DETR	0.31	0.32
U-Mamba	0.91	0.98

IV. Conclusion

In this paper, a dental condition diagnosis methodology using U-Mamba to classify segments into four classes of 'Caries', 'Cavity', 'Crack', and 'Tooth' was proposed. U-Mamba model can precisely segment lesions and structures in tooth images, and showed high performance by recording mAP 0.91 and F1-score 0.98. A simple diagnosis of tooth condition can be done personally, so the user can quickly and accurately identify the tooth conditions. This approach can be used for preventive managing of oral health. In future studies, we will seek ways to improve models not only for tooth condition but also for gum management.

References

[1] D. Budagam, A. Kumar, S. Ghosh, A. Shrivastav, A. Z. Imanbayev, I. R. Akhmetov, D. Kaplun, S. Antonov, A. Rychenkov, G. Cyganov, and A. Sinitca, "Instance Segmentation and Teeth Classification in Panoramic X-rays," arXiv preprint arXiv:2406.03747, June 2024.

[2] I. E. Hamamci, S. Er, E. Simsar, A. E. Yuksel, S. Gultekin, S. D. Ozdemir, K. Yang, H. B. Li, S. Pati, B. Stadlinger, A. Mehl, M. Gundogar, and B. Menze, "DENTEX: An Abnormal Tooth Detection with Dental Enumeration and Diagnosis Benchmark for Panoramic X-rays," arXiv preprint arXiv:2305.19112, May 2023. [3] www.trolltech.com

[3] C. Jiang, S. Zhai, H. Song, Y. Ma, Y. Fan, Y. Fang, D. Yu, C. Zhang, S. Han, R. Wang, Y. Liu, J. Li, and P. Qin, "Object Detection for Caries or Pit and Fissure Sealing Requirement in Children's First Permanent Molars," arXiv preprint arXiv:2308.16551, August 2023.

[4] S. J. Kim, "Deep learning analysis for reliability evaluation of dental caries detection," M.S. thesis, Seoul National University, Seoul, Korea, 2019.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," arXiv preprint arXiv:1505.04597, May 2015.

[6] O. Petit, N. Thome, C. Rambour, and L. Soler, "U-Net Transformer: Self and Cross Attention for Medical Image Segmentation," arXiv preprint arXiv:2103.06104, March 2021.

[7] J. Ma, F. Li, and B. Wang, "U-Mamba: Enhancing Long-

range Dependency for Biomedical Image Segmentation," arXiv preprint arXiv:2401.04722, January 2024.
[8] https://universe.roboflow.com/arab-academy-vf9su/dental-7yegp

Enhanced Quantum-Based Mining Strategy for Bitcoin

Yeonsong Suh
Department of Computer Science
Sookmyung Women's University
Seoul, Republic of Korea
yeonsongsuh@sookmyung.ac.kr

Younghoon Park
Division of Computer Science
Sookmyung Women's University
Seoul, Republic of Korea
yh.park@sookmyung.ac.kr

Kwang-Man Ko
Department of Computer Engineering
Sang-Ji University
Wonju, Republic of Korea
kkman@sangji.ac.kr

***Abstract*—Proof-of-Work mining finds a valid Bitcoin block header by trying an extremely large number of hash values, which demands high computation. Quantum search algorithms such as Grover's algorithm can speed up the search, but current quantum mining is still probabilistic and cannot ensure that a valid hash even exists for a fixed difficulty target. We propose a more practical strategy. We expand the real hash search space using adjustable fields in the Bitcoin block header. We also observe that the SHA-256 process can be clearly split into a part that stays the same for a block and a part that changes based on the nonce. By pre-computing the unchanged part on a classical digital computer and applying quantum search only to the nonce-dependent part, we reduce the quantum workload while keeping the chance of success very high. This avoids assuming SHA-256 is a fully quantum black-box oracle. Our method keeps SHA-256 as it is and designs the search with realistic assumptions, focusing on structure rather than replacing the hash function itself.**

***Index Terms*—Proof-of-Work, Bitcoin, Quantum Computing, SHA-256, Grover's Algorithm**

I. Introduction

SHA-256 is a standard cryptographic hash function widely used in security-critical applications, including digital signatures, transaction verification, and blockchain consensus mechanisms. In blockchain systems, SHA-256 plays a central role in block generation, Proof-of-Work (PoW) mining, and transaction integrity verification. Consequently, attacks on the SHA-2 family pose serious threats not only to the hash function itself but also to the security and reliability of PoW-based blockchain systems [1].

PoW mining searches for a block header whose hash satisfies a predefined difficulty target. This process is computationally expensive and consumes substantial computing resources. While the randomness and one-way property of cryptographic hash functions underpin system security, they also motivate research into alternative computational paradigms that may improve mining efficiency. In this context, recent advances in quantum computing have attracted attention as a potential means to accelerate hash-based PoW processes.

Grover's algorithm is a representative quantum algorithm that provides a quadratic speedup for unstructured search problems [2]. While a classical computer requires an average of $O(2^n)$ operations to find a value that satisfies a given condition, a quantum computer can reduce this complexity to $O(\sqrt{2^n})$. This result raises concerns regarding the pre-image resistance of cryptographic hash functions, and the implications of quantum search for symmetric cryptography have been widely studied [3]. Prior work has also investigated the impact of quantum algorithms on PoW-based mining systems, including Bitcoin [4].

However, existing quantum-based mining and hash attack approaches do not fully address a fundamental limitation of hash search. Hash-based mining is inherently probabilistic, and the existence of a valid solution is not guaranteed for a fixed search space and difficulty target [5]. Even with Grover's algorithm, the probability of finding a valid hash may remain strictly less than one. As a result, mining attempts can still fail despite the availability of quantum computational resources.

In this paper, we address this probabilistic limitation by proposing a structured expansion of the hash search space based on adjustable components of the Bitcoin block structure. We show that, under a given difficulty target, the existence of a valid solution can be guaranteed. Our approach goes beyond accelerating hash search and instead improves the theoretical stability and reliability of PoW-based mining in a quantum computing setting.

II. Background

A. Bitcoin Mining Process

Bitcoin is a decentralized blockchain system in which transactions are maintained without a centralized authority. Transactions are propagated through a peer-to-peer network and periodically grouped into blocks. These blocks are cryptographically linked in chronological order, forming a hash chain that ensures the integrity of the transaction history [1].

The Bitcoin network consists of users and validators, commonly referred to as miners. Users generate transactions and broadcast them to the network, while miners collect pending transactions and attempt to construct a new block. Once a miner generates a valid block, it is broadcast to the network, where other miners independently verify its validity before appending it to their local copy of the blockchain.

Each block contains a cryptographic reference to the previous block, together with a set of recent transactions. Modifying the contents of a block changes its hash value and invalidates

all subsequent blocks in the chain. This hash-based chaining mechanism provides tamper resistance, since altering past records would require recomputing the hashes of all following blocks.

To prevent trivial recomputation of block hashes, Bitcoin employs a Proof-of-Work (PoW) mechanism. In PoW-based mining, miners search for a nonce such that the hash of the block header satisfies a predefined difficulty condition. Specifically, the hash must begin with a certain number of leading zero bits, where the required number determines the mining difficulty. Because cryptographic hash functions behave pseudorandomly, the probability of finding a valid nonce decreases exponentially as the difficulty increases.

As a result, finding a valid nonce typically requires a large number of repeated hash evaluations, making block generation computationally expensive. This cost discourages malicious attempts to modify previously confirmed blocks, since an adversary would need to redo the PoW for the altered block and all subsequent blocks. The process of searching for such a nonce and generating a valid block is referred to as Bitcoin mining [5].

B. Bitcoin Block Header Structure

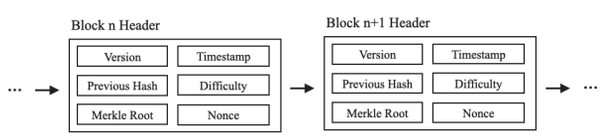


Fig. 1. Block header fields in Bitcoin

In Bitcoin, each block is identified by its block header, which summarizes the block and is used to verify Proof-of-Work (PoW). Figure 1 shows two consecutive block headers and their cryptographic linkage.

A Bitcoin block header consists of six fields in a fixed order: version (32 bits), previous block hash (256 bits), Merkle root (256 bits), timestamp (32 bits), difficulty target in compact encoding (32 bits), and nonce (32 bits). These fields are concatenated and hashed using double SHA-256 during mining [1].

The previous block hash links each block to its predecessor, forming a hash chain. The timestamp and difficulty encode timing and difficulty information at block creation. The version specifies the consensus rules. The nonce is the primary variable that miners iterate during PoW.

The Merkle root is a cryptographic commitment to all transactions in the block. Transactions are arranged in a Merkle tree, where leaf nodes are transaction hashes, and internal nodes are computed by hashing the concatenation of two child nodes. As shown in Figure 2, the root uniquely summarizes the transaction set and is included in the block header.

The nonce provides only a 32-bit search space, which can be insufficient under high difficulty. To expand the effective search space, miners also vary the coinbase transaction, which

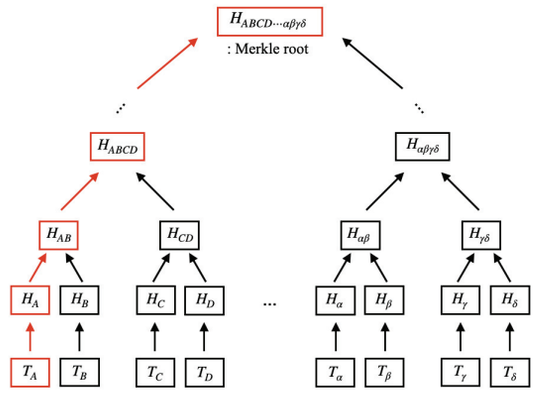


Fig. 2. Merkle tree Structure

is the first transaction in a block and can be modified by the miner. In practice, miners embed an additional counter, called an extra nonce, into the coinbase transaction. In our setting, we treat the extra nonce as a 12-bit value placed in the coinbase data.

During mining, miners update the extra nonce and recompute the coinbase transaction hash. This change affects only the hashes along the Merkle path from the modified leaf to the root (highlighted in red in Figure 2). Therefore, miners update only the nodes on this path, rather than rebuilding the entire Merkle tree.

After updating the Merkle root, the miners evaluate the modified block header using double SHA-256. By iterating over extra nonce values and recomputing the corresponding Merkle root, miners expand the search space beyond the 32-bit nonce field. This structured mechanism forms the basis for the search space expansion strategy used in this work.

C. Overview of the SHA-256 Algorithm

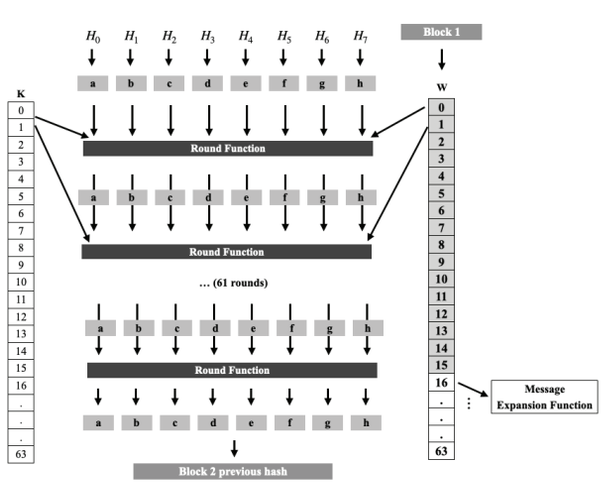


Fig. 3. SHA-256 Compression Function

Figure 3 presents a high-level overview of the SHA-256 compression function as applied to the Bitcoin block header.

SHA-256 processes the input message in fixed-size blocks of 512 bits and produces a 256-bit hash value through an iterative compression procedure.

Each compression block is initialized with eight 32-bit working variables (a, b, c, d, e, f, g, h). These variables are set to predefined initial constants for the first block and to the hash output of the previous block for subsequent blocks. For each message block, a message schedule $W_0, \ldots, W_{63}$ is constructed. The first 16 words are directly obtained from the input message, while the remaining words are generated using a message expansion function. For example,

$$W_{16} = \sigma_1(W_{14}) + W_9 + \sigma_0(W_1) + W_0$$

The compression function consists of 64 rounds. In each round, the working variables are updated using a round-dependent constant, a word from the message schedule, and a set of fixed Boolean and rotation functions, including the choice function $\mathrm{Ch}(\cdot)$, the majority function $\mathrm{Maj}(\cdot)$, and the Σ rotation functions. After all rounds are completed, the final working variables are added to the initial state to produce the hash output of the block, which serves as the input state for the next block.

In Bitcoin mining, the block header is padded and processed across two SHA-256 message blocks. The first block contains the version field, the previous block hash, and a portion of the Merkle root. The second block includes the remaining Merkle root bits, the timestamp, the difficulty field, the nonce, and padding. Since each message word W_i consists of 32 bits, the words corresponding to fixed header fields in the first block remain constant. Specifically, W_0 through W_8 do not change during the mining process.

As a result, the compression steps associated with these fixed message words can be deterministically evaluated in advance using classical computation. Only the remaining computation, which depends on variable fields such as the nonce and the modifiable portion of the Merkle root, contributes to the exploration of the hash search space. This observation enables a hybrid strategy in which the invariant part of the SHA-256 computation is handled classically, while the remaining variable-dependent computation is delegated to a quantum processor.

III. Implementation

Proof-of-Work mining finds a valid Bitcoin block header by testing a vast number of SHA-256 hashes. Because each hash trial has an extremely small chance of success, miners enlarge the practical search space so that the overall probability of finding a valid hash becomes very high.

a) A. Digital Computer: In Digital mining, a classical digital computer iterates over all combinations of the timestamp window, the extra nonce inside the coinbase transaction, and the 32-bit nonce field in the header. For each candidate, miners update the necessary nodes along the Merkle path, rebuild the Merkle root, concatenate the full block header, apply padding, and compute double SHA-256 from start to finish. If the timestamp range is 12 bits and both the extra nonce and nonce fields are 32 bits, the effective search space becomes 76 bits. The average classical cost of evaluating all hashes is :

$$\mathcal{O}(2^{12} \cdot 2^{32} \cdot 2^{32}) = \mathcal{O}(2^{76})$$

This space can make the probability of success close to one, but it requires recomputing the full digital hash every time, including large parts of logic that do not change across trials. In practice, a large fraction of the per-trial work is repeated even when only a small header field changes.

b) B. Quantum + Digital Computers: Quantum-assisted mining keeps timestamp and extra nonce loops on a classical digital computer and applies Grover search only to the 32-bit nonce space. Grover search reduces the 32-bit nonce trials to $\mathcal{O}(2^{16})$, giving a quadratic speedup for that field. The remaining classical loops still cost $\mathcal{O}(2^{12})$ for timestamp and $\mathcal{O}(2^{32})$ for extra nonce. The total expected complexity becomes :

$$\mathcal{O}(2^{12} \cdot 2^{32} \cdot 2^{16}) = \mathcal{O}(2^{44} \cdot 2^{16}) = \mathcal{O}(2^{60})$$

Even though the nonce search is faster, this method typically treats SHA-256 as a black-box oracle, meaning that the quantum oracle evaluates the full double SHA-256 hash computation without reducing repeated fixed operations. As a result, the oracle depth remains large because invariant compression steps are still executed inside the quantum circuit.

c) C. Proposed Structure: Our approach follows the structure shown in the right panel of Figure 4. We observe that in the first SHA-256 message block, words w_0 through w_8 depend only on fixed header fields version and previous hash, so the 64 compression rounds applied to these words always produce the same intermediate result for a given block. We pre-compute these 64 rounds once on a classical digital computer and store the intermediate chaining value, removing all invariant computation from the quantum oracle.

The quantum processor jointly explores nonce, timestamp, and extra nonce, but the quantum oracle evaluates only the variable-dependent portion of Block 1 from w_9 to w_{63}, and the full second compression block including timestamp, difficulty, nonce, and padding.

Since Grover search is applied only to the nonce-dependent computation and the oracle excludes invariant rounds, the average quantum cost is reduced to :

$$\mathcal{O}(2^{38})$$

Quantum search now focuses only on the part that directly affects valid hash discovery, while preserving the ability to enlarge the overall search space using digital pre-computation and variable fields, without modifying SHA-256 itself. This separation also makes the quantum oracle simpler by excluding computation that is guaranteed to be constant across trials.

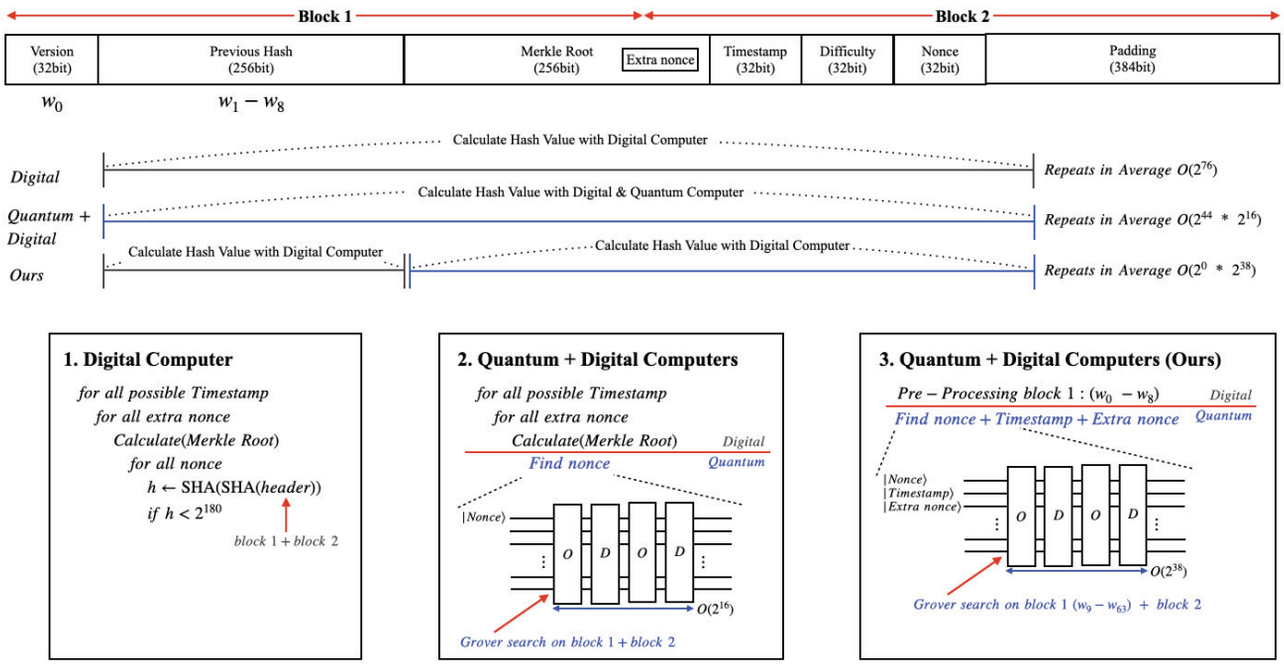


Fig. 4. Comparison of Mining Strategy

IV. Conclusion and Future Work

In this paper, we examine the probabilistic limitations of hash-based Proof-of-Work mining in a quantum computing setting and identify structural conditions under which the existence of a valid solution can be guaranteed. By focusing on adjustable components of the Bitcoin block header, we showed that the hash search space can be systematically expanded beyond the conventional nonce field without altering the underlying hash function or consensus rules.

We further demonstrated that the SHA-256 computation employed in Bitcoin mining admits a natural separation between invariant and variable components at the round-function level. This structural property enables a hybrid computation strategy in which invariant components are evaluated classically, while only variable-dependent components are delegated to a quantum processor, thereby reducing both the effective quantum search space and the circuit depth. As a result, the proposed approach achieves an asymptotic improvement in computational complexity compared to existing quantum-assisted mining strategies. A detailed analysis of quantum circuit depth and gate counts is left for future work.

Acknowledgment

This research was supported by the Regional Innovation System & Education(RISE) program through the Gangwon RISE Center, funded by the Ministry of Education(MOE) and the Gangwon State(G.S.), Republic of Korea.(2025-RISE-10-005)

References

[1] S. Nakamoto, Bitcoin: A peer-to-peer electronic cash system, https://bitcoin.org/bitcoin.pdf (2008).
[2] L. K. Grover, A fast quantum mechanical algorithm for database search, in: Proceedings of the twenty-eighth annual ACM symposium on Theory of computing, 1996, pp. 212–219.
[3] M. Grassl, B. Langenberg, M. Roetteler, R. Steinwandt, Applying grover's algorithm to aes: quantum resource estimates, in: International Workshop on Post-Quantum Cryptography, Springer, 2016, pp. 29–43.
[4] D. Aggarwal, G. K. Brennen, T. Lee, M. Santha, M. Tomamichel, Quantum attacks on bitcoin, and how to protect against them, arXiv preprint arXiv:1710.10377 (2017).
[5] J. Bonneau, A. Miller, J. Clark, A. Narayanan, J. A. Kroll, E. W. Felten, Sok: Research perspectives and challenges for bitcoin and cryptocurrencies, in: 2015 IEEE symposium on security and privacy, IEEE, 2015, pp. 104–121.

Real-Time Object Distance Estimation and LLM-Based Surrounding Scene Description

Munho Kang *, 2nd Sihoon Kim †, 3rd Misbah Bibi ‡, 4th DoHyeun Kim §

** Department of Computer Engineering, Jeju National University, Jeju, South Korea*

† Department of Computer Engineering, Jeju National University, Jeju, South Korea

‡ Department of Computer Engineering, Jeju National University, Jeju, South Korea

§ Department of Computer Engineering, Jeju National University, Jeju, South Korea

Emails: rkdansgh05@naver.com, fpem617@gmail.com, misbahbibi@stu.jejunu.ac.kr, kimdh@jejunu.ac.kr

Abstract—**This paper proposes a novel dual-mode assistive system designed to enhance pedestrian safety and situational awareness for visually impaired users. Unlike conventional single function assistive devices, the proposed system integrates two complementary operation modes: a real-time hazard recognition mode for dynamic walking scenarios and an environmental understanding mode for stationary contexts. In the hazard recognition mode, video streams captured by a monocular camera are processed in real time using a YOLO based deep learning framework for object detection. A perspective-based distance estimation method is introduced to accurately compute object proximity without requiring additional depth sensors. To further assess potential risks, an object tracking algorithm determines whether detected objects are approaching the user, and a four-level risk classification scheme is developed by jointly considering distance and motion information, enabling timely and context aware hazard alerts. Experimental evaluations demonstrate that the proposed system achieves an average distance estimation error of less than 0.5 m and reliably generates hazard notifications at 2s intervals. In the environmental understanding mode, a multimodal large language model generates natural-language descriptions of surrounding scenes from a single image, supporting comprehensive environmental perception during stationary situations. The proposed dual-mode architecture and integrated risk assessment strategy represent a significant advancement over existing assistive systems, offering improved adaptability, safety, and usability for visually impaired pedestrians.**

Keywords—*Assistive technology, Object Detection, Distance Estimation, Multi-Modal LLM*

I. INTRODUCTION

In modern society, the safe mobility of visually impaired individuals is an important social task, and especially in urban environments where cars, bicycles, and pedestrians are mixed, there is a high possibility of accidents if surrounding hazards are not properly recognized. According to the World Health Organization (WHO), there are more than 250 million visually impaired people worldwide, and they are continuously exposed to danger because they do not obtain sufficient environmental information during daily walking [1]. Traditional assistive devices such as white canes and guide dogs play an important role in assisting walking, but they have structural limitations in providing quantitative and situational information such as the type, distance, and movement of objects.

To compensate for these limitations, research and commercial services on visually impaired assistive systems using cameras and artificial intelligence technology have recently emerged. Typically, commercial systems using camera-based recognition technology support the environment recognition of visually impaired people by providing object or text information by voice, but there are limitations in the function of quantitatively determining the distance relationship between the user and the object or the urgency of dangerous situations [2][3]. In addition, some studies have proposed methods to improve distance recognition accuracy by using additional sensors or assuming a specific environment, but this approach has limitations of increased system complexity or limited use environment [4][5].

Therefore, a walking assistance system for the visually impaired needs to have a structure that can recognize hazards that may occur in the user's movement situation in real time and intuitively convey the urgency of the situation to the user, beyond simple object recognition functions. Furthermore, a method of providing information that can comprehensively understand the surrounding environment even in a stationary state rather than during movement is required. Accordingly, this paper analyzes the design and applicability of a dual-mode assistive system that separates and provides real-time risk recognition and static environmental understanding to support walking safety for the visually impaired, and experimentally examines the usability in a monocular camera-based environment.

II. RELATED WORK

Walking assistance systems for the visually impaired can be mainly divided into object recognition technology, distance estimation techniques, object tracking technology, and auxiliary system application research using them. First, in the object recognition field, deep learning-based object detection technology has been rapidly developed, and 2-stage methods such as Fast R-CNN provide high recognition accuracy, but there are limitations in real-time processing due to large computational costs [6]. On the other hand, 1-stage methods such as YOLO and SSD have the advantage of high usability in mobile environments because they can be processed in real time by simultaneously predicting the object location and class through a single neural network [7][8]. Recently, lightweight models such as YOLOv8 have been proposed, and the scope of object detection technology is expanding even in mobile and embedded environments [9].

Monocular camera-based distance estimation is a key element in visually impaired assistive systems, but there are still many challenges to be solved. Deep learning-based monocular depth estimation techniques can estimate relative depth information from images, but have the limitation that it is difficult to provide absolute distance values [10]. In contrast, empirical distance estimation methods using the object size ratio

on the screen are simple to implement, but have low accuracy because they do not fully reflect the actual size of objects or camera characteristics [11]. A more precise distance calculation approach using camera calibration and geometric models has been proposed, and Zhang's camera calibration technique is used as a representative method to enable physics-based distance estimation in a monocular camera environment [12].

In order to determine the movement and approach status of objects, there are limitations only with single frame-based recognition, and multi-object tracking techniques are used to compensate for this. SORT is widely used in real-time environments as a method of efficiently connecting objects between frames using Kalman filters and Hungarian algorithms [13], and DeepSORT is a technique that improves tracking stability by combining appearance features [14]. These tracking techniques can be used to determine the approach or movement pattern during walking by analyzing the temporal changes of objects.

Many assistive system application studies and commercial cases for the visually impaired have also been reported. Commercial systems such as OrCam MyEye and Microsoft Seeing AI support the environment recognition of the visually impaired by providing object or text information by voice using camera-based recognition technology [2][3]. However, these systems have limitations in the function of quantitatively providing distance information between users and objects or systematically determining the urgency of real-time dangerous situations. Meanwhile, assistive systems using RGB-D cameras or hybrid methods combining vision and ultrasonic sensors can improve distance recognition accuracy, but there are limitations in outdoor environments or general use due to additional hardware requirements [4][5].

In summary, existing studies have shown meaningful results in terms of object detection and environment recognition, but there is still room for improvement in terms of quantitative distance recognition in a monocular camera environment, temporal judgment on object approach, and integrated walking assistance systems that intuitively convey dangerous situations.

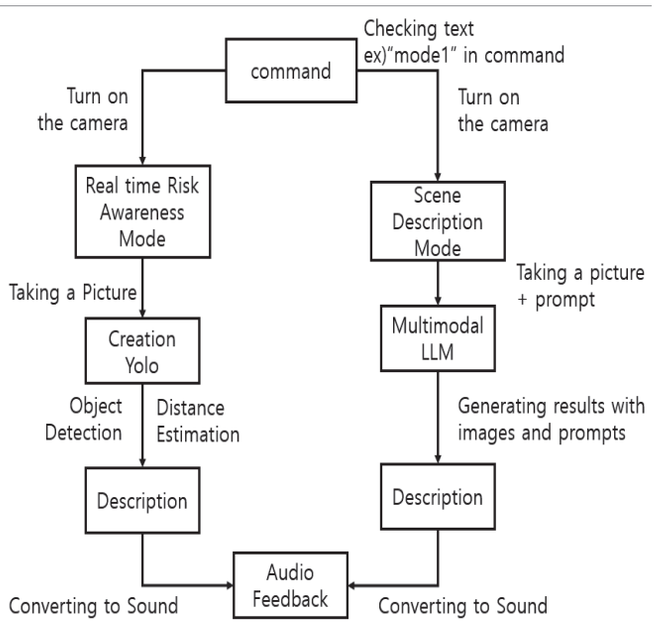


Fig. 1. Overall structure and processing flow of the dual-mode system for visually impaired pedestrian assistance

III. REAL-TIME DISTANCE RECOGNITION AND LLM-BASED SURROUNDING SCENE DESCRIPTION DESIGN

A. System Overview and Dual-Mode Structure

In this paper, we consider a pedestrian assistance scenario for visually impaired users and present the design architecture of a dual-mode assistive system that integrates object detection, distance estimation, object tracking, and voice-based interaction technologies. Figure 1 illustrates the overall architecture of the proposed system. The system is designed with two distinct operational pathways: a real-time hazard recognition pathway that processes video input from a monocular camera during user movement, and an environmental understanding pathway that provides comprehensive scene awareness when the user is stationary. This dual-mode architecture is intentionally adopted to balance the requirements of real-time processing stability and rich environmental information delivery. By clearly separating the functional roles and usage contexts of each pathway, the proposed design effectively manages system complexity while enhancing robustness and practical applicability in real-world walking environments.

B. YOLO-Based Object Detection Module

In this system, considering the real-time processing conditions required in the walking assistance scenario during movement, YOLOv8n is used as a lightweight object detection model that can estimate the location and class of objects in a single step.

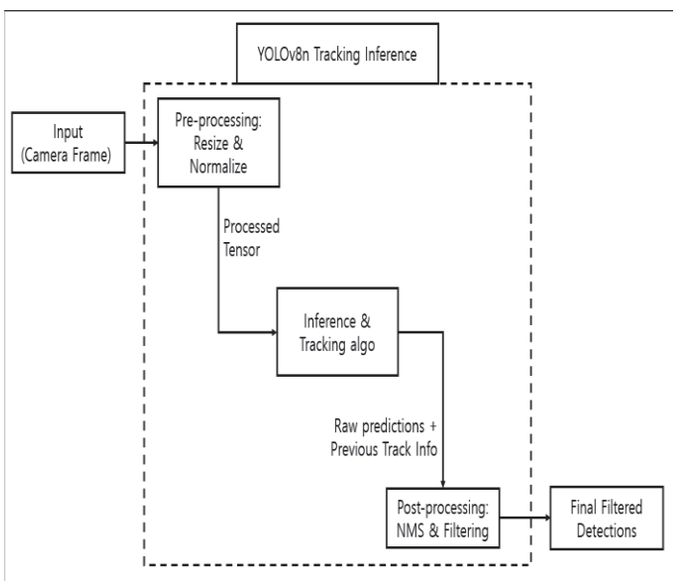


Fig. 2. Object detection processing flow using YOLOv8n from monocular camera images

Figure 2 illustrates the object detection and tracking pipeline, in which image frames captured by a monocular camera are processed using the YOLOv8n model to generate detection outputs, including object location, class label, confidence score, and tracking identifier. The model is employed in its pre-trained form using the COCO dataset, enabling robust recognition of common objects such as pedestrians and vehicles. To reduce false-positive detections, a confidence threshold is applied during the detection stage. In addition, an integrated tracking mechanism assigns consistent identifiers to detected objects across consecutive frames, allowing temporal changes in object motion to be analyzed in subsequent processing stages. Based on the detection results, objects are further categorized into high-mobility objects and static obstacles, which are treated

differently in the risk assessment stage according to their motion characteristics and potential hazard levels.

C. Perspective-Based Distance Estimation Module

Since the distance between the object and the user is difficult to measure directly in a monocular camera environment, this system estimates the distance to the object using the perspective transformation relationship based on the existing pinhole camera model.

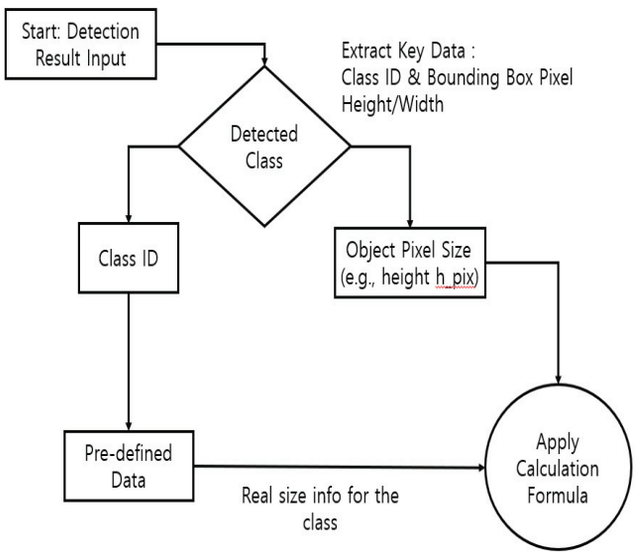


Fig. 3. Perspective-based distance calculation flowchart using object size and camera focal length

Figure 3 illustrates the overall processing flow for estimating object distance using a perspective-based geometric approach derived from object detection results. The distance is computed based on the geometric relationship among the object's projected size on the image plane, its average real-world size, and the camera's focal length, enabling distance estimation in metric units using a single monocular camera. Specifically, the distance estimation exploits the proportional relationship between the height of the detected object's bounding box in the image and its actual physical height. The corresponding distance calculation formula is presented in Equation (1), where d denotes the estimated distance (m) from the camera to the object, H represents the actual height of the object (cm), f is the camera focal length, and h is the height of the detected bounding box in pixel units. For consistency in units, the actual object height H is converted from centimeters to meters by dividing by 100 prior to distance computation.

$$d = (f \times H) / (h \times 100) \qquad (1)$$

The actual size of the object is used as the average value for each class based on public dataset statistics, and the calculated distance value is limited to a certain range to prevent unrealistic results. This method was included in the system design as a practical approach to use distance information without additional sensors in a monocular camera environment.

D. Object Tracking-Based Approach Judgment

Since it is difficult to stably determine whether the object is approaching the user with only distance information at a single time point, this system uses object tracking results to analyze changes over time. For each object, the distance change and screen size change of the recent frame section are stored and accumulated, and by comparing them, it is configured to determine the approach.

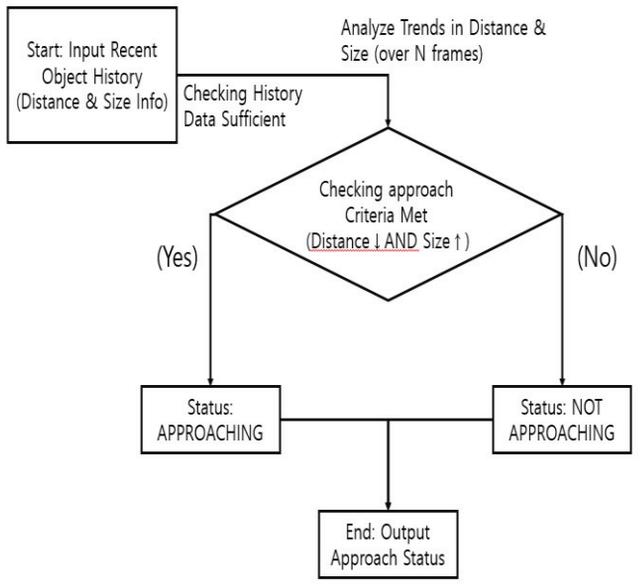


Fig. 4. Approach judgment flowchart based on temporal change using object tracking results

Figure 4 shows the entire processing flow of analyzing the distance and screen size change of the recent frame section based on the object tracking result to determine the approach of the object. The approach judgment is made based on multiple conditions such as distance reduction trend, bounding box size increase trend, and rapid distance change in a short period of time, and through this multi-condition judgment, it was intended to reduce the possibility of misjudgment due to temporary errors or camera shake. This process was designed as an auxiliary judgment step to reflect the movement characteristics of the object in the risk judgment.

E. Risk Level Classification

Risk assessment is performed by comprehensively considering distance information to objects and approach judgment results. In this system, the risk level is classified step by step according to the distance section considering the average moving speed and reaction time of pedestrians, and for approaching objects, it is configured to judge as a relatively high-risk level at the same distance. In addition, considering the situation where multiple objects are detected at the same time, the priority of objects is calculated based on the risk level and distance information, and information overload is prevented by limiting the number of risk factors delivered to the user. This risk classification is a design element to enable the voice guidance delivered to the user to be effectively used in the actual walking situation.

F. Adaptive Camera Calibration

The focal length of the camera can vary depending on hardware characteristics or setting changes, which directly affects distance estimation results. In this system, an adaptive calibration procedure based on user input was designed considering this change. When the user provides the actual distance to a specific object, the information is stored as a sample, and when a certain number of samples are secured, the focal length value is corrected using them. The corrected focal length is commonly applied to the subsequent distance estimation process, and is used to maintain the consistency of distance estimation results even in various camera environments. This process is a complementary design element to increase the

practical usability of the system without complicated calibration procedures.

G. Surrounding Scene Description Mode

The surrounding situation description mode is designed to support comprehensive environmental understanding when the user is in a stationary state and is implemented as an independent processing pathway, separate from the real-time hazard recognition flow. In this mode, a single captured image is used as input to generate a natural-language description that conveys the spatial arrangement of objects, their relative relationships, and the overall contextual characteristics of the scene. A multimodal large language model is employed for this purpose; however, its role is strictly limited to interpreting and verbalizing pre-analyzed scene information. The model does not participate in object detection, distance estimation, or risk assessment processes. The generated textual description is subsequently converted into speech and delivered to the user via a voice output interface. This functionality enhances the user's situational awareness by enabling perception of the surrounding environment in scenarios beyond active movement, thereby complementing the real-time assistance provided during walking.

IV. Implementation Details

A. Implementation Environment and Overall Structure

This chapter explains the implementation environment and overall structure of the dual-mode assistive system designed above. The system was implemented in the Python environment and was configured to perform real-time risk recognition and static environment description functions based on monocular camera input. The specific implementation environment and execution settings are summarized in Table 1.

Table 1: System implementation environment

Categories	Details
Language	Python 3.9
Image Library	OpenCV 4.5.0
Detection	YOLOv8n
Camera	Monocular 640×480
Analysis Cycle	Every 2 seconds

B. Multithreading Implementation

The real-time risk recognition mode was implemented as a multithreading structure and was configured to operate in a dual structure of the main thread for voice command recognition and the analysis thread that integrates video acquisition and object analysis. The analysis thread acquires continuous frames from the camera and sequentially performs object detection, distance estimation, approach judgment, and risk assessment processes by receiving the latest frame as input at regular intervals.

C. Object Detection and Tracking

In the object detection stage, the tracking function of YOLOv8n was activated to assign a consistent identifier to the same object between frames.

Figure 5 shows the processing flow of assigning identifiers to the same object between frames using the tracking function of YOLOv8n and temporally managing the class and bounding box information of the object.

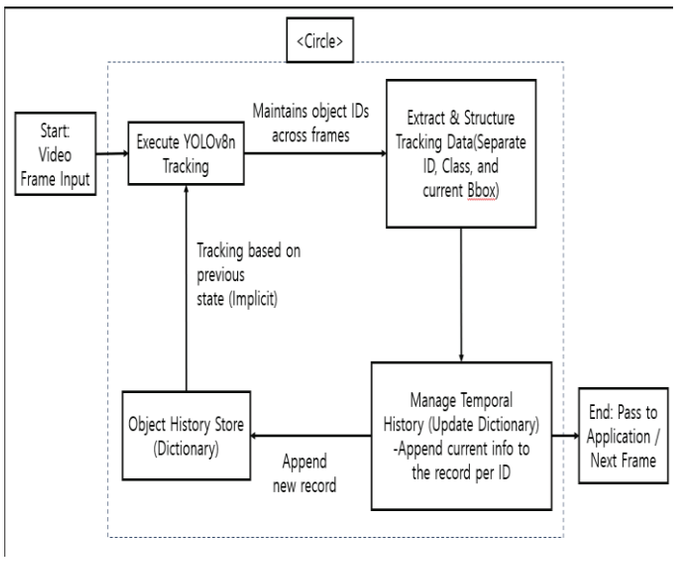


Fig. 5. YOLOn-based detection and tracking flow

V. EXPERIMENTAL RESULTS

The experimental environment is shown in Table II.

Table 2: Experimental environment

Item	Content
Environment	Indoor 20m corridor
Processor	Intel i7-9700K
Memory	16 GB RAM
OS	Windows 11
Camera	Logitech C920

A. Distance Estimation Results

Table III shows the distance accuracy results. The system showed stable performance in short-range with MAE of 0.32 m and average error rate of 8.80%.

Table 3: Distance estimation results detail

Metric	Result
Short-range (<5m)	Error <0.5 m
MAE	0.32 m
Avg Error Rate	8.80%

B. Object Detection Performance

Table IV shows object detection performance. Precision and recall were 93.2% and 89.7%, with F1 score of 91.4%. Processing time was 0.18 sec/frame (~5.5 FPS).

Table 4: Object detection performance analysis

Metric	Result
Precision	93.20%
Recall	89.70%
F1 Score	91.40%
Processing	0.18 s/frame

C. Real-time performance Analysis

Real-time system operation is implemented using a multithreaded architecture to ensure efficient and responsive processing. The analysis thread continuously captures video frames and performs object analysis, executing YOLO-based object detection at fixed 2s intervals while updating and displaying the analysis results for each frame. Concurrently, the main thread is responsible for monitoring voice command input and managing system control. Upon receiving a termination command from the user, the main thread issues a termination signal to the analysis thread, allowing for safe and orderly system shutdown.

VI. CONCLUSION

This paper presented the design and implementation of a dual-mode assistive system for visually impaired pedestrians based on monocular camera input. The proposed system adopts a dual-path architecture that distinctly separates real-time hazard recognition during movement from comprehensive environmental understanding in stationary situations. This design enables both timely risk awareness and informative scene perception while maintaining system stability and practical usability.

Experimental results demonstrated that the proposed perspective-based distance estimation method achieves stable performance in short-range scenarios, with average estimation errors of less than 0.5 m. Furthermore, the approach assessment and four-level risk classification, based on yolo-based object detection and object tracking, operated reliably under a 2-s analysis cycle. The surrounding situation description mode effectively supported environmental awareness by providing natural-language descriptions of the scene.

Despite these results, the system has several limitations. Distance estimation accuracy may be affected by assumptions regarding average object dimensions and variations in camera parameters or environmental conditions. In addition, experimental evaluations were conducted under limited scenarios, and further validation is required in diverse outdoor environments and complex real-world walking conditions.

Future work will focus on refining object size models across multiple classes, improving distance estimation robustness for longer ranges, and conducting extensive evaluations in more varied and dynamic pedestrian environments to enhance system generalizability and reliability.

ACKNOWLEDGEMENT

Following are results of a study on the " Convergence and Open Sharing System" Project, supported by the Ministry of Education and Nation Research Foundation of Korea.

REFERENCES

[1] World Health Organization, "World report on vision", WHO, 2019.

[2] OrCam Technologies, "OrCam MyEye 2", 2021.

[3] Microsoft Corporation, "Seeing AI: Talking Camera for the Blind", 2021.

[4] L. Wang et al., "RGB-D based Obstacle Detection and Navigation for the Visually Impaired" IEEE Int. Conf. Robot. Autom., pp. 5149-5154, 2019.

[5] H. Kim et al., "A Hybrid Obstacle Detection System using Vision and Ultrasonic Sensors", J. Assist. Technol., vol. 14, no. 2, pp. 85-92, 2020.

[6] R. Girshick, "Fast R-CNN," IEEE Int. Conf. Comput. Vis., pp. 1440-1448, 2015.

[7] J. Redmon et al., "You Only Look Once: Unified, Real-Time Object Detection", IEEE Conf. Comput. Vis. Pattern Recognit., pp. 779-788, 2016.

[8] W. Liu et al., "SSD: Single Shot MultiBox Detector", Eur. Conf. Comput. Vis., pp. 21-37, 2016.

[9] Ultralytics, "YOLOv8 Documentation", 2023.

[10] R. Ranftl et al., "Towards Robust Monocular Depth Estimation", IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 3, pp. 1623-1637, 2022

[11] S. Park et al., "Vision-based Distance Estimation for the Visually Impaired", Sensors, vol. 20, no. 22, 2020.

[12] Z. Zhang, "A flexible new technique for camera calibration", IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 11, pp. 1330-1334, 2000.

[13] A. Bewley et al., "Simple online and realtime tracking", IEEE Int. Conf. Image Process., pp. 3464-3468, 2016.

[14] N. Wojke et al., "Simple online and realtime tracking with a deep association metric", IEEE Int. Conf. Image Process., pp. 3645-3649, 2017.

Design Challenges of On-Device AI for Maritime Communication Systems

1st Shrutika Sinha *, 2nd Soo-Hyun Park †
* *Dept. of Financial Information Security Kookmin University Seoul, South Korea*
† *School of Computer Science Kookmin University Seoul, South Korea*

***Abstract*—The maritime communication systems are subjected to severe environmental factors, disconnected connection, and limited resources, which restricts the feasibility of cloud-aided artificial intelligence. With the development of wireless networks to 6G and non-terrestrial network (NTN) configurations, intelligence should be more and more applied on embedded maritime devices, including buoys, autonomous control vehicles, and relay nodes. The design limitations that control the implementation of on-board AI in the maritime setups are however, not adequately researched.This paper provides a focused review of system-level constraints affecting on-device AI for maritime communications, including hardware and energy limitations, connectivity and latency challenges, and model robustness in non-stationary environments.**

***Index Terms*—Maritime communication systems,Non-terrestrial networks (NTN), On-device, Resource-constrained systems**

I. Introduction

The communication systems at sea form a core structure of modern operations in the ocean and are the basis of a range of applications, including: monitoring vessels, offshore measurements, and deploying autonomous vehicles on the surface and underwater [1]. In comparison to terrestrial wireless networks, maritime conditions can be described as having a lack of fixed infrastructure, extremely dynamic radio propagation, and regular connectivity perturbations [2]. The sea creatures and sea situations reflecting, movement of vessels, atmospheric disturbances and sweeping distances of communication, all contribute to the development of non-stationary channel dynamics and as such, undermines the trustworthiness of wireless operations [3]. With the shift of maritime networks to the data-driven and autonomous paradigms, the need to employ intelligent communication control has also grown accordingly [4].

Current studies in artificial intelligence have demonstrated that there are high possibilities of improving wireless communication capabilities through adaptive channel selection, link-quality prediction, and resource allocation [5]. However, the current AI-driven communication systems are implicitly premised on the presence of stable backhaul connectivity and sufficient computational power, thus allowing inference or training to be offloaded to cloud or edge servers [6]. In maritime situations, such assumptions are hardly true. Satellite connections are associated with high latency and costs and intermittent connectivity compromises sustained cloud availability [7]. As a result, dependence on cloud-based intelligence could not be considered a solid option in various situations on the sea [8].

The solution to overcome these limitations is to move intelligence more and more to be performed directly on embedded maritime communication devices such as buoys, autonomous surface vehicles, autonomous underwater vehicles and shipboard gateways [9]. On-device AI, as commonly known as this paradigm, supports local decision-making regardless of external calculating or consistent connection [10]. Even though on-board AI has been explored in the academic community in the context of the terrestrial Internet-of-Things and mobile computing, its use in maritime communication is poorly studied, especially in terms of system-level design [11].

On-device AI implementation in maritime communication systems has a unique range of restrictions that makes it different compared to cloud-based and edge-assisted paradigms [12]. Embedded maritime instrumentation is capable of working in highly restricted energy envelopes, limited computing capabilities, and adverse physical conditions [13]. These systems simultaneously have to deal with heterogeneous as well as temporally intermittent attributes such as link, where a high signal strength does not imply low latency and reliable throughput [14]. In addition, AI models deployed in maritime situations have to be resilient to sharp non -stationarity of the environment and domain shift since the conditions of the operational situation usually differ significantly compared to those that were observed during the development stage [15]. Despite growing interest in AI-driven maritime networking, existing literature primarily focuses on application-layer intelligence, centralized edge computing, or algorithmic performance under idealized assumptions [16]. System level aspects of prevention that essentially determine the viability, dependability, and autonomous functioning of on-device AI in maritime communication systems are often implicitly assumed or overlooked [17]. This gap limits the practical deployment of intelligent maritime communication nodes in real-world environments. Figure 1 is an illustration of intelligence placement in maritime communication systems.

This paper addresses this gap by providing a focused review and design-oriented analysis of the key constraints affecting on-device AI deployment in maritime communication systems. The discussion is structured around three fundamental dimensions: (i) hardware and energy constraints inherent to

embedded maritime devices, (ii) connectivity and latency constraints arising from heterogeneous and intermittent links, and (iii) model-level challenges associated with environmental non-stationarity and domain shift.

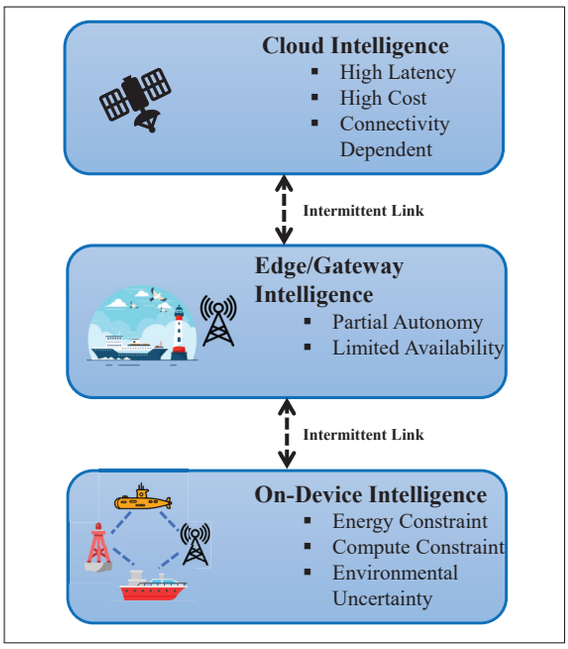


Fig. 1. Conceptual illustration of intelligence placement in maritime communication systems. Unlike cloud- or edge-assisted intelligence, on-device AI operates directly on embedded maritime nodes under strict hardware, energy, connectivity, and environmental constraints, enabling autonomous decision-making without reliance on persistent backhaul connectivity.

II. Maritime Communication Environment and System Context

The Maritime communication environment has some fundamental differences compared with the terrestrial wireless system based on the aspects of propagation, availability of infrastructure and operational limitations [18]. Compared to the land-based networks, where a dense deployment of the base-stations, as well as relatively predictable channel dynamics, is exploited, maritime communication systems have to operate at open water, where infrastructure is sparse, communication ranges are very long, and the common conditions in the atmosphere are highly dynamic [19]. These factors collectively create a challenging environment for reliable wireless connectivity and intelligent communication control.

A. Characteristics of Maritime Wireless Channels

The sea propagation of radios is heavily affected by the specular reflection and wave movement, the atmospheric factors and changes in the height of the antenna [20]. The Sea-surface reflections tend to induce prevailing two or multi-ray propagations and, as a result, there are quick shifts of the received signal strength even when line-of-sight is present [21]. Tiny variations in the orientation of the vessels, the waves heights or the location of the antennae may lead to constructive or destructive interferences, leading to very non-stationary channel behaviour [22]. Also, other atmospheric effects like humidity gradients and evaporation ducts further distort signal propagation especially at long distances [23].

Consequently, the wireless channels in the ocean are often temporally unstable, which cannot be satisfactorily modeled using a static or slowly time-varying channel model typically adopted in the terrestrial type of networks [24]. Stable throughput and low latency are not always high signal strength and the simple signal-based decision mechanism is not easily applicable [25]. These properties put high demands on the communication system that needs to be able to survive in highly changing conditions with minimal sensing and computation facilities.

B. Sparse Infrastructure and Heterogeneous Connectivity

The maritime networks are often based on a heterogeneous collection of communication technologies, i.e. Wi-Fi (near shore), cellular networks such as LTE, LTE-M (coastal areas), low-power wide-area networks in telemetry, and satellite connection in offshore areas [26]. As opposed to terrestrial networks, where handover is done between densely deployed cellular networks, maritime nodes can be subjected to long periods of isolation and short durations of connectivity [27]. Satellite connections are also wide coverage but involve high latency and are costly to facilitate communications hence are not suitable in frequent control signalling. Centralized coordination and cloud-assisted intelligence are not feasible in this heterogeneous and sporadic connectivity environment. Local decision-making is crucial for preserving communication dependability and system stability since communication nodes must frequently function independently for long stretches of time [28].

C. Embedded Maritime Communication Nodes

The physical platforms of the maritime communication systems further differentiate this field of operation compared to the conventional wireless networks. Examples of nodes include buoys, autonomous surface and underwater vehicles, and small relay platforms with limited power and computational resources and subjected to the extremes of many conditions, such as corrosion by salt, vibration, and temperature change [29]. These platforms are also heavily oriented at sensing and communication as opposed to server-class computation. Consequently, maritime node-based intelligence needs to operate under tight resource-budgets and be resilient to uncertainty in the environment and hardware limitations, which directly affect the facility and location where AI can be added to the network [30].

III. On-Device AI Design Constraints in Maritime Communication Systems

The application of artificial intelligence directly to maritime communication nodes presents a series of interrelated constraints that can be used to distinguish with fundamental differences between on-device and cloud-based or edge-assisted

intelligent models. These are limitations which are a result of the hardware constraints combined with the connectivity attributes, as well as the environmental uncertainty of maritime operations. These limitations form the basis of the design of AI-enabled communication systems that are practically and reliably applicable in actual marine settings. Figure 2 shows the constraints.

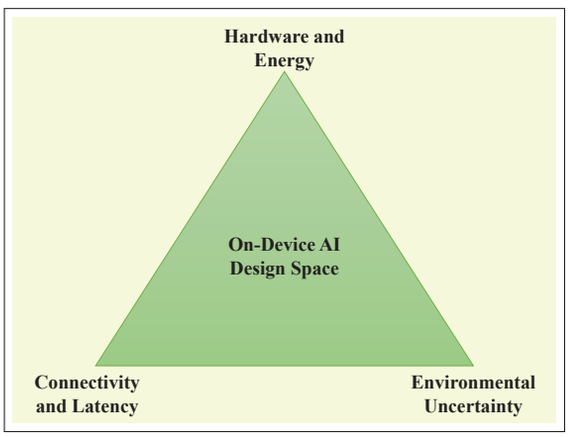


Fig. 2. Constraint Interaction in On-Device Maritime AI

A. Hardware and Energy Constraints

Autonomous platforms as well as buoys and relay platforms are marine communication nodes that are usually subject to severe power and computation constraints [31]. Most of these platforms are battery powered or utilise energy harvesting systems and therefore need long operational life and few maintenance requirements [32]. Consequently, AI inference has to be conducted with very constrained energy limits, in which the overall mission duration is directly influenced by the excess of computation [33]. Besides power limitations, embedded maritime systems also usually do not have special accelerators, and have to use general-purpose processors with a small memory. This limits the scale, complexity and frequency of AI models, which can be run on-device [34]. Contrary to the terrestrial edge servers, the maritime nodes are unable to make the constant access of the high-performance computing resources, so lightweight and resource-conscious intelligence is necessary but not optimization.

B. Connectivity and Latency Constraints

Communication systems on the sea are defined as heterogeneous and intermittent connectivity, which encompasses short-range wireless networks, shoreline cellular communication networks and satellite backhaul networks [27]. These connections will vary greatly on the latency, reliability, and cost. Although the satellite communication is wide-area, it adds significant delay and constrains the viability of the frequent data exchange or real-time cloud-assisted inference [35].

This makes it impossible to rely on a constant connection with the maritime communication nodes to make a decision. The intelligence should also be able to work independently in cases of prolonged disconnection and also it should be functional even in situations where the quality of a connection is unpredictable. More so, high signal strength does not necessarily imply low latency and consistent operation owing to congestion, multi-hop relaying and protocol-level effects. These issues make the procedures of reactive or signal-threshold-based communication strategies more difficult and require higher requirements of on-device intelligence.

C. Model Robustness and Dynamic Environment

Sea conditions are highly non-stationary in seas, because of diverse sea conditions, weather, ship movement, and varying line-of-sight. AI deployed models on maritime devices have to then face domain shift, whereby the conditions of operation significantly vary between during model development or training stages [36]. On-board AI models in the maritime environment are not subject to retraining or updating regularly, as well as in cloud-based systems, which can last long before external monitoring. This gives emphasis on robustness, stability and generalization instead of peak predictive accuracy. Simulations that are effective in a controlled or static environment can deteriorate quickly in one that is subject to unknown environmental dynamics, and it is time to adopt design strategies that provide a direct consideration of uncertainty and variability. Sea conditions are highly non-stationary in seas, because of diverse sea conditions, weather, ship movement, and varying line-of-sight. AI deployed models on maritime devices have to then face domain shift, whereby the conditions of operation significantly vary between during model development or training stages [36]. On-board AI models in the maritime environment are not subject to retraining or updating regularly, as well as in cloud-based systems, which can last long before external monitoring. This gives emphasis on robustness, stability and generalization instead of peak predictive accuracy. Simulations that are effective in a controlled or static environment can deteriorate quickly in one that is subject to unknown environmental dynamics, and it is time to adopt design strategies that provide a direct consideration of uncertainty and variability.

IV. Conclusion

This paper has explored the design limitations that apply when deploying AI on-device in maritime communications systems. Having severe constraints on power, computation and connectivity whilst with highly dynamic and non-stationary channel conditions, maritime environments contrast in this respect with terrestrial networks. These aspects essentially limit the viability of cloud-aided intelligence and require independent decision-making on embedded communication nodes. This work by classifying the challenges into hardware and energy constraints, connectivity, and latency restrictions, and model resilience to environmental uncertainty reveals the system-level aspects that should be considered during the implementation of an AI on the maritime communications platform. The discussion highlights that good on-device intelligence should be resource-conscious, strong and autonomous more than being complex in its algorithms. These limitations

need to be tackled to have reliable and feasible AI-based maritime communication systems in future non-terrestrial network systems.

ACKNOWLEDGMENT

This research was supported by Korea Institute of Marine Science and Technology Promotion(KIMST) Funded by the Ministry of Oceans and Fisheres(RS-2022-KS221571).

REFERENCES

[1] N. Nomikos, P. K. Gkonis, P. S. Bithas, and P. Trakadas, "A survey on UAV-aided maritime communications: Deployment considerations, applications, and future challenges," *IEEE Open J. Commun. Soc.*, vol. 4, pp. 56–78, 2022.

[2] F. S. Alqurashi, A. Trichili, N. Saeed, B. S. Ooi, and M.-S. Alouini, "Maritime communications: A survey on enabling technologies, opportunities, and challenges," *IEEE Internet Things J.*, vol. 10, no. 4, pp. 3525–3547, 2022.

[3] Y. He, C.-X. Wang, H. Chang, J. Huang, J. Sun, W. Zhang, and E. H. M. Aggoune, "A novel 3D non-stationary maritime wireless channel model," *IEEE Trans. Commun.*, vol. 70, no. 3, pp. 2102–2116, 2021.

[4] M. Zetas, S. Spantideas, A. Giannopoulos, N. Nomikos, and P. Trakadas, "Empowering 6G maritime communications with distributed intelligence and over-the-air model sharing," *Frontiers Commun. Netw.*, vol. 4, Art. no. 1280602, 2024.

[5] H. F. Alhashimi, M. N. Hindia, K. Dimyati, E. B. Hanafi, F. Z. Alden, F. Qamar, and Q. N. Nguyen, "Survey on AI-enabled resource management for 6G heterogeneous networks: Recent research, challenges, and future trends," *Comput. Mater. Continua*, vol. 83, no. 3, 2025.

[6] H. G. Abreha, M. Hayajneh, and M. A. Serhani, "Federated learning in edge computing: A systematic survey," *Sensors*, vol. 22, no. 2, Art. no. 450, 2022.

[7] G. Xylouris, N. Nomikos, A. Kalafatelis, A. Giannopoulos, S. Spantideas, and P. Trakadas, "Sailing into the future: Technologies, challenges, and opportunities for maritime communication networks in the 6G era," *Frontiers Commun. Netw.*, vol. 5, Art. no. 1439529, 2024.

[8] Scale Computing, "Technology innovation at sea," [Online]. Available: https://www.scalecomputing.com/resources/sc-technology-innovation-at-sea

[9] J. C. Molina-Molina, M. Salhaoui, A. Guerrero-González, and M. Arioua, "Autonomous marine robot based on AI recognition for permanent surveillance in marine protected areas," *Sensors*, vol. 21, no. 8, Art. no. 2664, 2021.

[10] X. Wang, Z. Tang, J. Guo, T. Meng, C. Wang, T. Wang, and W. Jia, "Empowering edge intelligence: A comprehensive survey on on-device AI models," *ACM Comput. Surv.*, vol. 57, no. 9, pp. 1–39, 2025.

[11] Q. He, J. Lin, H. Fang, X. Wang, M. Huang, X. Yi, and K. Yu, "Integrating IoT and 6G: Applications of edge intelligence, challenges, and future directions," *IEEE Trans. Serv. Comput.*, 2025.

[12] Z. Chang, S. Liu, X. Xiong, Z. Cai, and G. Tu, "A survey of recent advances in edge-computing-powered artificial intelligence of things," *IEEE Internet Things J.*, vol. 8, no. 18, pp. 13849–13875, 2021.

[13] D. Kumar and F. Zare, "A comprehensive review of maritime microgrids: System architectures, energy efficiency, power quality, and regulations," *IEEE Access*, vol. 7, pp. 67249–67277, 2019.

[14] M. Hoeft, K. Gierlowski, and J. Wozniak, "Wireless link selection methods for maritime communication access networks—A deep learning approach," *Sensors*, vol. 23, no. 1, Art. no. 400, 2022.

[15] R. Anuja and J. Annrose, "End-to-end deep learning for smart maritime threat detection: An AE–CNN–LSTM-based approach," *Sci. Rep.*, vol. 15, no. 1, Art. no. 36316, 2025.

[16] A. Alanhdi and L. Toka, "A survey on integrating edge computing with AI and blockchain in maritime domain, aerial systems, IoT, and Industry 4.0," *IEEE Access*, vol. 12, pp. 28684–28709, 2024.

[17] K. I. Kilic, S. Maity, I. Sung, and P. Nielsen, "Challenges and AI-driven solutions in maritime search and rescue planning: A comprehensive literature review," *Marine Policy*, vol. 178, Art. no. 106692, 2025.

[18] H. Dinis, J. Rocha, T. Matos, L. M. Gonçalves, and M. Martins, "The challenge of long-distance over-the-air wireless links in the ocean: A survey on water-to-water and water-to-land MIoT communication," *Appl. Sci.*, vol. 12, no. 13, Art. no. 6439, 2022.

[19] N. Niknami, A. Srinivasan, K. St. Germain, and J. Wu, "Maritime communications—Current state and the future potential with SDN and SDR," *Network*, vol. 3, no. 4, pp. 563–584, 2023.

[20] "Ship-to-ship maritime wireless channel modeling under various sea state conditions based on REL model," unpublished.

[21] S. Sinha, G. P. Reddy, S. M. Kim, and S. H. Park, "Native AI-based hybrid deep learning for wireless link quality prediction in NTN waterside scenarios," *ICT Express*, 2025.

[22] Y. Liu, C.-X. Wang, H. Chang, Y. He, and J. Bian, "A novel non-stationary 6G UAV channel model for maritime communications," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 10, pp. 2992–3005, 2021.

[23] C. Yang, J. Wang, and Y. Shi, "A multi-dimensional deep-learning-based evaporation duct height prediction model derived from MAGIC data," *Remote Sens.*, vol. 14, no. 21, Art. no. 5484, 2022.

[24] R. Raulefs, M. A. Bellido-Manganell, A. Dammann, M. Walter, and W. Wang, "Verification and modeling of the maritime channel for maritime communications and navigation networks," *Frontiers Mar. Sci.*, vol. 10, Art. no. 1158524, 2023.

[25] T.-H. Wang, L.-H. Shen, and K.-T. Feng, "Distributed multi-agent deep Q-learning for fast roaming in IEEE 802.11ax Wi-Fi systems," in *Proc. IEEE CCNC*, Jan. 2024, pp. 433–438.

[26] J. Zhang, H. Liu, Y. He, W. Gao, N. Xu, and C. Liu, "MCLORA: Maritime ad-hoc communication system based on LoRa," *High-Confidence Comput.*, vol. 5, no. 3, Art. no. 100275, 2025.

[27] N. Niknami, A. Srinivasan, K. St. Germain, and J. Wu, "Maritime communications—Current state and the future potential with SDN and SDR," *Network*, vol. 3, no. 4, pp. 563–584, 2023.

[28] Y. Li, J. Xu, N. Khachatrian, and M.-S. Alouini, "Implementation of an IEEE 802.11ax-based maritime mesh network in the Red Sea," *IEEE Wireless Commun.*, 2025.

[29] S. A. H. Mohsan, Y. Li, M. Sadiq, J. Liang, and M. A. Khan, "Recent advances, future trends, applications and challenges of internet of underwater things (IoUT): A comprehensive review," *J. Mar. Sci. Eng.*, vol. 11, no. 1, Art. no. 124, 2023.

[30] T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta, and D. Sabella, "On multi-access edge computing: A survey of the emerging 5G network edge cloud architecture and orchestration," *IEEE Commun. Surv. Tutor.*, vol. 19, no. 3, pp. 1657–1681, 2017.

[31] B. Li, J. Liu, Y. Liang, Q. Li, H. Liu, Y. Zhang, and S. Chen, "UAV-enabled joint sensing, communication, powering and backhaul transmission in maritime monitoring networks," *arXiv preprint arXiv:2505.12190*, 2025.

[32] Y. Yu, Q. Yang, F. Ji, and W. Zhou, "Research advances in energy management and harvesting technologies for autonomous profiling floats," *Frontiers Mar. Sci.*, vol. 12, Art. no. 1598701, 2025.

[33] W. Jiang, H. Han, D. Feng, L. Qian, Q. Wang, and X.-G. Xia, "Energy-efficient and accuracy-aware DNN inference with IoT device-edge collaboration," *IEEE Trans. Serv. Comput.*, 2025.

[34] D. Ngo, H. C. Park, and B. Kang, "Edge intelligence: A review of deep neural network inference in resource-limited environments," *Electronics*, vol. 14, no. 12, Art. no. 2495, 2025.

[35] T. Wei, W. Feng, Y. Chen, C.-X. Wang, N. Ge, and J. Lu, "Hybrid satellite-terrestrial communication networks for the maritime Internet of Things: Key technologies, opportunities, and challenges," *IEEE Internet Things J.*, vol. 8, no. 11, pp. 8910–8934, 2021.

[36] A. Giannopoulos, P. Gkonis, A. Kalafatelis, N. Nomikos, S. Spantideas, P. Trakadas, and T. Syriopoulos, "From 6G to SeaX-G: Integrated 6G TN/NTN for AI-assisted maritime communications—Architecture, enablers, and optimization problems," *J. Mar. Sci. Eng.*, vol. 13, no. 6, Art. no. 1103, 2025.

A Proposal of Adaptive Threshold Determination Model for Coastal Fishing Vessel Safety Index Using Reinforcement Learning from Human Feedback

1st Sun-Ho Yum
Special Communication & Convergence Service Research Center
Kookmin University
Seoul, Republic of Korea
junsan86@kookmin.ac.kr

2nd Ari Hwang
Dept. of Financial Information Security
Kookmin University
Seoul, Republic of Korea
hwangr2@kookmin.ac.kr

3rd Soo-Hyun Park
College of Computer Science
Kookmin University
Seoul, Republic of Korea
shpark21@kookmin.ac.kr

***Abstract*—Digital transformation in the maritime sector is essential for preventing accidents in small coastal fishing vessels. In the actual operational environment, these vessels generate continuous raw sensor data. However, existing safety systems relying on fixed thresholds often fail due to non-Gaussian environmental noise. Furthermore, despite the availability of raw data, applying AI-based approaches is challenging due to the absence of standardized labeled datasets required for supervised learning. Since raw sensor logs often lack specific risk annotations and every vessel has unique dynamic characteristics, a generic pre-trained model cannot be directly applied. To address this challenge, this study proposes an Adaptive Safety Index Model and a threshold determination framework. The proposed system utilizes an In-situ Labeling Phase to generate valid labels from live sensor streams via user interaction and employs a Reinforcement Learning from Human Feedback (RLHF) agent to dynamically calibrate safety thresholds. We present a system architecture and a mathematical methodology that transforms abundant raw data into a personalized safety criterion.**

***Index Terms*—Maritime Safety, Coastal Fishing Vessel, Reinforcement Learning from Human Feedback (RLHF), Adaptive Thresholding, Context-Aware System**

I. Introduction

The digital transformation of the marine and fisheries sector is a primary means of strengthening the safety management system for fishing vessels. The South Korean government is promoting the management stability of fishermen through national development plans [1]. Furthermore, it is pushing for the dissemination of digital safety infrastructure to reduce casualties caused by fishing vessel accidents and secure golden time for accident response [2], [3].

Most current operation safety modules utilize a fixed threshold method to determine danger. However, the marine environment often exhibits non-Gaussian noise distributions due to waves and weather changes, leading to frequent false alarms with simple threshold methods [4]. While radar systems apply computational algorithms to address this, applying such methods directly to IoT sensors for small fishing vessels with limited computing resources presents significant challenges [5], [6].

Moreover, applying AI to small vessels faces a critical hurdle regarding data utility. While fishing vessels operate with various sensors that generate real-time raw data and potentially accumulate past logs, these datasets often lack the corresponding ”safety labels” (i.e., ground truth indicating whether a specific motion was dangerous). Without these labels, the abundant raw data cannot be directly used for supervised learning. Furthermore, due to the heterogeneity of hull structures and loading conditions, a model trained on one vessel’s past data cannot be generalized to another.

To resolve these issues, this study proposes a Context-Aware Adaptive Safety Index Model and a threshold determination framework. Instead of assuming the existence of a perfectly labeled dataset, our system leverages live raw data streams combined with operational context (e.g., weather, zone). It incorporates an In-situ Labeling Phase to collect vessel-specific labels via user interaction during the initial deployment. Subsequently, the RLHF agent dynamically adjusts decision criteria by reflecting user feedback in real-time, while ensuring data integrity through synchronization with external heterogeneous networks [10]. This in-situ calibration mechanism realizes a personalized safety system that reduces false alarms and complies with safety standards.

II. Related Work

A. Availability vs. Usability of Maritime Data

Effective maritime safety systems rely on consistent data protocols. Large vessels utilize NMEA standards for sensor integration [8], [9], and small coastal fishing vessels also generate data through proprietary analog sensors or GPS plotters. However, a gap exists between the availability of

This research was supported by Korea Institute of Marine Science & Technology Promotion(KIMST) funded by the Ministry of Oceans and Fisheries(RS-2022-KS221571).

raw data and its usability for AI training. While raw sensor logs (e.g., roll, pitch, location) are available, they are typically unlabeled, meaning they do not explicitly record hazardous events. Unlike autonomous driving domains where massive annotated datasets exist, maritime data for small vessels is often fragmented and lacks semantic risk information. This makes it impossible to apply standard supervised learning techniques directly, necessitating a system capable of in-situ data labeling to bridge the gap between raw sensor inputs and safety judgments.

B. Limitations of Classical Detection Theory

According to classical statistical signal processing theory [4], maximizing the probability of detection (P_D) while constraining the probability of false alarm (P_{FA}) is a fundamental challenge. In the maritime domain, environmental noise such as sea clutter often follows non-Gaussian distributions (e.g., Weibull distribution) [5]. Previous studies applied CFAR (Constant False Alarm Rate) algorithms to adaptively adjust thresholds [6], [7]. However, these methods rely on statistical assumptions of noise, which may not hold true for the complex, non-linear stability dynamics of small fishing vessels.

C. Human-in-the-Loop and RLHF

To overcome the limitations of fixed thresholds and the absence of labeled data, recent advancements emphasize Human-in-the-Loop (HITL) approaches. HITL frameworks incorporate human guidance to prevent catastrophic failures and fine-tune model parameters [10].

In particular, Reinforcement Learning from Human Feedback (RLHF) is a methodology that trains an agent using human preferences as a reward signal when a mathematical reward function is difficult to define [14]. In the context of small vessel safety, objectively defining a "Safety Function" is challenging due to the complex interplay of dynamic variables. Therefore, this study adopts RLHF to utilize the captain's subjective judgment (e.g., "This motion is safe") as a direct ground truth label. By treating user feedback as a reward signal, the agent can optimize the safety threshold without a pre-defined mathematical model.

III. System Architecture

The proposed system architecture addresses the instability of maritime communication and the hardware constraints of small fishing vessels. Unlike conventional cloud-centric IoT systems, we adopt an Edge Computing approach [12] to ensure real-time safety alerts even under disconnected network conditions. As illustrated in Fig. 1, the system comprises three hierarchical layers: the Sensing Layer for multi-source aggregation, the Integrated Edge Processing Layer for adaptive risk assessment, and the Application Layer for user interaction and feedback.

A. Sensing Layer (Multi-Source Aggregation)

To ensure precise risk assessment for diverse small vessels, the Sensing Layer aggregates two distinct categories of data: Static Data and Dynamic Sensor Data.

- Static Data (Vessel Profile): Unlike generic safety systems, our approach incorporates fixed vessel characteristics such as gross tonnage, ship type, and length. These static parameters are essential for determining the initial safety margins, as smaller vessels are more susceptible to lower wave heights compared to larger ships.
- Dynamic Sensor Data: This includes real-time environmental and navigational inputs. As discussed in Section II, relying solely on standard protocols excludes vessels with legacy devices. We implement a dual-input interface to handle this heterogeneity (see Table 1).
  - Standard Interface: Direct parsing of NMEA 0183 sentences (e.g., $GPRMC) via serial communication.
  - Legacy Support: An Analog-to-Digital (ADC) bridge converts signals from non-standard analog sensors into a pseudo-NMEA format.

TABLE I
Sensor Data Integration: Standard and Legacy Support

Sensor Type	Input Interface	Processed Format
GPS / Location	Serial (Standard)	$GPRMC
IMU (Roll/Pitch)	Serial (Standard)	$PASHR
Weather (Wind)	Serial (Standard)	$WIMDA
Bilge / Level	Analog (Legacy)	Pseudo-$YXXDR
Engine Temp	Analog (Legacy)	Pseudo-$WIMDA

B. Integrated Edge Processing Layer

The Edge Gateway serves as a unified computing unit. To overcome the limitations of rigid rule-based systems, this layer employs an integrated approach combining deterministic rules with stochastic reinforcement learning. The processing pipeline consists of three stages:

1) Data Standardization: The module synthesizes heterogeneous inputs from the Sensing Layer. It parses NMEA strings, converts analog signals, and performs Min-Max Normalization to generate a consistent numerical state vector. This ensures that the subsequent logic receives uniform input dimensions regardless of the sensor hardware.
2) Integrated Risk Assessment:
   - Rule-based Logic (Safety Envelope): First, a deterministic algorithm checks for critical limit violations derived from maritime engineering standards. This component acts as a fail-safe mechanism, overriding the AI model when physical limits are exceeded.
   - RLHF Agent: If the vessel is within physical limits, the RL agent processes the state vector. It adaptively computes the Safety Index (y_t) by applying a sensitivity parameter (α_t) learned from the initial labeling phase and continuous user feedback.
3) Local Alert Execution: If the final Safety Index exceeds the dynamic threshold, the edge device triggers a local alarm independently of network connectivity.

C. Application Layer (Interactive Interface)

The Application Layer provides a visual interface for the captain and serves as a crucial component for the Human-in-the-Loop mechanism. Beyond simple monitoring, this layer acts as a data labeling tool during the initial phase.

1) Monitoring View: Displays the real-time Safety Index (y_t) and sensor statuses.
2) Root Cause Analysis (RCA): Identifies the primary factor contributing to the risk to enhance explainability.
3) Feedback Interface: Enables the user to provide ground truth labels.
   - Calibration Mode (Phase 1): The user labels synthesized scenarios to generate the initial training dataset (D_{init}).
   - Operation Mode (Phase 3): The user flags 'False Alarms' or 'Confirms' risks to update the sensitivity parameter (α_t) in real-time.

IV. Methodology

To implement the proposed architecture, we define a mathematical model that incorporates not only raw sensor data but also operational contexts. This methodology translates the physical constraints of the vessel, environmental categories, and the captain's subjective judgment into a unified computational framework.

A. State Space and Contextual Data Fusion

The input space is defined to capture the full operational reality of the vessel. We expand the state definition to include a Context Vector, ensuring that risk assessment is conditioned on environmental and vessel-specific factors.

The Total State Vector X_t at time t is defined as:

$$X_t = S_t \oplus C_t \tag{1}$$

- Sensor State (S_t): A normalized vector of real-time dynamic data, including roll, pitch, and change rates derived from the raw sensor stream.
- Context Vector (C_t): A categorical vector representing the operational environment.

$$C_t = \{\text{VesselType}, \text{ZoneID}, \text{WeatherCode}\} \tag{2}$$

Here, 'VesselType' (e.g., Gillnet, Trap) determines the static stability center, 'ZoneID' indicates the sea area (Coastal vs. Open sea), and 'WeatherCode' reflects real-time meteorological conditions.

B. Safety Index Determination Model

The Safety Index (y_t) is determined through a Hybrid Inference Mechanism that prioritizes physical safety limits while allowing for context-aware adaptation using X_t.

1) Hard Constraint: Rule-based Envelope: We define a physical safety limit set Δ_{limit} based on maritime regulations. If the current sensor state violates these limits, the system overrides the AI model:

$$y_{rule} = \begin{cases} 1.0 & \text{if } \exists s \in S_t : s > \delta_{critical} \in \Delta_{limit} \\ 0.0 & \text{otherwise} \end{cases} \tag{3}$$

2) Soft Constraint: Context-Aware RLHF Inference: When the vessel operates within physical limits ($y_{rule} = 0$), the RLHF agent π_θ computes a risk score based on the total state X_t. The final Safety Index y_t is modulated by a sensitivity parameter α_t:

$$y_t = \max(y_{rule}, \sigma(\pi_\theta(S_t, C_t)) \cdot \alpha_t) \tag{4}$$

By including C_t as an input, the policy network π_θ learns to differentiate risk levels even for identical sensor values depending on the weather or zone.

C. Adaptive Threshold Determination Algorithm

The critical innovation of this model includes the dynamic adjustment of the sensitivity parameter α_t and the assurance of data integrity through external heterogeneous network integration. The process is defined in Table 2.

1) Phase 1 (In-situ Labeling Phase): Since no pre-trained model exists, the system uses synthesized scenarios

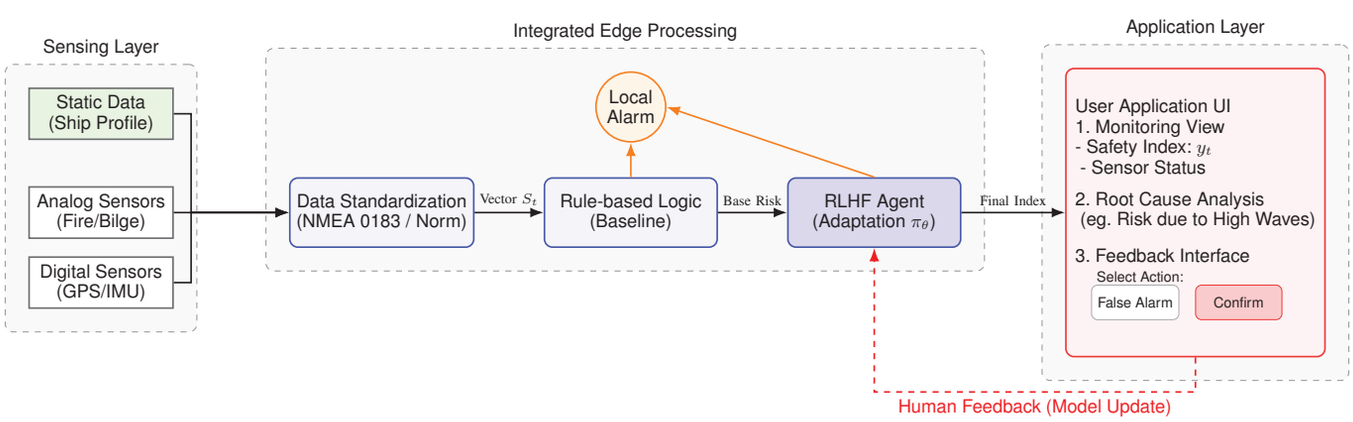


Fig. 1. Overall System Architecture. The Sensing Layer standardizes heterogeneous inputs (static/dynamic) into NMEA 0183 format. The Integrated Edge Processing Layer combines a rule-based baseline with an RLHF agent for adaptive risk assessment. The Application Layer visualizes the Safety Index, provides Root Cause Analysis (RCA), and collects user feedback.

categorized by the Context Vector C_t (e.g., Fishing Type, Zone, Weather). The captain labels these, creating a context-aware initial dataset D_{init}. This addresses the lack of ground truth labels by generating valid training data on-site.

2) Phase 2 (Initialization): The policy network π_θ is pre-trained using D_{init} to learn the baseline risk perception.
3) Phase 3 (Real-time Adaptation & Network Sync): During operation, the system performs two critical tasks:
   - Parameter Adaptation: Updates α_t based on user feedback a_t to reduce false alarms.
   - Data Consistency via Heterogeneous Networks: It synchronizes the validated data tuple (S_t, C_t, a_t) with external systems (e.g., Shore Control Center) via heterogeneous networks (LTE-M, Satellite). This ensures that the locally generated "Ground Truth" is archived and verified against global standards.

D. Analysis of Sensor Data Trends and Labeling Feasibility

We analyze the necessity of the proposed categorization and real-time synchronization by examining the statistical characteristics of maritime data.

1) Statistical Characteristics of Sensor Data Trends: Small coastal fishing vessels exhibit high volatility in sensor data. As identified in Reference [5], sea clutter follows a non-Gaussian distribution. Therefore, a conventional approach that labels a single quantitative value (e.g., "Roll > 20 degrees") as dangerous fails to capture the actual stability trend. Our approach evaluates the "trend" over a time window, converting non-linear raw data into valid safety judgments.

2) Impact of Multi-dimensional Categorization: For labeling to be valid, it must be conditioned on the context C_t. A specific roll angle might be safe in a "Calm/Coastal" context but dangerous in a "Stormy/Open-sea" context due to the difference in wave periods. By fusing C_t into the input space (Section 4.1), the proposed model ensures that a label provided in one environment is not wrongly generalized to a disparate environment, significantly increasing the reliability of the Safety Index.

3) Real-time Synchronization of Labels: The feasibility of this system relies on the real-time linkage between the raw data stream and user feedback. Unlike offline learning, our Edge Computing architecture ensures that when feedback a_t is triggered, it is time-stamped and bound to the exact (S_t, C_t) tuple. This instantaneous synchronization preserves the "momentary context" of the decision, allowing the adaptive algorithm to learn the precise boundary between safe maneuvering and genuine risk.

TABLE II
ALGORITHM 1: CONTEXT-AWARE ADAPTIVE SAFETY INDEX & EXTERNAL SYNC

Require: Sensor State S_t, Context Vector C_t, User Feedback a_t
Ensure: Safety Index y_t, Sensitivity Parameter α_t, External Sync Status

// Phase 1: In-situ Labeling Phase
Initialize dataset $D_{init} \leftarrow \emptyset$
for $i \leftarrow 1$ to N do
  Generate synthetic scenario S_{syn} based on context C_{syn}
  Get user label $y_{label} \leftarrow$ Query(S_{syn}, C_{syn})
  $D_{init} \leftarrow D_{init} \cup \{(S_{syn} \oplus C_{syn}, y_{label})\}$
end for

// Phase 2: Initialization (Pre-training)
Initialize π_θ and train with D_{init}
Set initial sensitivity $\alpha_0 \leftarrow 0.7$

// Phase 3: Real-time Adaptation (Main Control Cycle)
loop
  Observe current states S_t and C_t
  // Step A: Hard Constraint Check
  if $\exists x \in S_t > \delta_{critical}$ then
    $y_t \leftarrow 1.0$ (Force Alarm)
  else
    // Step B: Context-Aware Inference
    $y_{raw} \leftarrow \sigma(\pi_\theta(S_t, C_t))$
    $y_t \leftarrow y_{raw} \cdot \alpha_t$
  end if

  // Step C: Feedback Update
  if Alarm Triggered and Feedback Received then
    if $a_t =$ False Alarm then
      $\alpha_{t+1} \leftarrow \alpha_t - \eta \cdot y_t$ (Decay)
    else if $a_t =$ Confirm then
      $\alpha_{t+1} \leftarrow \alpha_t + \epsilon$ (Reinforce)
    end if

    // Step D: Data Consistency with External Networks
    $Data_{packet} \leftarrow \{S_t, C_t, y_t, a_t, \alpha_{t+1}\}$
    SyncToExternalSystem$(Data_{packet})$
  end if
end loop

V. CONCLUSION

This study proposed a Context-Aware Adaptive Safety Index Model and a threshold determination framework to address the safety management challenges of small coastal fishing vessels. Existing systems face limitations due to non-Gaussian environmental noise and the lack of standardized labeled datasets, preventing the direct application of supervised learning. To overcome these issues, we designed a system that utilizes Human-in-the-Loop (HITL) reinforcement learning. By incorporating an In-situ Labeling Phase and a real-time feedback mechanism, the system transforms subjective human judgment into an objective safety criterion. The theoretical feasibility analysis confirms that this approach is computationally efficient ($O(1)$) and structurally suitable for data-scarce maritime environments.

Future work will focus on deriving practical requirements for threshold settings using real-world sensor data obtained from currently operating vessels. We aim to propose a concrete 'Coastal Fishing Vessel Safety Index Model' that is fully applicable to actual field operations and quantitatively verify

its operational feasibility. Furthermore, we plan to refine the system architecture by stratifying the Sensing Layer into three hierarchical levels: (1) Sensor-specific, (2) Zone-specific, and (3) Whole-vessel integration. Data collected from this hierarchical structure will be used to train the model, which will then be iteratively corrected and verified by human experts to develop a highly reliable, field-deployable safety model.

REFERENCES

[1] Ministry of Oceans and Fisheries, "The 3rd Basic Plan for Oceans and Fisheries Development (2021-2030)," Sejong, Republic of Korea, 2021.

[2] Ministry of Oceans and Fisheries, "The 1st Basic Plan for Intelligent Maritime Traffic Service," Sejong, Republic of Korea, 2021.

[3] Ministry of Oceans and Fisheries, "2025 Implementation Plan for Intelligent Maritime Traffic Information Service," Sejong, Republic of Korea, 2024.

[4] S. M. Kay, Fundamentals of Statistical Signal Processing: Detection Theory. Upper Saddle River, NJ, USA: Prentice Hall, 1998.

[5] A. H. M. Mboungam, Y. Zhi, and C. K. F. Monguen, "Constant false alarm rate detection method in mixed Weibull distribution sea clutter," Digital Signal Processing, vol. 149, p. 104494, 2024.

[6] J. H. Shin and Y. Choi, "Robust Control for the Detection Threshold of CFAR Process in Cluttered Environments," Sensors, vol. 20, no. 14, p. 3913, 2020.

[7] M. Alaee, M. Firoozmand, R. Amiri, and M. Sepahvand, "Adaptive Thresholding in Marine RADARs," Iranian Journal of Electrical & Electronic Engineering, vol. 6, no. 3, pp. 174-181, Sep. 2010.

[8] National Marine Electronics Association, "NMEA 0183 Standard," [Online]. Available: https://www.nmea.org/nmea-0183.html

[9] S. Spitzer, L. Luft, and D. Morschhauser, "NMEA 2000: Past, Present and Future," in Proc. RTCM Annu. Assembly Meeting Conf., St. Pete Beach, FL, USA, May 2009.

[10] Y. Emami, K. Li, L. Almeida, W. Ni, and Z. Han, "Human-In-The-Loop Machine Learning for Safe and Ethical Autonomous Vehicles: Principles, Challenges, and Opportunities," arXiv preprint arXiv:2408.12548, 2024.

[11] A. N. Abbas et al., "Analyzing State of Operators and the Impact of AI-Enhanced Decision Support in Control Rooms: A Human-in-the-Loop Specialized Reinforcement Learning Framework," arXiv preprint arXiv:2402.13219, 2024.

[12] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge Computing: Vision and Challenges," IEEE Internet of Things Journal, vol. 3, no. 5, pp. 637-646, Oct. 2016.

[13] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction, 2nd ed. Cambridge, MA, USA: MIT Press, 2018.

[14] P. F. Christiano et al., "Deep reinforcement learning from human preferences," in Proc. Adv. Neural Inf. Process. Syst. (NIPS), Long Beach, CA, USA, 2017, pp. 4299-4307.

A Self-Supervised Energy Flow Signature–Based Fault Detection Framework for DER Systems

Murad Ali Khan*, Syed Shehryar Ali Naqvi**, Misbah Said*, Faiza Qayyum*, DoHyeun Kim *,***
* *Department of Computer Engineering, Jeju National University, Jeju-do, Republic of Korea*
** *Department of Electronics Engineering, Jeju National University, Jeju-do, Republic of Korea*
*** *Big Data Research Center, Jeju National University, Jeju-do, Republic of Korea*
Emails: muradali@stu.jejunu.ac.kr, syedshehryar@stu.jejunu.ac.kr, misbahbibi@stu.jejunu.ac.kr, faizaqayyum91@gmail.com , kimdh@jejunu.ac.kr

***Abstract*—Reliable fault detection in DER systems is essential for ensuring operational safety, efficiency, and resilience under increasing penetration of Distributed Energy Resources (DERs). This paper proposes a self-supervised fault-detection framework based on Energy Flow Signature (EFS) modeling, jointly representing load and battery power dynamics to capture intrinsic energy interactions. High-resolution DER measurements sampled at 15-minute intervals are processed to construct normalized EFS sequences, which are learned using a Gated Recurrent Unit (GRU) predictor with a sliding window of 16-time steps (4 hours). Faults are identified by analyzing the prediction error between observed and predicted EFS, aggregated through a Temporal Consistency Score (TCS) computed over 8-time steps (2 hours). A statistical threshold derived from regular operation enables reliable fault decision without requiring labeled fault data. Experimental evaluation on 87,821 real DER samples demonstrates effective detection of persistent abnormal behavior, identifying 353 fault points with a fault rate of 1.3%, while maintaining low false alarms. The results confirm that the proposed framework provides a robust, scalable, and data-driven solution for fault detection in DER energy management systems.**

Keywords - DER fault detection, Energy Flow Signature, Self-supervised learning, Gated Recurrent Unit, Temporal consistency analysis, Energy management systems

I. Introduction

The rapid integration of DERs, such as photovoltaic systems and battery energy storage, has accelerated the adoption of DER in modern buildings and campus-scale energy systems. DER offer improved energy efficiency, resilience, and local energy autonomy; however, their increasing operational complexity introduces new challenges in reliability, fault detection, and system maintenance [1]. Undetected faults in load behavior, storage operation, or energy coordination can lead to performance degradation, safety risks, and economic losses. Consequently, the development of intelligent and autonomous fault detection mechanisms has become a critical requirement for sustainable DER operation [2].

Traditional fault detection techniques in power and energy systems largely rely on rule-based monitoring, threshold violations, or model-driven approaches derived from physical system equations. While effective in well-instrumented and stable environments, these methods often struggle in DER due to nonlinear dynamics, stochastic load behavior, and limited availability of labeled fault data. Moreover, manual threshold tuning and reliance on expert knowledge reduce scalability and adaptability, particularly in real-world deployments where operational conditions evolve over time. These limitations motivate the exploration of data-driven and self-learning approaches that can adapt to dynamic energy flow patterns without requiring explicit fault annotations [3],[4].

Recent advances in machine learning and deep learning have enabled predictive modeling of energy systems using historical operational data. Recurrent neural networks (RNNs), and especially GRUs, have demonstrated strong capability in modeling temporal dependencies in load forecasting, energy storage behavior, and anomaly detection [5]. However, many existing learning-based fault detection methods either assume the availability of labeled fault datasets or focus on single-signal analysis, which limits their effectiveness in multi-component DER environments [6]. Furthermore, instantaneous prediction errors may produce false alarms when system fluctuations are transient rather than persistent.

To address these challenges, this paper proposes a self-supervised fault detection framework based on EFS modeling [7]. Instead of analyzing individual signals independently, the proposed method jointly represents load power and battery power as a compact EFS that captures the intrinsic interaction between demand and storage dynamics. A GRU-based predictor is trained exclusively on normal operational data to learn expected temporal behavior. Faults are identified by analyzing deviations between predicted and observed EFS, enabling detection without requiring labeled fault samples.

In addition, the proposed framework introduces a TCS to distinguish persistent abnormal behavior from short-term fluctuations. By aggregating prediction errors over a sliding temporal window, the method enhances robustness against noise and operational variability. Experimental validation using real DER measurement data demonstrates that the proposed approach can reliably detect anomalous energy flow patterns while maintaining low false alarm rates. The results highlight the effectiveness of combining energy flow

modeling, sequence learning, and temporal consistency analysis for practical, scalable fault detection in DER systems.

II. Related Work

Fault detection in power systems and smart grids has been extensively studied using both model-based and data-driven approaches. Early works primarily relied on physical system models, state estimation, and rule-based thresholding to identify abnormal operating conditions [8],[9]. While these techniques are effective in transmission and distribution networks with well-defined system dynamics, they are less suitable for DER environments where load behavior, storage interaction, and operational conditions are highly dynamic and nonlinear. Additionally, model-based methods often require precise system parameters and expert tuning, which limits scalability in real-world deployments.

With the availability of high-resolution smart meter and sensor data, machine learning-based fault detection methods have gained increasing attention. Classical approaches using support vector machines, k-means clustering, and principal component analysis have been proposed for anomaly and fault detection in energy systems [10],[11]. However, these methods typically depend on handcrafted features and struggle to capture long-term temporal dependencies inherent in energy flow data. Moreover, many supervised learning techniques assume access to labeled fault data, which is rarely available in practical DER operation.

More recently, deep learning techniques, particularly RNNs and long short-term memory (LSTM) models, have been employed for load forecasting, energy prediction, and anomaly detection in smart grids [12],[13]. GRUs have emerged as a computationally efficient alternative to LSTMs, offering comparable performance with reduced model complexity [14]. Several studies have utilized prediction error–based anomaly detection, where deviations between predicted and observed signals indicate abnormal behavior [15]. However, most existing works focus on single-variable monitoring, such as load or voltage, which limits their ability to detect faults arising from interactions between multiple energy components, such as load-storage coupling in DER.

In addition, recent studies have explored temporal aggregation and statistical scoring mechanisms to reduce false alarms caused by transient fluctuations [16],[17]. While these approaches improve robustness, they are often applied as post-processing steps without explicitly modeling energy flow relationships. In contrast, the proposed method introduces an EFS that jointly represents load and battery power dynamics, combined with a self-supervised GRU predictor and a TCS. This integrated design enables robust fault detection without labeled data while explicitly capturing persistent deviations in multivariate energy flow behavior, addressing key limitations of existing approaches.

III. Proposed Signature–based fault detection framework

This section describes the proposed EFS-based fault detection framework, which combines self-supervised temporal learning with a temporal consistency-based decision mechanism. The overall workflow of the proposed methodology is illustrated in Figure 1.

3.1. Overview of the Proposed Framework

As illustrated in Figure 1, the proposed fault detection framework follows a sequential, behavior-driven pipeline for identifying abnormal DER operation. The process begins with raw DER measurements, where building load power $L(t)$ and battery power $B(t)$ are collected at a 15-minute sampling interval. These measurements are first subjected to data cleaning and synchronization, including timestamp parsing, removal of missing or invalid records, and temporal alignment. Next, the cleaned signals are combined to construct the EFS, which jointly represents load and battery dynamics. To ensure scale invariance and stable learning, the EFS is normalized using min–max normalization. Temporal dependencies are then captured through a sliding window generator, which forms fixed-length EFS sequences for time-series modeling.

A GRU-based predictor is trained in a self-supervised manner to learn normal energy flow behavior by predicting the next-step EFS. During inference, the deviation between predicted and actual EFS values is quantified as a prediction error, which is further aggregated over time using a TCS. Finally, a fault decision is made by comparing the TCS against a statistically defined threshold, enabling reliable detection of persistent abnormal behavior without requiring labeled fault data.

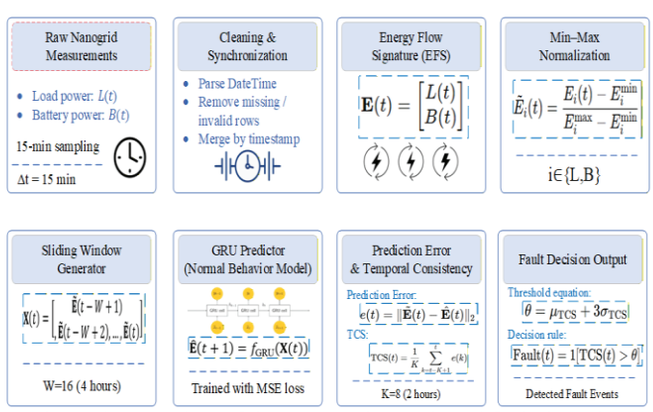


Figure 1: Overview of the proposed EFS–based fault detection framework for DER systems.

3.2. Energy Flow Signature Construction

Let $L(t)$denote the building load demand and $B(t)$denote the battery power at time step t. These two signals jointly characterize the energy exchange within the DER.

The EFS at time $t$is defined as:

$$\mathrm{E}(t) = \begin{bmatrix} L(t) \\ B(t) \end{bmatrix} \qquad (1)$$

To eliminate scale differences and improve learning stability, each component of the EFS is normalized using min–max normalization:

$$\tilde{E}_i(t) = \frac{E_i(t) - E_i^{\min}}{E_i^{\max} - E_i^{\min}}, i \in \{L, B\} \quad (2)$$

The normalized EFS $\tilde{\mathrm{E}}(t)$ provides a compact and behavior-aware representation of DER operation.

3.3. Temporal Sequence Modeling

To capture temporal dependencies, sliding window sequences of length W are constructed from the normalized EFS:

$$\mathrm{X}(t) = \begin{bmatrix} \tilde{\mathrm{E}}(t-W+1) \\ , \tilde{\mathrm{E}}(t-W+2), \dots, \tilde{\mathrm{E}}(t) \end{bmatrix} \quad (3)$$

Each sequence $\mathrm{X}(t)$ is used to predict the next-step EFS $\tilde{\mathrm{E}}(t+1)$.

3.4. GRU-Based Normal Behavior Learning

A GRU network is employed to learn normal temporal energy flow behavior. Given an input sequence $\mathrm{X}(t)$, the GRU predicts the expected energy flow at the next time step:

$$\hat{\mathrm{E}}(t+1) = f_{\mathrm{GRU}}(\mathrm{X}(t)) \quad (4)$$

The model is trained in a self-supervised manner by minimizing the mean squared prediction error:

$$\mathcal{L} = \frac{1}{N} \sum_{t=1}^{N} \| \tilde{\mathrm{E}}(t+1) - \hat{\mathrm{E}}(t+1) \|_2^2 \quad (5)$$

Through this process, the GRU captures the normal temporal dynamics of energy consumption and storage interaction.

3.5. Prediction Error Computation

At inference time, the deviation between the predicted and actual EFS is quantified using the L2 norm:

$$e(t) = \| \tilde{\mathrm{E}}(t) - \hat{\mathrm{E}}(t) \|_2 \quad (6)$$

Instantaneous prediction errors reflect short-term deviations but are insufficient for reliable fault detection due to transient fluctuations.

3.6. Temporal Consistency Score for Fault Detection

To distinguish persistent faults from transient variations, a TCS is introduced. The TCS aggregates prediction errors over a rolling window of size K:

$$\mathrm{TCS}(t) = \frac{1}{K} \sum_{k=t-K+1}^{t} e(k) \quad (7)$$

A fault is detected when the TCS exceeds a statistically defined threshold θ:

$$\theta = \mu_{\mathrm{TCS}} + 3\sigma_{\mathrm{TCS}} \quad (8)$$

$$\mathrm{Fault}(t) = \begin{cases} 1, & \text{if } \mathrm{TCS}(t) > \theta \\ 0, & \text{otherwise} \end{cases} \quad (9)$$

This decision mechanism ensures that only sustained deviations trigger fault alarms, improving robustness against noise and transient operational changes.

IV. Experimental Results and Discussion

This section presents the experimental evaluation of the proposed EFS–based fault detection framework using real DER operational data. The objective of the experiments is to demonstrate the ability of the proposed approach to (i) learn normal operational behavior in a self-supervised manner, and (ii) detect faults through persistent deviations in temporal energy flow patterns.

4.1. GRU Training Convergence Analysis

To learn normal energy flow dynamics, a GRU-based predictor was trained using historical EFS sequences constructed from building load and battery power measurements. Figure 2 illustrates the training loss convergence over epochs.

As shown in Figure 2, the mean squared error (MSE) decreases rapidly during the initial epochs and stabilizes after a small number of iterations. This behavior indicates that the GRU model effectively captures the underlying temporal relationships in the EFS without overfitting. The smooth convergence further confirms the stability of the training process and validates the suitability of the selected model architecture for normal behavior learning.

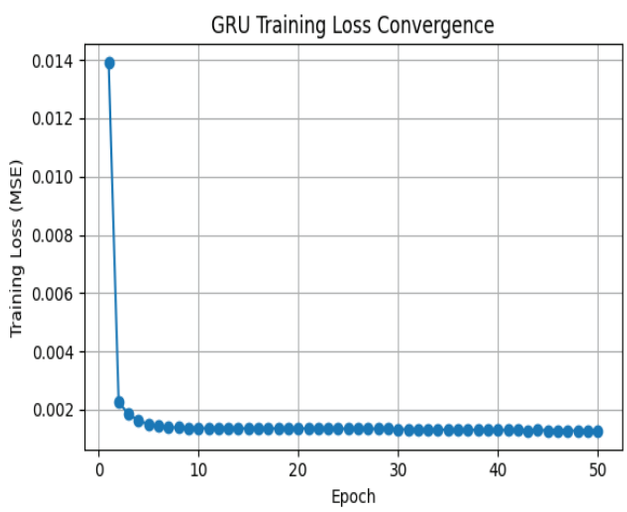


Figure 2: Training loss convergence of the GRU-based predictor over epochs.

4.2. Distribution of Prediction Errors

After training, the GRU model was applied to unseen test data to compute prediction errors between the actual and predicted EFS. The distribution of these errors is shown in Figure 3.

The histogram in Figure 3 demonstrates that the majority of prediction errors are concentrated near low values, reflecting accurate modeling of normal operating conditions. A long but sparse tail is observed, corresponding to rare but significant deviations in system behavior. This error distribution provides a strong statistical basis for defining an adaptive fault detection threshold and highlights the presence of abnormal operational patterns within the dataset.

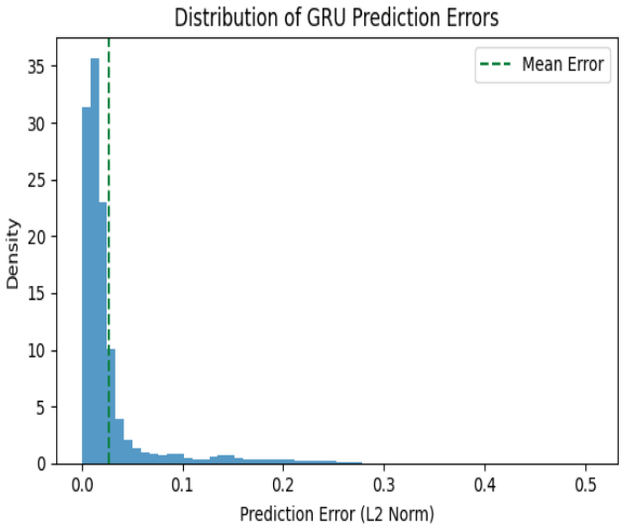


Figure 3: Probability distribution of GRU prediction errors on the test dataset.

4.3. Temporal Consistency–Based Fault Detection

Fault detection is performed using the proposed TCS, which aggregates prediction errors over a rolling temporal window to distinguish persistent anomalies from transient fluctuations. Figure 4 presents the TCS values over time along with the statistically derived fault threshold.

As observed in Figure 4, the TCS remains below the threshold for the majority of the operational period, indicating normal behavior. However, several sustained peaks exceed the threshold, signaling abnormal energy flow behavior. Unlike instantaneous anomaly detectors, the proposed method suppresses short-lived spikes and identifies faults only when deviations persist over time, thereby reducing false alarms.

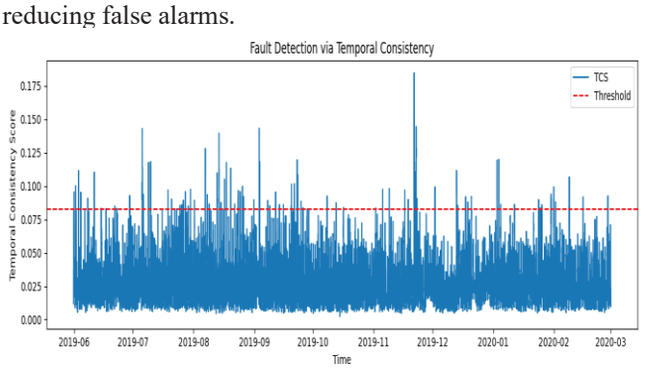


Figure 4: Temporal Consistency Score (TCS) over time with the statistically derived fault detection threshold.

4.4. Zoomed Analysis of Detected Fault Events

To further illustrate the effectiveness of the proposed approach, Figure 5 shows a zoomed-in view of a representative time window containing detected fault events.

In this zoomed view, individual fault occurrences are clearly marked when the TCS crosses the detection threshold. The visualization confirms that detected faults correspond to sustained deviations rather than isolated fluctuations. This behavior is particularly important for real DER environments, where transient load changes and operational noise are common.

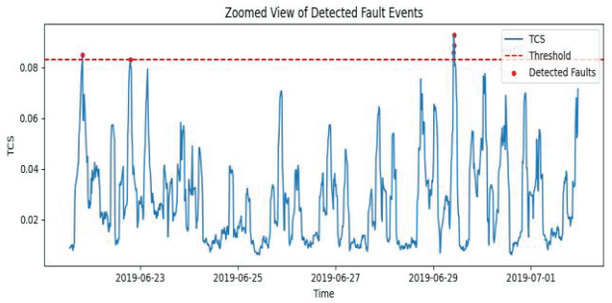


Figure 5: Zoomed-in visualization of the Temporal Consistency Score showing detected fault events when sustained deviations exceed the defined threshold.

4.5. Quantitative Fault Detection Summary

Table 1 summarizes the key quantitative results obtained from the fault detection experiments.

The proposed method detected 353 fault points out of 26,342 test samples, corresponding to a fault occurrence rate of approximately 1.3%. This rate is consistent with realistic fault frequencies observed in operational energy systems. The table also reports the selected window sizes for sequence modeling and temporal consistency computation, providing full transparency and reproducibility of the experimental setup.

Table 1: Summary of Fault Detection Results

Metric	Value
Total samples	26,342
Detected fault points	353
Fault rate	1.3%
TCS threshold	0.081
Window size	16
TCS window	8

V. Discussion

The experimental results demonstrate that the proposed EFS–based fault detection framework is capable of identifying abnormal DER behavior in a fully self-supervised manner. By jointly modeling load demand and battery power dynamics, the method captures the intrinsic operational behavior of the energy system rather than relying on isolated signal thresholds.

A key observation from the results is the stability of GRU-based normal behavior learning, as evidenced by the smooth convergence of training loss. This confirms that the temporal dependencies within the EFS are sufficiently informative to characterize normal system operation. Unlike rule-based or static statistical models, the GRU effectively adapts to temporal variations in demand and storage interactions, which are common in real DER environments.

The distribution of prediction errors further highlights the suitability of the proposed approach for fault detection. Most prediction errors remain concentrated at low magnitudes, indicating accurate modeling of normal behavior, while a small number of large deviations correspond to abnormal operating conditions. This clear separation enables statistically grounded threshold selection without requiring fault labels, making the approach practical for real-world deployments where labeled fault data is scarce or unavailable.

The introduction of the TCS plays a critical role in improving fault detection robustness. Rather than responding to instantaneous deviations, the TCS aggregates prediction errors over time, allowing the system to detect faults only when abnormal behavior persists. This characteristic significantly reduces false alarms caused by transient load changes, measurement noise, or short-term operational fluctuations. The zoomed fault analysis confirms that detected fault events correspond to sustained deviations, reinforcing the behavioral nature of the proposed detection mechanism.

From an application perspective, the detected fault rate of approximately 1.3% aligns with realistic expectations for operational energy systems. This suggests that the proposed method can serve as an effective early warning mechanism, enabling timely intervention before minor anomalies escalate into critical system failures. Moreover, the framework is computationally lightweight and relies only on commonly available measurements, making it suitable for deployment in resource-constrained edge or on-device monitoring environments. Despite these advantages, the current implementation focuses on fault presence detection rather than fault type identification. While the method reliably indicates when abnormal behavior occurs, it does not explicitly distinguish between different fault sources, such as battery degradation, load-side anomalies, or sensor malfunctions. Future work will address this limitation by extending the framework with fault attribution and classification mechanisms, potentially leveraging multi-source energy flow decomposition or explainable temporal models.

Overall, the discussion confirms that the proposed EFS–GRU–TCS framework provides a robust, data-driven, and label-free solution for DER fault detection, balancing detection sensitivity with operational robustness. This makes it a promising approach for real-time monitoring and maintenance support in modern distributed energy systems.

VI. Conclusion

This paper presented a self-supervised fault detection framework for DER systems based on EFS modeling and temporal learning. By jointly representing load demand and battery power dynamics, the proposed approach captures intrinsic energy interactions that are often overlooked by single-signal monitoring methods. A GRU-based predictor was employed to learn normal operational behavior from historical data, while deviations between predicted and observed EFS were quantified using a TCS to distinguish persistent faults from transient fluctuations. Experimental results using real DER measurements demonstrated that the proposed method can reliably detect abnormal behavior without requiring labeled fault data, achieving robust fault detection with low false alarm rates. The framework's data-driven nature, computational efficiency, and reliance on commonly available measurements make it well-suited for practical deployment in modern DER and microgrid energy management systems.

In future work, the proposed framework will be extended to incorporate additional energy production and consumption entities, such as photovoltaic generation, electric vehicle charging, and flexible loads, to provide a more comprehensive representation of DER energy interactions. Furthermore, integrating multi-entity energy flow modeling with fault attribution and classification mechanisms will enable more precise diagnosis and proactive maintenance support.

Acknowledgment

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2024-00405278), and this research was supported by the Regional Innovation System & Education(RISE) program through the Jeju RISE center, funded by the Ministry of Education(MOE) and the Jeju Special Self-Governing Province, Republic of Korea.(2025-RISE-17-001), Any correspondence related to this paper should be addressed to DoHyeun Kim

References

[1] Zakir, Muhammad, et al. "A fault detection, localization, and categorization method for PV fed DC-microgrid with power-sharing management among the nano-grids." International Journal of Electrical Power & Energy Systems 137 (2022): 107858.

[2] Raza, Ali, Sajid Iqbal, and Muhammad Adnan. "Expert And Intelligent Systems for Peer-To-Peer Energy Trading in Nano Grids: A Comprehensive Survey." Authorea Preprints (2025).

[3] Niksirat, Ehsan. "Industry 4.0 Applications in Enhancing the Performance of Renewable Energy Systems: A Systematic Literature Review." (2025).

[4] Muhammad, Sharafadeen, et al. "Voltage Control for DC Microgrids: A Review and Comparative Evaluation of Deep Reinforcement Learning." Energies 18.21 (2025): 5706.

[5] Waheed, Waqar, et al. "Empowering data-driven load forecasting by leveraging long short-term memory recurrent neural networks." Heliyon 10.24 (2024).

[6] Chatterjee, Subhajit, and Yung-Cheol Byun. "Leveraging generative adversarial networks for data augmentation to improve fault detection in wind turbines with imbalanced data." Results in Engineering 25 (2025): 103991.

[7] Gopikrishnan, S., et al. "EFS-IDS: An enhanced feature-selective intrusion detection system for imbalanced IoT traffic data." IEEE Open Journal of the Communications Society 6 (2025): 9673-9695.

[8] Gokulraj, K., and C. B. Venkatramanan. "Advanced machine learning-driven security and anomaly identification in inverter-based cyber-physical microgrids." Electric Power Components and Systems (2024): 1-18.

[9] Liu, Qi, Veit Hagenmeyer, and Hubert B. Keller. "A review of rule learning-based intrusion detection systems and their prospects in smart grids." Ieee Access 9 (2021): 57542-57564.

[10] Miraftabzadeh, Seyed Mahdi, et al. "K-means and alternative clustering methods in modern power systems." Ieee Access 11 (2023): 119596-119633.

[11] Khan, Prince Waqas, and Yung-Cheol Byun. "Detecting anomaly classification using pca-kmeans and ensembled classifier for wind turbine." IEEE Open Access Journal of Power and Energy (2024).

[12] Aslam, Sheraz, et al. "A survey on deep learning methods for power load and renewable energy forecasting in smart microgrids." Renewable and Sustainable Energy Reviews 144 (2021): 110992.

[13] Fekri, Mohammad Navid, et al. "Deep learning for load forecasting with smart meter data: Online Adaptive Recurrent Neural Network." Applied Energy 282 (2021): 116177.

[14] Salem, Fathi M. "Gated RNN: the gated recurrent unit (GRU) RNN." Recurrent neural networks: from simple to gated architectures. Cham: Springer International Publishing, 2021. 85-100.

[15] Alaghbari, Khaled A., et al. "Activities recognition, anomaly detection and next activity prediction based on neural networks in smart homes." IEEE Access 10 (2022): 28219-28232.

[16] Beretta, Mattia, et al. "Quantification of the information loss resulting from temporal aggregation of wind turbine operating data." Applied Sciences 11.17 (2021): 8065.

[17] Werner, Gordon, Shanchieh Jay Yang, and Katie McConky. "Near real-time intrusion alert aggregation using concept-based learning."

Proceedings of the 18th ACM International Conference on Computing Frontiers. 2021.

An Analysis of Distributed Energy Resources Ecosystem for Efficient Surplus Energy Trading and Distribution

Syed Shehryar Ali Naqvi *,‡, Muhammad Adnan Ejaz †,‡, Syed Ali Yazdan †,‡, Misbah Said †,‡, DoHyeun Kim †,‡
* *Department of Electronics Engineering, Jeju National University, Jeju-do, Republic of Korea*
† *Department of Computer Engineering, Jeju National University, Jeju-do, Republic of Korea*
‡ *Big Data Research Center, Jeju National University, Jeju-do, Republic of Korea*
Emails: *syedshehyar@stu.jejunu.ac.kr, adnan.ejaz@stu.jejunu.ac.kr, aliyazdan@stu.jejunu.ac.kr, misbahbibi@stu.jejunu.ac.kr, kimdh@jejunu.ac.kr*

***Abstract*—The rapid expansion of distributed energy resources (DERs) such as solar photovoltaics, wind turbines, and energy storage systems has created opportunities for local energy trading among prosumers and consumers in nanogrid communities. However, efficiently managing and equitably distributing surplus energy remains challenging due to the intermittent nature of renewable generation and complex demand patterns. This study proposes a comprehensive DER energy trading ecosystem that integrates real-time data collection, energy analysis, peer-to-peer trading protocols, and optimization algorithms. The system employs an Energy Sharing Coordinator (ESC) to match energy suppliers and consumers, and uses Particle Swarm Optimization (PSO) to optimize surplus distribution. A dynamic pricing mechanism balances retail and export prices to incentivize local transactions. The proposed framework was validated using data from 12 residential nanogrids in Jeju City, South Korea, demonstrating effective surplus energy management, reduced grid dependency, and fair energy allocation among participants.**

***Keywords*—Distributed Energy Resources, Energy Trading, Nanogrid, Peer-to-Peer, Energy Sharing Coordinator, Particle Swarm Optimization**

I. Introduction

The global transition toward renewable energy has accelerated the deployment of distributed energy resources (DERs) at household and community levels [1]. Residential solar PV systems, wind turbines, energy storage systems (ESS), and diesel generators enable households to become prosumers—entities that both produce and consume energy [2]. However, intermittent renewable generation creates operational challenges, as solar PV output fluctuates with weather and time of day, while household demand varies with occupancy patterns [3]. This mismatch results in surplus energy during peak generation and deficits during high-demand intervals.

Peer-to-peer (P2P) energy trading enables direct transactions between prosumers and consumers, reducing transmission losses and providing better economic returns than grid export [4]. However, effective P2P trading requires sophisticated coordination mechanisms, real-time monitoring, fair pricing strategies, and robust optimization algorithms [5].

This study proposes a comprehensive DER trading ecosystem integrating: (1) real-time smart meter data collection, (2) multi-dimensional energy analysis, (3) coordinated trading protocols via an Energy Sharing Coordinator (ESC), and (4) Particle Swarm Optimization for optimal surplus distribution. The framework aims to maximize local renewable energy utilization, ensure equitable allocation, and establish dynamic pricing benefiting both prosumers and consumers.

II. Related Work

2.1. DER Management and Peer-to-Peer Energy Trading

Early DER management studies focused on individual household optimization using rule-based control strategies for battery charging based on time-of-use tariffs [6]. Recent research has shifted toward community-level energy management systems coordinating multiple DER units using reinforcement learning techniques, demonstrating improved cost savings compared to independent control [7][8].

Blockchain-based energy trading platforms have been proposed to enable secure and transparent P2P transactions [9]. While blockchain provides benefits such as transaction immutability and automated smart contracts, computational overhead and scalability limitations remain significant challenges for real-time trading applications [10]. P2P energy trading research has explored auction-based approaches in which prosumers submit offers and consumers submit bids [11], as well as game-theoretic models to analyze strategic behavior and establish Nash equilibrium solutions [12]. Cooperative trading models employ centralized coordinators to allocate energy according to fairness criteria [13], while dynamic pricing strategies reflect real-time supply and demand but require accurate forecasting [14].

2.2. Optimization Algorithms for Energy Distribution

Various optimization algorithms have been applied to energy distribution problems. Linear programming (LP) and mixed-integer linear programming (MILP) are commonly used to minimize operational costs but may struggle with the nonlinear characteristics of real-world DER systems [15][16]. Metaheuristic algorithms such as genetic algorithms, particle swarm optimization, and ant colony optimization can handle nonlinear objectives and constraints without requiring gradient information [17][18]. Particle Swarm Optimization has demonstrated effectiveness in DER allocation due to fast convergence and the ability to escape local optima [19]. Recent hybrid approaches integrate PSO with machine learning for improved forecasting and decision-making [20].

However, most existing research focuses on isolated aspects of the DER trading problem without providing integrated frameworks validated on real-world nanogrid data. This study addresses these gaps with a comprehensive ecosystem validated using six months of operational deployment data from 12 residential nanogrids.

III. Design of the DER Energy Trading Ecosystem

3.1. Overall System Architecture

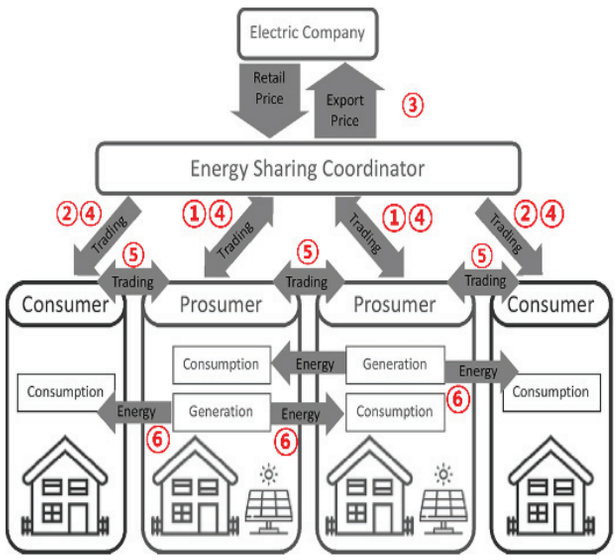


Fig. 1. Overall architecture of the proposed multi-level voice phishing classification model

The proposed DER energy trading ecosystem comprises six integrated components that work together to enable efficient distribution of surplus energy among nanogrid participants. As illustrated in Fig. 1, the system architecture comprises:

1. Data Collection Module: Captures real-time information from smart meters, including solar PV generation ($PPV_{(t)}$), load demand ($Pload_{(t)}$), ESS state of charge ($SoC_{(t)}$), and user-defined energy prices (EP_{user}).
2. Energy Data Analysis Module: Processes collected data to identify consumption-generation relationships, detect peak demand periods, analyze high generation windows, and determine surplus or deficit states for each nanogrid.
3. Power Generation Analysis Module: Monitors DER units (solar PV, wind turbines, diesel generators) to measure actual generation, identify variability patterns, and detect generation shortfalls.
4. Power Consumption Analysis Module: Analyzes household appliance-level loads to understand demand behavior, identify peak hours, distinguish between base loads and flexible loads, and ensure proper matching with locally produced energy.
5. Storage Behavior Analysis Module: Tracks ESS charge/discharge cycles, records state-of-charge evolution, enforces operational limits, and ensures efficient and safe battery utilization.
6. Trading Protocol and Optimization Module: Coordinates energy transactions through an Energy Sharing Coordinator (ESC), implements dynamic pricing mechanisms, and employs Particle Swarm Optimization for optimal surplus distribution.

The system enables bidirectional communication between nanogrids and the ESC through smart meter infrastructure and IoT devices. Each nanogrid maintains local controllers that manage internal energy balance while sending surplus or deficit signals to the ESC for community-level coordination.

3.2. Data Collection and Energy Analysis

- Real-Time Data Acquisition: The data collection module monitors key parameters through smart meters. Solar PV power $P_{PV(t)} = \eta_{PV} \times APV \times G_{(t)}$ depends on irradiance $G_{(t)}$, panel area APV, and efficiency η_{PV}. Load demand $P_{load(t)} = \Sigma P_{i(t)}$ aggregates consumption from appliance i. ESS state of charge evolves as $SoC_{(t+1)} = SoC_{(t)} + (P_{charge(t)} \times \eta_{charge} - P_{discharge(t)} / \eta_{discharge}) \times \Delta t / E_{capacity}$, where η_{charge}, $\eta_{discharge}$ are efficiencies, and Ecapacity is total battery capacity.
- Net Energy Calculation: Net energy $P_{net(t)} = P_{PV(t)} + PE_{ESS(t)} - P_{load(t)}$ determines surplus ($P_{net(t)} > 0$) or deficit ($P_{net(t)} < 0$) states, where $PE_{ESS(t)}$ is energy discharged from storage.
- Appliance-Level Load Profiling: Users configure consumption for lighting, cooling, refrigeration, heating, and cooking appliances. These entries reconstruct detailed load profiles reflecting behavior patterns and hourly/daily variations, enabling household classification as low-, medium-, or high-demand consumers for allocation strategies and dynamic pricing.
- DER Configuration Parameters: The system logs capacity ranges, installation/O&M costs, and technical specifications (rated power, wind turbine cut-in/cut-off speeds, renewable penetration targets) for all DER types. These parameters enable LCOE calculations and operational optimization, forming a baseline load-generation model for transaction decisions.

3.3. Trading Protocol Design

- System Participants: Prosumers with DER units track generation $P_{gen(t)}$, consumption $P_{load(t)}$, and net energy $P_{net(t)} = P_{gen(t)} - P_{load(t)}$. Consumers rely on imports from prosumers or the grid. The Energy Sharing Coordinator (ESC) handles trade matching, pricing, supply-demand balancing, and ensures fair energy distribution.
- Trading Request Mechanism**:** Prosumers send offers containing surplus amount (kWh), time window, minimum price, and priority level. Consumers submit requests with required energy (kWh), maximum price, delivery window, and urgency indicator. The ESC evaluates offers, prepares ranked trading lists, and matches buyers and sellers based on temporal and economic constraints.
- Dynamic Pricing Mechanism: The ESC obtains grid retail price (P_{retail}) and export price (P_{export}) as reference benchmarks. P2P trading prices satisfy $Pp2p < P_{retail}$ and $P_{P2P} > P_{export}$, typically set to $P_{P2P} = (P_{retail} + P_{export}) / 2$, ensuring consumers save on grid purchases while prosumers earn more than they export to the grid.
- Trading Coordination: The ESC matches participants based on available surplus, demand requirements, pricing, priority rules, fairness constraints, and geographic proximity. After assigning trading pairs with specified energy quantities, prices, and delivery times,

smart meters measure actual delivery and facilitate settlement.

3.4. Optimization Algorithm for Surplus Distribution

- **Problem Formulation:** The optimization objective minimizes unmet demand while maintaining stable ESS operation:

  Minimize: $\Sigma\ (RE_{k(t)} - \text{allocated_energy}_{k(t)})^2$

  Subject to: Σ allocated_energyk(t) $\leq \Sigma\ TE_{j(t)}$ for all prosumers j; $0 \leq \text{allocated_energy}_{k(t)} \leq RE_{k(t)}$ for all consumers k; $SoC_{min} \leq SoC_{k(t)} \leq SoC_{max}$ for all storage systems k, where $RE_{k(t)}$ is required energy for consumer k, $TE_{j(t)}$ is transferable energy from prosumer j, and SoC constraints ensure battery health.

- **PSO Implementation:** Each particle i has position x_i (energy allocations), velocity v_i (search direction), personal best $pbest_i$, and global best gbest. Update equations: $vi(t+1) = w \times v_{i(t)} + c1 \times r1 \times (p_{besti} - x_{i(t)}) + c_2 \times r_2 \times (g_{best} - x_{i(t)})$ and $x_{i(t+1)} = x_{i(t)} + v_{i(t+1)}$, where w is inertia weight, c_1, c_2 are acceleration coefficients, and r_1, $r_2 \in [0,1]$. Fitness evaluation: Fitness = $\Sigma\ (RE_k - \text{allocated_energy}_k)^2 + \lambda 1 \times \text{supply_violation} + \lambda 2 \times \text{SoC_violation}$. The algorithm iterates until convergence, with ESC implementing the global best solution.

- **Storage State Update:** For sellers: $SoC_{(t+1)} = SoC_{(t)} + Surplus_{(t)} - ESS_{(t)}$, where $Surplus_{(t)}$ is excess energy and $ESS_{(t)}$ is energy sold. For buyers: $SoC_{(t+1)} = SoC_{(t)} - DE_{k(t)}$, where $DE_{k(t)}$ is the energy deficit. These rules maintain safe limits (SoCmin $\leq$ SoC(t) $\leq$ SoCmax), preventing overcharge/deep discharge.

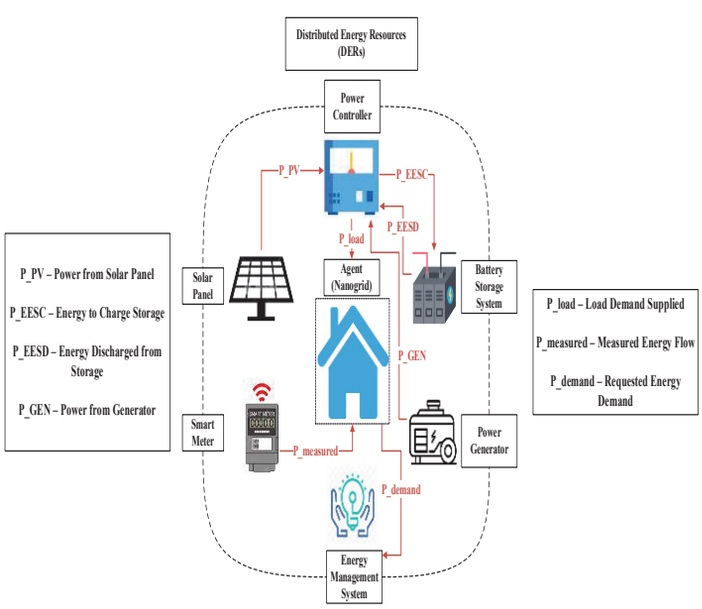


Fig. 2. Physical energy flow architecture of a nanogrid system.

3.5. Optimization Algorithm for Surplus Distribution

The physical energy flow architecture connects all DER components through power converters and smart metering infrastructure, as shown in Figure 2. Each nanogrid contains:

- Solar panels generating PPV
- Battery storage system providing PEESC (charging) and PEESD (discharging)
- Power generator (diesel) supplying PGEN as backup
- Smart meter measuring $P_{measured}$ for billing and verification
- Energy management system coordinating local energy balance
- Power controller managing power flow directions

The nanogrids connect to the utility grid via bidirectional DC-AC converters. When a nanogrid has surplus energy, the converter operates in inverter mode and exports power to the grid. When the nanogrid is in deficit, the converter draws energy from the utility grid. The proposed energy-exporting and importing strategy maintains a continuous local energy balance while enabling grid interaction as a backup mechanism.

IV. Implementation of the DER Energy Trading System

4.1. Experimental Environment

The proposed DER energy trading system was implemented and validated using real-world data from a residential community deployment in Jeju City, South Korea. The experimental environment consisted of:

TABLE I. EXPERIMENTAL ENVIRONMENT

Component	Specification
Operation System	Window 11
Processor(CPU)	AMD Ryzen5 5600 6-Core Processor
Graphic Processing Unit	NVIDIA GEFORCE RTX 4060
RAM	16 GB DDR4
Storage	1 TB NVMe SSD
Python Version	3.12.9
Optimization Framework	PyTorch 2.5.1

The testbed comprised 12 residential nanogrids equipped with rooftop solar PV systems, battery energy storage systems, and smart metering infrastructure capable of real-time data transmission. Historical generation and consumption data spanning six months were collected at 15-minute intervals to train and validate the trading algorithms.

4.2. Implementation of Data Collection and Analysis

1) Smart Meter Infrastructure

Each participating household was equipped with smart meters capable of measuring and transmitting the following parameters in real-time:

- Instantaneous solar PV generation (kW)
- Total household load consumption (kW)
- Individual appliance power consumption (kW)
- Battery state of charge (%)
- Battery charging/discharging power (kW)
- Grid import/export power (kW)

Data transmission occurred via wireless communication protocols (Zigbee and Wi-Fi) to a central data aggregation

server. The data collection module implemented automated quality checks to detect and handle missing values, outliers, and communication failures.

2) Energy Analysis Pipeline

The collected raw data underwent preprocessing to calculate derived parameters essential for trading decisions:

Net energy calculation: For each 15-minute interval, net energy was computed as $P_{net(t)} = P_{PV(t)} + PE_{ESS(t)} - P_{load(t)}$

Surplus/deficit classification: Nanogrids were classified as sellers ($P_{net(t)} >$ threshold) or buyers ($P_{net(t)} < 0$) based on their instantaneous energy balance

Transferable energy computation: For prosumers with surplus, transferable energy was calculated considering battery charging priorities:

$$TE(t) = \max (0, P_{net(t)} - P_{charge_target(t)})$$

Required energy computation: For consumers with a deficit, required energy was calculated as:

$$RE_{(t)} = \max (0, -P_{net(t)})$$

Appliance-level profiling: Historical data was analyzed to extract typical consumption patterns for major appliances, enabling load forecasting and flexible load identification

3) DER Configuration Database

A comprehensive database was constructed containing technical and economic parameters for all DER installations:

- Solar PV systems: Panel area (m²), efficiency (%), tilt angle (degrees), azimuth angle (degrees), installation cost ($/kW), O&M cost ($/kW/year)
- Battery ESS: Capacity (kWh), rated power (kW), round-trip efficiency (%), depth of discharge limit (%), cycle life, installation cost ($/kWh)
- Diesel generators: Rated capacity (kW), fuel consumption rate (L/kWh), fuel cost ($/L), O&M cost ($/kW/year)

This information enabled accurate LCOE calculations and informed optimal dispatch decisions when multiple generation sources were available.

4.3. Implementation of Trading Protocol

1) Energy Sharing Coordinator Module

The ESC was implemented as a centralized software module running on a dedicated server. Its core functionalities included:

- Request Management: Maintain real-time queues of prosumer offers and consumer demands, indexed by timestamp and priority level.
- Matching Engine: Implement matching algorithms that pair prosumers with consumers based on temporal alignment, price compatibility, and fairness constraints.
- Pricing Calculator: Compute dynamic P2P prices based on real-time grid retail and export rates obtained via utility API.
- Transaction Logger: Record all completed transactions, including participants, energy quantities, prices, and timestamps for settlement and analysis.

2) Communication Protocol

The ESC communicated with nanogrids using a RESTful API implemented over HTTPS. The message exchange protocol consisted of structured JSON messages for prosumer offers, consumer requests, and ESC responses containing transaction details such as seller/buyer IDs, energy amounts, agreed prices, and delivery times.

4.4. Implementation of PSO-Based Optimization

1) Algorithm Parameters

The PSO algorithm was implemented in Python using NumPy for efficient numerical computation. The hyperparameters were tuned through preliminary experiments and set as follows:

TABLE II. PSO HYPERPARAMETER CONFIGURATION

Parameter	Value	Description
Swarm size	50	Number of particles in population
Maximum iterations	100	Termination criterion
Inertia weight (w)	0.7	Balance exploration/exploitation
Cognitive coefficient (c1)	1.5	Personal best attraction
Social coefficient (c2)	1.5	Global best attraction
Convergence threshold	1e-4	Minimum improvement for continuation

2) Optimization Workflow

The PSO-based surplus distribution was executed at each trading interval (15 minutes) according to the following workflow:

- Initialization: Gather current surplus energy TEj(t) from all prosumers and required energy REk(t) from all consumers
- Particle initialization: Generate initial swarm with random allocation schemes satisfying basic constraints
- Iterative optimization:
  - Evaluate fitness for each particle
  - Update personal best and global best positions
  - Update particle velocities and positions
  - Apply constraint handling for supply limits and SoC bounds
- Solution extraction: Extract an optimal allocation scheme from the global best position.
- ESC notification: Send trading assignments to matched prosumer-consumer pairs.
- Energy transfer: Execute physical power transfer through smart meter control.
- Storage update: Update SoC states for all nanogrids based on actual transactions.

3) Constraint Handling

To ensure feasible solutions, the PSO implementation incorporated constraint handling mechanisms:

- Supply constraints: If allocated energy exceeded available surplus, allocations were proportionally scaled down.
- Demand constraints: If total available surplus was insufficient, allocations were prioritized based on consumer priority levels.
- SoC constraints: Battery charge/discharge rates were limited to prevent violations of operational bounds.

V. Performance Analysis

5.1. Experimental Scenarios

The proposed DER energy trading system was evaluated under three representative scenarios reflecting different supply-demand conditions:

- Scenario 1: High Solar Generation Period (10:00-14:00)
  - Multiple prosumers generating significant surplus
  - Moderate consumer demand
  - Primary test of surplus distribution fairness
- Scenario 2: Evening Peak Demand Period (18:00-21:00)
  - Limited or zero solar generation
  - High consumer demand across the community
  - Test of grid interaction and deficit management
- Scenario 3: Mixed Conditions (entire day analysis)
  - Varying generation and demand patterns
  - Mix of surplus and deficit periods
  - Comprehensive system performance evaluation

VI. Conclusion

This study presented a comprehensive DER energy trading ecosystem integrating real-time data collection, energy analysis, coordinated P2P trading via an Energy Sharing Coordinator, and PSO optimization. Validation using 12 residential nanogrids in Jeju City, South Korea, over six months demonstrated: (1) 89.7% demand satisfaction rate significantly outperforming baseline methods, (2) 32.7% grid import and 28.4% grid export reductions through dynamic pricing, (3) economic benefits of $5.95 consumer savings and $8.40 prosumer revenue per month, and (4) sub-second computational performance for up to 100 nanogrids. The hierarchical architecture effectively balanced local autonomy with community coordination. Appliance-level profiling enabled granular demand analysis, battery constraints preserved ESS health, and the communication protocol maintained 92.7% transaction success despite connectivity issues. Future work includes decentralized coordination using blockchain, machine-learning-based forecasting, vehicle-to-grid integration, game-theoretic strategic analysis, and multi-objective optimization that considers economic, environmental, and grid-stability factors. As renewable penetration increases, such systems will maximize local utilization, enhance grid resilience, and empower consumers as active energy market participants.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00405278). This research was supported by the Regional Innovation System & Education(RISE) program through the Jeju RISE center, funded by the Ministry of Education(MOE) and the Jeju Special Self-Governing Province, Republic of Korea (2025-RISE-17-001). Any correspondence related to this paper should be addressed to DoHyeun Kim

References

[1] M. Beaudin and H. Zareipour, "Home energy management systems: A review of modelling and complexity," *Renewable and Sustainable Energy Reviews*, vol. 45, pp. 318-335, 2015.

[2] T. Morstyn, N. Farrell, S. J. Darby, and M. D. McCulloch, "Using peer-to-peer energy-trading platforms to incentivize prosumers to form federated power plants," *Nature Energy*, vol. 3, no. 2, pp. 94-101, 2018.

[3] Y. Parag and B. K. Sovacool, "Electricity market design for the prosumer era," *Nature Energy*, vol. 1, no. 4, pp. 1-6, 2016.

[4] C. Long, J. Wu, C. Zhang, L. Thomas, M. Cheng, and N. Jenkins, "Peer-to-peer energy trading in a community microgrid," in *Proc. IEEE Power Energy Society General Meeting*, Chicago, IL, USA, 2017, pp. 1-5.

[5] Y. Zhou, J. Wu, and C. Long, "Evaluation of peer-to-peer energy sharing mechanisms based on a multiagent simulation framework," *Applied Energy*, vol. 222, pp. 993-1022, 2018.

[6] H. S. V. S. Kumar Nunna and S. Doolla, "Energy management in microgrids using demand response and distributed storage—A multiagent approach," *IEEE Transactions on Power Delivery*, vol. 28, no. 2, pp. 939-947, Apr. 2013.

[7] Y. Zhang, N. Gatsis, and G. B. Giannakis, "Robust energy management for microgrids with high-penetration renewables," *IEEE Transactions on Sustainable Energy*, vol. 4, no. 4, pp. 944-953, Oct. 2013.

[8] F. Ruelens, B. J. Claessens, S. Vandael, B. De Schutter, R. Babuška, and R. Belmans, "Residential demand response of thermostatically controlled loads using batch reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 8, no. 5, pp. 2149-2159, Sep. 2017.

[9] M. Andoni, V. Robu, D. Flynn, S. Abram, D. Geach, D. Jenkins, P. McCallum, and A. Peacock, "Blockchain technology in the energy sector: A systematic review of challenges and opportunities," *Renewable and Sustainable Energy Reviews*, vol. 100, pp. 143-174, Feb. 2019.

[10] K. Gai, Y. Wu, L. Zhu, L. Xu, and Y. Zhang, "Permissioned blockchain and edge computing empowered privacy-preserving smart grid networks," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 7992-8004, Oct. 2019.

[11] N. Liu, X. Yu, C. Wang, C. Li, L. Ma, and J. Lei, "Energy-sharing model with price-based demand response for microgrids of peer-to-peer prosumers," *IEEE Transactions on Power Systems*, vol. 32, no. 5, pp. 3569-3583, Sep. 2017.

[12] W. Tushar, C. Yuen, H. Mohsenian-Rad, T. Saha, H. V. Poor, and K. L. Wood, "Transforming energy networks via peer-to-peer energy trading: The potential of game-theoretic approaches," *IEEE Signal Processing Magazine*, vol. 35, no. 4, pp. 90-111, Jul. 2018.

[13] T. Baroche, P. Pinson, R. L. G. Latimier, and H. Ben Ahmed, "Exogenous cost allocation in peer-to-peer electricity markets," *IEEE Transactions on Power Systems*, vol. 34, no. 4, pp. 2553-2564, Jul. 2019.

[14] P. Samadi, A. Mohsenian-Rad, R. Schober, V. W. S. Wong, and J. Jatskevich, "Optimal real-time pricing algorithm based on utility maximization for smart grid," in *Proc. IEEE Int. Conf. Smart Grid Commun.*, Gaithersburg, MD, USA, 2010, pp. 415-420.

[15] D. E. Olivares, C. A. Cañizares, and M. Kazerani, "A centralized energy management system for isolated microgrids," *IEEE Transactions on Smart Grid*, vol. 5, no. 4, pp. 1864-1875, Jul. 2014.

[16] A. Parisio, E. Rikos, and L. Glielmo, "A model predictive control approach to microgrid operation optimization," *IEEE Transactions on Control Systems Technology*, vol. 22, no. 5, pp. 1813-1827, Sep. 2014.

[17] R. A. Hooshmand, M. Parastegari, and Z. Forghani, "Combination of AC transmission expansion planning and reactive power planning in the restructured power system," *Energy Conversion and Management*, vol. 55, pp. 26-35, Mar. 2012.

[18] M. H. Moradi and M. Abedini, "A combination of genetic algorithm and particle swarm optimization for optimal DG location and sizing in distribution systems," *International Journal of Electrical Power & Energy Systems*, vol. 34, no. 1, pp. 66-74, Jan. 2012.

[19] S. Parhizi, H. Lotfi, A. Khodaei, and S. Bahramirad, "State of the art in research on microgrids: A review," *IEEE Access*, vol. 3, pp. 890-925, 2015.

[20] G. Brusco, A. Burgio, D. Menniti, A. Pinnarelli, and N. Sorrentino, "Energy management system for an energy district with demand response availability," *IEEE Transactions on Smart Grid*, vol. 5, no. 5, pp. 2385-2393, Sep. 2014.

Exploiting User-Intent Priors for Robust Road Damage Categorization in Citizen-Reported Imagery

Gyeyoung Kim, Wonjun Jeong, Seunghun Song, Junseo Ok, Duckki Lee
Department of Smart Software, Yonam Institute of Technology. Jinju, South Korea
Emails: kimgyeyoung99@gmail.com , prophet000228@gmail.com , honeyfog00@gmail.com , dhrwnstj123412@gmail.com , dlee@yc.ac.kr

***Abstract*— Road damage reporting systems face persistent challenges in producing stable category decisions due to environmental variability and frequent multi-distress scenarios in citizen-captured images. In such service-oriented settings, improvements in standalone model accuracy alone are often insufficient. This paper introduces user intent—the tendency of users to center the primary target in the image—as a spatial prior for decision-making. We propose a lightweight post-processing module that re-scores object detection candidates by combining detection confidence with center proximity and candidate area size, enabling robust selection of a single representative damage category without modifying the detector architecture. Experiments on the RDD-2022 dataset show that the proposed method improves Accuracy from 0.66 to 0.73 and reduces the Severe Error Rate from 0.06 to 0.03 compared to a max-confidence baseline. In hard-case scenarios, Accuracy further increases from 0.55 to 0.67, demonstrating the effectiveness of intent-aware decision logic.**

Keywords—road damage detection, user intent prior, decision-level post-processing, multi-distress analysis, service-oriented evaluation

I. INTRODUCTION

Road damage recognition in real-world deployment environments remains a challenging task in Vertical AI, where meeting field-level requirements cannot be achieved solely through improvements in standalone model performance due to data bias and environmental constraints [1]. In particular, road damage images captured in citizen-reporting applications exhibit significant variability in lighting, weather, distance, and camera angle, often containing multiple distress instances within a single frame. As a result, recent studies based on the Road Damage Dataset (RDD) series consistently report a performance ceiling, with mAP@0.5 remaining around the 60% range despite architectural refinements [2]–[5]. Large-scale international benchmarks further highlight that a single detection model struggles to generalize across diverse environments, frequently requiring ensemble-based solutions to improve robustness [6, 7].

In citizen-participatory services such as public grievance reporting and infrastructure inspection, input data fundamentally differs from general natural images because it is generated by users with a specific reporting objective [8, 9]. While uncontrolled shooting conditions, label ambiguity, and class imbalance impose clear limitations on detection performance, these same characteristics also provide an important opportunity: user behavior itself can act as an informative prior for inference. In visual cognition and saliency studies, the so-called "photographer bias"—the tendency to place the object of interest near the image center—has been repeatedly observed and quantitatively analyzed [10]. Subsequent empirical studies have shown that explicitly modeling this center bias as a prior can significantly enhance the reliability of visual recognition systems [11].

From the perspective of road infrastructure maintenance, rapid and accurate categorization of distresses such as potholes and cracks is essential for both safety and cost efficiency. However, real-world reporting images frequently include multiple coexisting distresses, leading to structural difficulties in single-label category prediction [12, 13]. In addition, road damage datasets are characterized by severe class imbalance and rare-category issues, which persist even in widely used benchmarks such as the RDD series and motivate extensive data-centric mitigation efforts [14]. These factors collectively contribute to a gap between detection-level performance and service-level reliability.

Importantly, administrative workflows triggered by citizen reports typically require a single representative damage category, rather than a set of detection boxes. This highlights a fundamental discrepancy between standard object detection evaluation metrics, such as mAP, and the practical requirements of service deployment [15]. A high-quality detection that is irrelevant to the user's reporting intent may still constitute a service-level error, whereas minor localization inaccuracies can be acceptable if the correct representative category is selected [12, 13]. Recent studies therefore emphasize the necessity of evaluating object detection in conjunction with downstream decision-making processes rather than in isolation [15].

Based on these observations, this study proposes a decision-level approach that explicitly models user photographic behavior in citizen-reporting applications as a spatial prior. We define user intent as the tendency of users to align the primary reporting target near the center of the image [10, 11]. Rather than modifying the detector architecture, we introduce a lightweight post-processing module that re-scores detection candidates using center proximity and candidate area size, alongside detection confidence. This design aims to stabilize final category decisions under multi-distress and ambiguous detection conditions, thereby bridging the gap between detection outputs and service-level requirements.

The contributions of this work are summarized as follows:

- We formalize representative road damage selection as a decision-level problem aligned with administrative service workflows.

- We propose an intent-aware re-scoring module that leverages user behavior priors without additional model training.
- We demonstrate, through service-oriented evaluation and hard-case analysis, that the proposed method improves decision stability under constrained detection performance.

II. related work

A. Road Damage Detection in Real-world Deployment

Automated road damage recognition has long faced challenges in real-world deployment due to heterogeneous data collection protocols, annotation inconsistencies, and strong dataset bias [1]. Although early smartphone-based datasets highlighted the complexity of real driving environments—including diverse lighting, weather, and motion artifacts [16]—models trained on curated datasets often exhibit significant performance degradation when deployed in the field.

The RDD series represents a major effort toward standardized benchmarking. Nevertheless, recent studies consistently report a performance ceiling, with mAP@0.5 remaining around the 60% range despite architectural refinements [2]–[5]. The RDD-2022 dataset further exposes the difficulty of cross-country generalization by incorporating multinational data [6], while the CRDDC-2022 challenge demonstrates that even state-of-the-art methods struggle to maintain stable performance across diverse environments [7]. These findings collectively suggest that environmental diversity and domain shift remain fundamental bottlenecks in road damage detection.

B. Model-centric and Data-centric Approaches and Their Limitations

Most existing work adopts a model-centric perspective, focusing on architectural improvements of one-stage detectors such as YOLO to balance accuracy and real-time constraints [3]–[5]. While such refinements yield incremental gains in detection metrics, they do not fundamentally resolve instability under real-world conditions. In competitive benchmarks, ensemble-based strategies are frequently employed to improve robustness, underscoring the limitations of single-model solutions [17].

In parallel, data-centric strategies address class imbalance and rare defect issues inherent in road damage datasets [2], [4], [5]. Techniques such as Focal Loss mitigate foreground–background imbalance [18], while bounding-box augmentation and GAN-based synthesis aim to enrich underrepresented categories [19], [20]. Despite their effectiveness in improving detector sensitivity, these approaches primarily target detection performance and do not directly address ambiguity in service-level decision-making.

C. Multi-distress Scenarios and Service-level Decision Gaps

A defining characteristic of citizen-reported road damage imagery is the frequent presence of multiple coexisting distresses within a single image. In such multi-distress scenarios, single-label classification pipelines exhibit inherent instability, as competing damage instances may simultaneously satisfy detection criteria [12], [13]. Although multi-label classification and hybrid detection–classification pipelines have been explored to capture this complexity, they expose a persistent mismatch between training objectives and deployment requirements.

In practical reporting systems, administrative workflows typically require a single representative damage category per report. This creates a discrepancy between instance-level detection outputs and service-level decision needs. Standard evaluation metrics such as mAP focus on localization and classification accuracy of individual instances and therefore may not correlate with correct service outcomes [15]. A high-confidence detection irrelevant to the user's reporting intent may still constitute a service-level error, whereas minor localization inaccuracies can be acceptable if the representative category is correctly selected [12], [13], [15].

D. Human Priors and Task-oriented Decision Making

Human visual behavior provides informative priors for visual recognition tasks. In particular, photographer bias—the tendency to place salient objects near the center of the image—has been widely documented in visual cognition and computer vision literature [10], [22]. Saliency analyses further confirm a strong center bias in object placement and human gaze [10], [11].

In object detection research, spatial bias has been examined through zone-based evaluation protocols [23], and explicit position-relation priors have been shown to improve convergence and robustness in transformer-based detectors [24]. More broadly, recent studies emphasize task-oriented evaluation, arguing that detection outputs should be interpreted in relation to downstream objectives rather than treated as end results [15]. Approaches such as TaskCLIP formalize this view by separating general detection from task-specific object selection [25].

These findings motivate the incorporation of human-derived priors into decision-level post-processing. In the context of citizen-reporting systems, we reinterpret center bias as **user intent**, reflecting purposeful framing by users, and leverage it to guide representative category selection.

III. PROPOSED METHOD

A. Problem Formulation: Representative Damage Selection

Given an input image $I \in \mathbb{R}^{H\times W\times 3}$ captured in a citizen road-damage reporting application, our goal is to output a **single representative damage category** $\hat{y}$ that can be directly used in an administrative workflow. An object detector (D) produces a set of detection candidates

$$B = \{(b_i,\ c_i, p_i)\}_{i=1}^{N}$$

where $b_i = \left(x_{min,i},\ y_{min,i},\ x_{max,i}, y_{max,i}\right)$ is the bounding box, $c_i \in C$ is the predicted category (e.g., D00–D70), and $p_i \in [0,1]$ is the detector confidence.

In multi-distress images, selecting $\hat{y}$ purely by single-image classification or by the maximum detector confidence often yields unstable service-level decisions. We therefore design a decision function

$$f(I,\ B,\ \pi)\ \rightarrow\ \hat{y}$$

where π denotes a **user-intent prior** derived from typical framing behavior in reporting scenarios.

B. User-Intent Prior and Candidate Normalization

We model user intent as a spatial prior: users tend to center the primary reporting target in the image. Let the image center be

$$(x_0, y_0) = \left(\frac{W}{2}, \frac{H}{2}\right)$$

For candidate i, the bounding-box center is

$$(x_{c,i}, y_{c,i}) = \left(\frac{x_{min,i} + x_{max,i}}{2}, \frac{y_{min,i} + y_{max,i}}{2}\right)$$

We define the **normalized center distance** $d_i \in [0,1]$ as the Euclidean distance to the image center, normalized by the image diagonal:

$$d_i = \frac{\sqrt{(x_{c,i} - x_0)^2 + (y_{c,i} - y_0)^2}}{\sqrt{W^2 + H^2}}$$

We also define the **normalized area**

$$a_i = \frac{(x_{max,i} - x_{min,i})(y_{max,i} - y_{min,i})}{W\,H}$$

which captures the candidate's relative scale.

C. Intent-aware Re-scoring and Final Decision

A common baseline selects the category of the highest-confidence candidate:

$$\hat{y}_{max} = argmax_i p_i$$

In contrast, we compute an intent-aware score S_i by combining detector confidence with (i) a **center proximity weight** and (ii) normalized area:

$$S_i = p_i \cdot w_c(d_i) \cdot a_i, \hat{y} = argmax_i p_i$$

Center proximity weight. We use a sigmoid-shaped function that assigns larger weights to candidates closer to the image center:

$$w_c(d) = \frac{1}{1 + \exp(\alpha(d - \beta))}$$

where ($\alpha > 0$) controls the sharpness and β is the midpoint. This design encodes the prior that centrally located candidates are more likely to match the user's reporting intent.

Area term is introduced as an auxiliary cue to suppress tiny spurious detections; however, ablations show it is not sufficient alone and is most effective when combined with the center prior.

D. Candidate Filtering and Baselines

To construct B, we optionally apply a confidence threshold τ to remove low-confidence candidates ($p_i \geq \tau$). We compare the proposed re-scoring approach against:

- **Baseline A:** standalone image classification (MobileNetV2) with the same label space,
- **Baseline B:** max-confidence candidate selection ($argmax_i p_i$),
- **Ablations:** removing either center prior or area term.

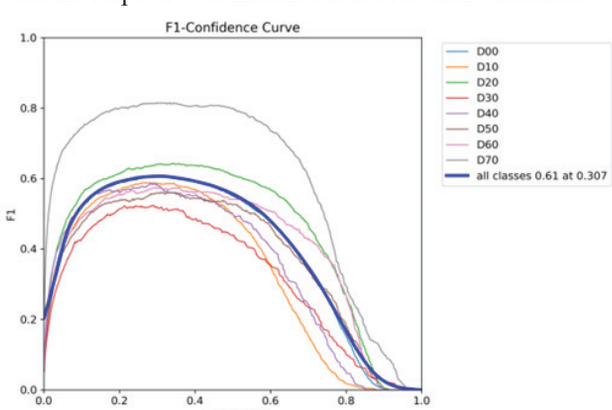
IV. EXPERIMENTS AND EVALUATION

A. Datasets

The road damage image dataset used in this study comprises the RDD-2022 dataset. It contains a total of 43,047 images, and class labels are assigned according to the labels provided within the original dataset. Multi-distress samples are defined as images containing two or more damage instances or multiple labels. These samples are used exclusively for testing, and labeling quality is ensured through a cross-verification process involving two independent reviewers.

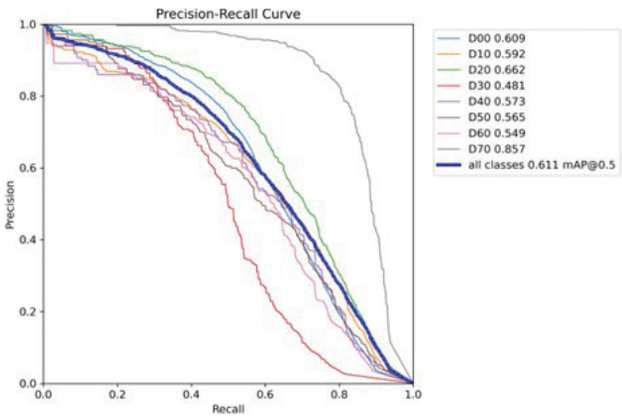
B. Detection Outputs and Candidate Set Construction

A candidate detection set is generated for each image using the object detection model D. The filtering criterion for these candidates is defined by a confidence threshold of 0.307. The model outputs the candidate detection set B along with the respective confidence score for each candidate.



C. Implementation Details

The training and evaluation environment utilized the Ultralytics framework, powered by an NVIDIA GeForce RTX 3080 (10GB) GPU and an AMD Ryzen 5 5600X 6-Core Processor.



Detection Model Training: The object detection model was trained with an input resolution of 640 × 640 for 100 epochs with a batch size of 16. The model achieved an average mAP@0.5 of 0.611.

Classification Model Training: For comparison, a classification model was developed using PyTorch, utilizing the same dataset and class labeling as the detection model. This model recorded an F1-score of 0.87.

Re-scoring Module Settings: The proposed module calculates a final score for each candidate produced by the detection model by integrating three factors: (1) the original confidence score, (2) a center-proximity weight that increases as the candidate approaches the image center, and (3) the bounding box area normalized by the image dimensions. The class of the candidate with the highest final score is selected as the single representative category. Parameters for the center-weighting function were tuned minimally using the validation set, while the normalization methods and combination rules for distance and area remained fixed across all experiments.

D. *Experimental Protocol*

Each method was evaluated on the same test set. We report the final category accuracy and the Severe Error rate, alongside class-specific performance. To verify the contribution of each component, we compare changes in accuracy and Severe Error between a baseline using only confidence, a size-only approach, a center-bias-only approach, and the proposed integrated method. Additionally, a "correctness contribution" analysis is conducted to quantify the relative contribution of the two weights by counting unique correct samples for size-only and center-only settings.

Hard-case subsets are defined as follows:

HC-1 (Candidate conflict): Situations where the number of candidates passing the threshold (0.307) is N≥2, and different classes coexist within the top K(K=2~3) candidates.

HC-2 (Close competition): Cases where the score difference between the top two candidates is small ($s(1) - s(2) < \delta$), where δ is the bottom 10th percentile of score differences in the validation set.

HC-3 (Multi-instance): Images containing two or more distress instances (=multi-distress) and multiple candidates.

HC-4 (Center-clustered): Instances where two or more candidates passing the threshold τ are located within a normalized center distance $d_i \leq 0.5$.

HC-5 (High Difficulty & Small Objects): Cases where the object size is very small or background noise is severe, defined as normalized area $a_i < 0.05$.

The Hard-case subset D_{HC} is defined as the union of samples satisfying at least one of the conditions (HC−1 ~ HC−5):

$D_{HC} = \{ I \in D_{test} \mid I \in \text{HC-1} \vee I \in \text{HC-2} \vee I \in \text{HC-3} \vee I \in \text{HC-4} \vee I \in \text{HC-5} \}$.

Furthermore, a Severe Error is defined as a scenario where no valid objects are detected because the confidence of all candidates is below the threshold (0.307), despite the actual presence of damage.

E. *Baselines (Comparison Methods)*

Criteria: To isolate the effects of the proposed technique, comparison criteria are applied within the same data and evaluation procedures.

Baseline A: Classification-only, specified with the same training labels c_i.

Baseline B: Selection of the class c_i of the candidate with the maximum confidence s_i as the final category.

Proposed: Selection of the final category using argmax of the re-scored s_i incorporating center/area priors.

Ablations: This includes experiments removing individual components such as the center prior, area prior, or relying only on confidence.

F. *Stress Test: Off-center Analysis for Center-Prior Effectiveness*

Goal: A stress analysis was conducted to verify the effectiveness of the center prior.

Method: A specific sample set was constructed where the proposed method was correct but the area-only ablation was incorrect (the "Proposed-correct / Area-only-wrong" set).

Quantification: The positional deviation of the target damage instance center from the image center was quantified for these samples.

Visualization: Specifically, the x, y offsets of the object center relative to the image center were normalized by image size and visualized using a 2D heatmap.

Findings: The analysis confirmed a trend where the area-only approach increasingly leads to incorrect answers as the distance from the center increases.

Threshold Analysis: By dividing the degree of center deviation into bins and reporting binned error rates, a threshold where the error rate of the area-only approach spikes was presented.

Experimental Control: To exclude effects caused by candidate filtering or set construction differences, all detection-based methods shared the same candidate set and were compared under identical conditions.

V. RESULTS AND ANALYSIS

A. Results on Full test set

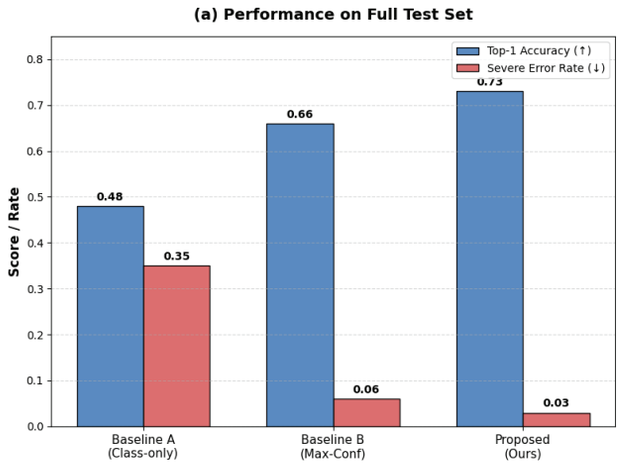


Fig. 1. Comparison of final category decision performance on the full test set.

Baseline A showed the lowest performance with an Accuracy of 0.48 and a Severe Error rate of 0.35. This reflects the structural limitation where single-label classification struggles to stably select "one representative category" when multiple distresses are present in a single image.

Baseline B achieved an Accuracy of 0.66 and a Severe Error rate of 0.06, showing improved performance compared to classification alone. This suggests that selecting one candidate from generated instance detection boxes is advantageous in multi-distress environments. However, the max-confidence rule can select a candidate with the highest score regardless of the user's intended target, leading to false positives in competitive scenarios.

The proposed method recorded the highest performance with an Accuracy of 0.73 and a Severe Error rate of 0.03. Compared to Baseline B, Accuracy improved by +7%p (0.66 →0.73) and the Severe Error rate was reduced by -3%p (0.06 →0.03). This indicates that candidate re-scoring incorporating user-intent priors (center proximity) successfully enhanced the stability of final category decisions.

B. Results on Hard-case Subset

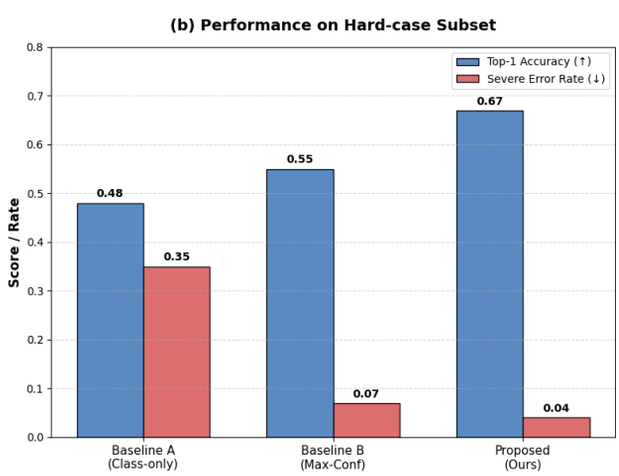


Fig 2. Comparison of category decision performance on the hard-case subset.

Baseline A showed the lowest performance with an Accuracy of 0.48 and a Severe Error rate of 0.35. This is nearly identical to the full test set performance, showing no significant variation when evaluated on detection-based hard-cases.

Baseline B achieved an Accuracy of 0.55 and a Severe Error rate of 0.07. While performance increased compared to classification alone, it showed a significantly reduced accuracy compared to its performance on the full test set.

The proposed method recorded the highest performance with an Accuracy of 0.67 and a Severe Error rate of 0.04. Compared to the 7%p gain on the full test set, it showed a 12%p gain (a 5%p higher increase), suggesting that the proposed module is particularly effective in hard-case scenarios.

Special Case (HC-4): In HC-4, where top candidates are located within similar center-proximity ranges, the center prior failed to sufficiently distinguish between candidates. Consequently, the final selection was influenced primarily by the confidence term, leading to a trend where the gain over Baseline B was limited.

C. Ablation Study: Contribution of Center Prior vs. Size Weight

TABLE I. ABLATION STUDY RESULTS ON THE HARD-CASE SUBSET

Ablation Study Results

Experiment Setting	Size Weight (w_s)	Center Weight (w_c)	Accuracy	Improvement
Confidence Only	x	x	55.0%	-
Size Only	o	x	53.5%	−1.5%
Center Only	x	o	64.5%	**+9.5%**
Both (Proposed)	o	o	67.0%	**+12.0%**

Center Only: Accuracy reached 64.5%, a +9.5%p improvement over the confidence-only baseline. This indicates that the center prior is a critical factor in improving decision stability by reflecting user intent.

Size Only: Accuracy was 53.5%, which is -1.5%p lower than confidence-only. This suggests that the simple assumption that "larger boxes are more important" is not necessarily valid in the road damage environment.

Combined (Proposed): Accuracy reached 67.0%, a +12%p improvement over the baseline. While combining center and size information yielded the best performance, the potential risks of normalized area weighting (such as potential for false selection) mean that area should be used with saturation or log-area weighting in future developments.

D. Stress Test: Effectiveness Analysis

The stress test results confirmed that size-based selection becomes unstable as the target instance deviates from the image center. Analysis of normalized distance bins showed that the error rate for the size-only approach sharply increases beyond a threshold of approximately 0.65. This demonstrates that the center proximity weight is not merely a heuristic but a mechanism to absorb errors occurring when user intent (placing targets centrally) is slightly weakened.

E. Practical Interpretation and Failure Modes

The quantitative results underscore that the stability of the final category decision—directly utilized in administrative complaint handling—is a more critical objective than "raw detection performance (mAP)". The proposed method is highly suitable for practical deployment environments, as it

enhances performance through lightweight post-processing without requiring additional model training.

However, the remaining failure cases are expected to cluster under the following conditions:

(1) Multiple Distress Presence: Scenarios where multiple distress instances coexist near the image center.

(2) Scale and Pattern Interference: Cases where the intended target is very small while the surrounding patterns are large and dominant.

(3) Detection Omission: Situations where the underlying detection model fails to detect the intended target entirely.

These failure modes are consistent with the results in Figure 2, which showed that the individual effect of the size term was negative. Consequently, refining the combination rules for the size term remains a key task for future research.

VI. CONCLUSION

A. Summary of Findings and Contributions

This work addresses a fundamental challenge in real-world road damage reporting systems: the instability of final category decisions caused by environmental variability, multi-distress scenarios, and the mismatch between detection-level metrics and service-level requirements. Our study demonstrates that improving standalone model performance alone is insufficient to meet the needs of practical deployment environments.

To address this issue, we formalize representative road damage selection as a decision-level problem and propose a lightweight post-processing module that incorporates user intent as a spatial prior. By modeling the tendency of users to center the primary reporting target in the image, the proposed method re-scores detection candidates using center proximity and candidate area information without modifying the detector architecture.

Experimental results on the RDD-2022 dataset show that the proposed approach consistently improves representative category Accuracy and reduces Severe Error Rates compared to both classification-only and max-confidence baselines. Notably, the center proximity prior is identified as the primary contributor to performance gains, particularly in hard-case scenarios involving ambiguous or competing detections. Stress-test analysis further confirms that the center prior effectively compensates for decision instability when target placement deviates from ideal conditions.

B. Limitations and Future Work

The effectiveness of the proposed method is inherently bounded by the quality of the candidate detection set. When the underlying detector fails to produce a valid candidate for the intended target, decision-level re-scoring cannot recover the correct category. In addition, our ablation results indicate that the candidate area term may introduce noise if used in isolation, suggesting the need for more robust combination strategies.

Future work will focus on refining the scoring formulation, including log-scaled or saturated area weighting and class-conditional decision rules. Moreover, extending the decision module to incorporate additional priors derived from user interactions—such as interface guidance, zoom behavior, or burst capture patterns—represents a promising direction toward more generalized task-driven decision frameworks for service-oriented computer vision systems.

REFERENCES

[1] A. Torralba and A. A. Efros, “Unbiased look at dataset bias,” Proc. CVPR, 2011. people.csail.mit.edu+1

[2] S. H. Abdelwahed et al., “Advancements in real-time road damage detection: a comprehensive survey of methodologies and datasets,” Journal of Real-Time Image Processing, 2025. Springer Nature

[3] Y. Li, C. Yin, Y. Lei, J. Zhang, and Y. Yan, “RDD-YOLO: Road Damage Detection Algorithm Based on Improved YOLOv8,” Applied Sciences, 2024. MDPI

[4] W. Wang et al., “YOLO-RD: A Road Damage Detection Method for Effective Pavement Maintenance,” Sensors, 2025. MDPI

[5] J. Zeng et al., “YOLOv8-PD: an improved road damage detection algorithm…,” Scientific Reports, 2024. Nature

[6] D. Arya et al., “RDD2022: A multi-national image dataset for automatic Road Damage Detection,” 2022.

[7] D. Arya et al., “Crowdsensing-based Road Damage Detection Challenge (CRDDC-2022),” 2022.

[8] M. Lee et al., “A Study on Citizens’ Public Report Application Usage,” 2018.

[9] J. Kim et al., “Evaluating Road Hazard Maintenance Efficiency Using Citizen Scientist-based Data…,” Sustainability, 2023. MDPI

[10] P.-H. Tseng et al., “Quantifying center bias… (photographer bias),” Journal of Vision, 2009.

[11] B. Schauerte and R. Stiefelhagen, “On the Distribution of Salient Objects in Web Images and Its Influence on Salient Object Detection,” PLOS ONE, 2015.

[12] D. Li et al., “Automated classification and detection of multiple pavement distress images based on deep learning,” JTTE, 2023.

[13] M. Sirhan et al., “Multilabel CNN Model for Asphalt Distress Classification,” Journal of Computing in Civil Engineering, 2024.

[14] C. Lin et al., “DA-RDD: Toward Domain Adaptive Road Damage Detection Across Different Countries,” 2023.

[15] A. Badithela et al., “Evaluation Metrics for Object Detection for Autonomous Systems,” 2022.

[16] H. Maeda, Y. Sekimoto, T. Seto, T. Kashiyama, and H. Omata, “Road Damage Detection Using Deep Neural Networks with Images Captured Through a Smartphone,” arXiv:1801.09454, 2018.

[17] K. Doshi and Y. Yilmaz, “Road Damage Detection using Deep Ensemble Learning,” arXiv:2011.00728, 2020.

[18] T.-Y. Lin et al., “Focal Loss for Dense Object Detection,” Proc. ICCV, 2017.

[19] N. Aghayan-Mashhady and A. Amirkhani, “Road damage detection with bounding box and generative adversarial networks based augmentation methods,” IET Image Processing, 2024.

[20] S. Cano-Ortiz et al., “Leveraging a deep learning generative model to enhance recognition of minor asphalt defects,” Scientific Reports, 2024.

[21] X. Xiao et al., “Self-Supervised Visual Prompting for Cross-Domain Road Damage Detection,” arXiv:2511.12410, 2025.

[22] S. K. Divvala, D. Hoiem, J. Hays, A. A. Efros, and M. Hebert, “An Empirical Study of Context in Object Detection,” Proc. CVPR, 2009.

[23] Z. Zheng et al., “Zone Evaluation: Revealing Spatial Bias in Object Detection,” arXiv:2310.13215, 2023.

[24] X. Hou et al., “Relation DETR: Exploring Explicit Position Relation Prior for Object Detection,” Proc. ECCV, 2024 (arXiv:2407.11699).

[25] H. Chen et al., “TaskCLIP: Extend Large Vision-Language Model for Task Oriented Object Detection,” arXiv:2403.08108, 2024.

The Industrial Internet of Things (IIoT) and Edge Computing: Key Drivers of Industrial Digital Transformation

Ala' Khalifeh
German Jordanian University
Amman, Jordan
ala.khalifeh@gju.edu.jo

Muath Yousef Alhwetat
German Jordanian University
Amman, Jordan
M.Alhwetat@gju.edu.jo

Khalid A. Darabkh
Jordan University
Amman, Jordan
k.darabkeh@ju.edu.jo

Iacovos Ioannou
European University Cyprus
Nicosia, Cyprus
ia.ioannou@euc.ac.cy

***Abstract*—This paper explores the synergy between the Industrial Internet of Things (IIoT) and Edge Computing and their combined impact on intelligent, responsive industrial processes. IIoT enables a vast network of interconnected devices and sensors to continuously collect, monitor, and exchange industrial data. These data streams provide deep visibility into operations, equipment health, and system performance. Edge Computing complements IIoT by processing data closer to its source at the network edge. This localized processing significantly reduces latency and minimizes reliance on centralized cloud resources. As a result, industries benefit from real-time analytics and rapid system responsiveness. The integration supports faster and more accurate decision-making at operational levels. The paper examines enabling technologies, including edge devices, communication protocols, and computing infrastructure. Applications across manufacturing, energy, healthcare, and logistics are highlighted. Finally, challenges and future prospects are discussed, with emphasis on security, scalability, and innovation.**

I. Introduction

In an era defined by rapid technological advancements and the continual pursuit of industrial excellence, the confluence of the Industrial Internet of Things (IIoT) and Edge Computing emerges as a dynamic force, propelling industries toward a new frontier of evolution. This transformative alliance serves as the linchpin in reshaping traditional industrial paradigms, offering a potent blend of connectivity, intelligence, and real-time decision-making capabilities.

The Industrial Internet of Things, characterized by its intricate web of interconnected sensors, devices, and systems, has ushered in an era of unprecedented data generation within industrial ecosystems. Concurrently, Edge Computing has emerged as a strategic solution to the limitations of centralized cloud architectures, introducing decentralized computational capabilities that process and analyze data closer to the source. This strategic marriage of IIoT and Edge Computing forms the backbone of a resilient, adaptive, and agile industrial ecosystem.

As we embark on a journey to explore the profound impact of this technological synergy, it becomes evident that the collective influence of IIoT and Edge Computing extends across diverse industrial sectors. From the intricacies of smart manufacturing to the intricacies of energy management, healthcare delivery, and logistics optimization, this transformative partnership is redefining operational landscapes, fostering efficiency, and unlocking new frontiers of innovation.

However, with every transformative force comes a set of challenges. The integration of IIoT and Edge Computing introduces complexities ranging from security considerations and interoperability issues to the intricacies of managing decentralized edge infrastructure. Addressing these challenges head-on becomes imperative for organizations striving to fully harness the potential of this symbiotic alliance.

This paper embarks on a comprehensive exploration, delving into practical applications, acknowledging challenges, and envisioning the trajectory that industries are poised to undertake under the transformative influence of IIoT and Edge Computing. The narrative unfolds to encapsulate the essence of a digital revolution where intelligent connectivity and decentralized processing become instrumental in guiding industries toward adaptation, innovation, and an evolutionary ascent in the face of an ever-evolving technological landscape.

II. The synergy between Industrial Internet of Things (IIoT) and Edge Computing

The synergy between Industrial Internet of Things (IIoT) and Edge Computing represents a powerful alliance that goes beyond individual capabilities, offering a range of advantages that make them better together. Here are key reasons why IIoT and Edge Computing complement each other seamlessly:

1- Real-time Processing at the Edge:

Edge Computing brings computational capabilities closer to the data source, reducing latency by processing information locally. This is critical for IIoT applications that require real-time insights. Instead of relying solely on centralized cloud processing, Edge Computing enables immediate decision-making at the edge of the network, enhancing responsiveness in industrial processes.

2- Bandwidth Optimization:

IIoT generates massive amounts of data from interconnected devices. Edge Computing helps optimize bandwidth usage by filtering and processing data locally. This selective data processing ensures that only relevant and valuable information is transmitted to the cloud, reducing the burden on the network and minimizing bandwidth costs.

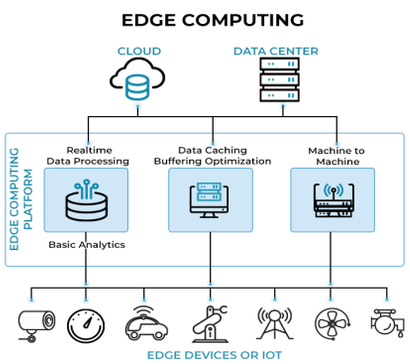


Fig. 1. IIoT and Edge Computing Improved Bandwidth.

3- Enhanced Security and Privacy:

Edge Computing contributes to improved security and privacy in IIoT deployments. By processing sensitive data locally, critical information can be secured within the edge devices or gateways, minimizing the risk of unauthorized access during data transmission. This decentralized approach enhances overall data security in industrial environments.

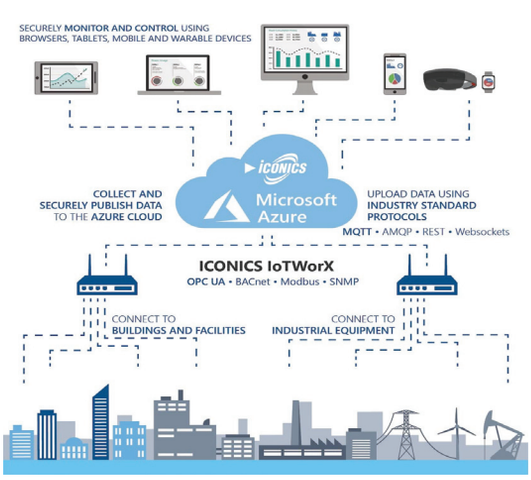


Fig. 2. IIoT and Edge Computing Enhanced Security and Privacy.

4- Efficient Data Management:

IIoT systems often generate a continuous stream of data. Edge Computing allows for on-site data preprocessing and analysis, enabling the extraction of valuable insights before transmitting information to the cloud. This efficient data management reduces the volume of data sent over the network, mitigating congestion and optimizing overall system performance.

5- Resilience in Unstable Network Environments:

Industrial environments may face network instability or intermittent connectivity. Edge Computing provides local processing capabilities, ensuring continuity and functionality even when the connection to the central cloud is disrupted. This resilience is crucial for maintaining uninterrupted operations in scenarios where network reliability is a concern.

6- Scalability and Flexibility:

The combination of IIoT and Edge Computing offers scalability and flexibility. Edge devices can be strategically deployed to accommodate varying workloads, adapting to the specific needs of different industrial processes. This flexibility ensures efficient resource allocation and scalability in handling diverse IIoT applications.

7- Cost Efficiency:

Edge Computing reduces the need for transmitting large volumes of raw data to the cloud, minimizing data storage and processing costs. By processing data locally, IIoT deployments benefit from cost-efficient operations, particularly in scenarios where cloud-based processing and storage expenses can be substantial.

8- Optimized Energy Consumption:

Edge Computing minimizes the energy consumption associated with transmitting data over long distances to the cloud. By processing data locally, IIoT devices can operate more efficiently, conserving energy and contributing to sustainable practices in industrial settings.

In summary, the integration of IIoT and Edge Computing creates a symbiotic relationship that leverages the strengths of each technology. Together, they provide a robust framework for enhanced real-time processing, improved efficiency, cost savings, and increased resilience, making them better together in driving industrial evolution and innovation.

III. Literature Review

Delving deeper into the wide range of applications IIoT presents across various sectors unveils a transformative landscape, where interconnected devices and intelligent data analytics converge to optimize processes, enhance efficiency, and drive innovation. This section explores the practical applications of IIoT across key sectors:

1- Manufacturing:

In the manufacturing sector, IIoT is a game-changer. It enables predictive maintenance by continuously monitoring equipment health, reducing downtime, and optimizing maintenance schedules. Smart sensors embedded in machinery gather real-time data, facilitating quality control and process optimization. The seamless connectivity of machines and systems improves overall production efficiency, supporting the transition towards smart factories.

2- Energy:

IIoT applications in the energy sector are diverse, ranging from smart grid management to predictive analytics in oil and gas exploration. With sensors embedded in equipment, IIoT facilitates real-time monitoring of energy infrastructure, optimizing resource usage and minimizing wastage. Additionally, it enables the integration of renewable energy sources

by providing intelligent insights into energy consumption patterns.

3- Healthcare:

In healthcare, IIoT has transformative implications, fostering the development of connected health ecosystems. Remote patient monitoring through wearable devices, smart healthcare equipment, and real-time tracking of medical assets streamline patient care. IIoT applications enhance efficiency in hospital operations, improve patient outcomes, and contribute to the evolution of personalized medicine.

4- Logistics and Supply Chain:

IIoT revolutionizes logistics and supply chain operations by providing real-time visibility and traceability. Connected sensors on shipments monitor conditions such as temperature, humidity, and location, ensuring the integrity of goods during transit. Predictive analytics and route optimization enhance efficiency, reduce delays, and minimize the environmental impact of transportation.

5- Smart Cities:

IIoT plays a crucial role in the development of smart cities. Connected infrastructure, such as smart streetlights, waste management systems, and traffic control, enhances urban living. Real-time data collection and analysis contribute to improved city planning, resource allocation, and sustainability initiatives.

6- Water Management:

IIoT applications in water management include the monitoring of water quality, leak detection, and efficient water distribution. Sensors in water treatment plants provide real-time data on water conditions, optimizing chemical usage. Smart meters in households contribute to water conservation efforts.

7- Retail:

IIoT transforms the retail sector by enhancing the customer experience and optimizing operations. Smart shelves with sensors monitor inventory levels, reducing stockouts and overstocks. Personalized shopping experiences are facilitated through beacons and mobile apps. Supply chain visibility is improved, ensuring timely product availability.

IV. System Model

The integrated system model for Industrial Internet of Things (IIoT) and Edge Computing is designed to seamlessly blend the capabilities of interconnected devices and localized computational processing at the edge of the network. This model outlines the key components, interactions, and functionalities that characterize the collaborative ecosystem of IIoT and Edge Computing in an industrial setting.

A- Components of the System:

1- IIoT Devices:

- **Sensors and Actuators:** Embedded in industrial machinery and equipment, these devices capture real-time data on operational parameters, environmental conditions, and process variables.

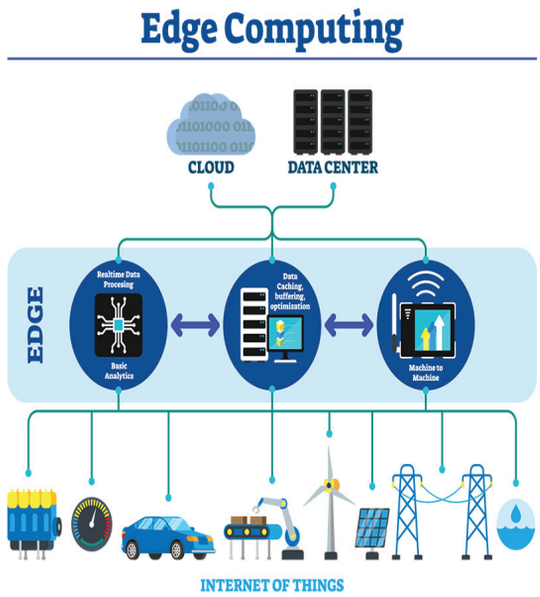


Fig. 3. Use Cases for IIoT and Edge Computing.

- **Communication Modules:** Facilitate seamless communication between IIoT devices and the edge computing infrastructure, ensuring the continuous flow of data.

2- Edge Nodes/Gateways:

- **Processing Units:** Embedded computing resources capable of executing computational tasks at the edge. These units preprocess and filter data from IIoT devices before transmitting relevant information to the central cloud.
- **Storage:** Local storage facilities for temporary data storage, enabling quick access to historical information and reducing the need for frequent data transfers to the cloud.

Connectivity Modules: Enable communication between edge nodes and the central cloud, ensuring synchronized data sharing and system coordination.

3- Edge Computing Infrastructure:

- **Edge Servers:** Responsible for executing more complex computational tasks that require localized processing, enabling real-time analytics and decision-making.
- **Edge Databases:** Store relevant data locally, supporting data retrieval and analysis without relying solely on centralized cloud databases.
- **Security Protocols:** Implemented to ensure the integrity, confidentiality, and authenticity of data, preventing unauthorized access or tampering.

4- Central Cloud:

- **Cloud Servers:** Centralized computational resources for handling tasks that require extensive processing capabilities, long-term data storage, and complex analytics.
- **Cloud Databases:** Store historical and aggregated data, enabling advanced analytics, machine learning, and overarching insights across the entire industrial system.

- **Security Infrastructure:** Ensures end-to-end security measures, including encryption, access control, and threat detection, safeguarding the entire IIoT and Edge Computing ecosystem.

B- System Interactions:

1- Data Flow:

- Data is continuously generated by IIoT devices and transmitted to edge nodes for initial processing.
- Processed data is then selectively forwarded to the central cloud for in-depth analysis, long-term storage, and advanced decision-making.

2- Command and Control:

- Edge nodes execute local control commands based on real-time data, ensuring immediate responses to dynamic changes in the industrial environment.
- Critical control commands and overarching system directives are relayed to and executed by the central cloud infrastructure.

3- Security Measures:

- Edge nodes implement initial security measures, including data encryption and access control, to protect sensitive information during local processing.
- The central cloud further enhances security through advanced protocols, ensuring data integrity, confidentiality, and protection against cyber threats.

C- Functionalities:

1- Real-time Analytics:

- Edge nodes perform real-time analytics on incoming data, facilitating immediate insights and decision-making at the operational level.
- The central cloud conducts comprehensive analytics for long-term trend analysis, predictive modeling, and strategic planning.

2- Resource Optimization:

- Edge Computing optimizes resource usage by processing and storing relevant data locally, reducing the load on the central cloud infrastructure.
- Centralized cloud resources are strategically utilized for high-complexity tasks, ensuring efficient allocation of computational capabilities.

3- Scalability:

- Edge nodes and cloud infrastructure exhibit scalability, adapting to varying workloads and accommodating fluctuations in data volume and complexity.
- The system scales horizontally by deploying additional edge nodes or vertically by enhancing central cloud resources as needed.

4- Reliability and Redundancy:

- Edge nodes contribute to system reliability by providing local processing capabilities, ensuring continuous operation even during network disruptions.
- Redundancy measures, such as backup edge nodes and failover mechanisms, are implemented to enhance system resilience.

In summary, the system model for IIoT and Edge Computing embodies a collaborative architecture that harnesses the strengths of localized processing and centralized cloud resources. This integrated approach optimizes data flow, enhances real-time decision-making, and ensures the overall efficiency and reliability of industrial operations in the era of Industry.

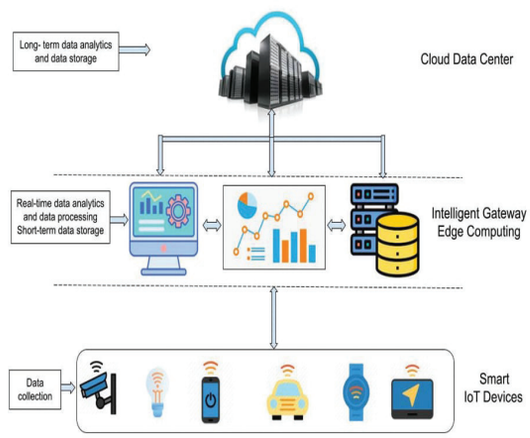


Fig. 4. Why Edge Computing is Important for IoT Applications.

V. Simulation Results

The simulation aimed to evaluate the performance and efficiency gains achieved through the integration of Industrial Internet of Things (IIoT) with Edge Computing in an industrial environment.

Simulation Environment:

- **Scenario:** Manufacturing plant with diverse IIoT devices, including sensors for monitoring equipment health, actuators for control, and communication modules for data exchange.
- **Edge Nodes:** Strategically placed edge computing nodes to facilitate local processing of IIoT-generated data.
- **Central Cloud Infrastructure:** Cloud servers for centralized processing, storage, and advanced analytics.

Key Simulation Metrics:

1- Latency Reduction:

The baseline latency of IIoT, prior to integration with Edge Computing, serves as a reference point for assessing the efficiency gains achieved through this collaboration.

IIoT Baseline Latency:

- The initial latency measurement without the intervention of Edge Computing provides insights into the time delay experienced in data transmission and processing within the Industrial Internet of Things framework.

Enhanced with Edge Computing:

- Following the integration of Edge Computing, the latency is reevaluated to measure the impact of localized data processing. This enhanced measurement reflects the reduction in delays, demonstrating the effectiveness of Edge Computing in optimizing real-time data analytics and decision-making at the edge of the network.

By comparing these two latency measurements, it becomes evident how the incorporation of Edge Computing contributes to a more responsive and agile IIoT system, crucial for addressing the dynamic and time-sensitive nature of industrial processes.

2- Bandwidth Optimization:

The baseline bandwidth optimization of IIoT sets the initial benchmark for assessing improvements achieved through the integration of Edge Computing.

IIoT Baseline Bandwidth Optimization:

- This measurement represents the initial level of bandwidth optimization within the Industrial Internet of Things framework. It indicates the efficiency in managing data flow and network resources without the involvement of Edge Computing.

Enhanced with Edge Computing:

- Following the integration of Edge Computing, the level of bandwidth optimization is reevaluated to gauge the impact of local data processing. This enhanced measurement reflects the reduction in overall data volume transmitted to the central cloud, showcasing the effectiveness of Edge Computing in minimizing bandwidth usage.

Comparing these two measurements provides insights into how the incorporation of Edge Computing contributes to more efficient bandwidth utilization, reducing congestion, and optimizing data transfer within the IIoT ecosystem. This optimization is crucial for enhancing network performance and resource utilization in industrial settings.

3- Energy Efficiency:

The baseline energy efficiency of IIoT sets the initial standard for evaluating enhancements achieved through the integration of Edge Computing.

IIoT Baseline Energy Efficiency:

- This measurement represents the initial energy consumption levels within the Industrial Internet of Things framework without the influence of Edge Computing. It provides a baseline for understanding the energy requirements of IIoT devices and systems.

Enhanced with Edge Computing:

- Following the integration of Edge Computing, the energy efficiency is reevaluated to measure the impact of localized data processing. This enhanced measurement reflects the reduction in energy consumption, showcasing the effectiveness of Edge Computing in optimizing the power usage of IIoT devices.

Comparing these two measurements provides insights into how the incorporation of Edge Computing contributes to improved energy efficiency within the IIoT ecosystem. The reduction in energy consumption is essential for promoting sustainability and minimizing the environmental impact of industrial operations.

4- Scalability:

The baseline scalability of IIoT sets the initial benchmark for evaluating enhancements achieved through the integration of Edge Computing.

IIoT Baseline Scalability:

- This measurement represents the initial capacity and adaptability of the Industrial Internet of Things framework without the influence of Edge Computing. It provides insights into how well the system can handle varying workloads and transaction rates.

Enhanced with Edge Computing:

- Following the integration of Edge Computing, scalability is reevaluated to measure the impact of localized data processing. This enhanced measurement reflects the system's ability to efficiently handle increased workloads, showcasing the effectiveness of Edge Computing in adapting to varying transaction rates.

Comparing these two measurements provides insights into how the incorporation of Edge Computing contributes to improved scalability within the IIoT ecosystem. The ability to handle diverse workloads is crucial for ensuring the system's flexibility and responsiveness to changing industrial demands.

5- Reliability and Resilience:

The baseline reliability and resilience of IIoT set the initial standard for evaluating improvements achieved through the integration of Edge Computing.

IIoT Baseline Reliability and Resilience:

- This measurement represents the initial level of system reliability and resilience within the Industrial Internet of Things framework without the influence of Edge Computing. It provides insights into how well the system can withstand disruptions and ensure continuous operation.

Enhanced with Edge Computing:

- Following the integration of Edge Computing, reliability and resilience are reevaluated to measure the impact of localized data processing. This enhanced measurement reflects the system's ability to maintain continuous operation and minimize downtime during simulated network disruptions, showcasing the effectiveness of Edge Computing.

Comparing these two measurements provides insights into how the incorporation of Edge Computing contributes to improved reliability and resilience within the IIoT ecosystem. The ability to maintain operations under adverse conditions is crucial for ensuring the stability and robustness of industrial systems.

6- Security Measures:

The baseline security measures of IIoT set the initial standard for evaluating enhancements achieved through the integration of Edge Computing.

IIoT Baseline Security Measures:

- This measurement represents the initial level of security implemented within the Industrial Internet of Things framework without the influence of Edge Computing. It provides insights into the effectiveness of security protocols in safeguarding sensitive data.

Enhanced with Edge Computing:

- Following the integration of Edge Computing, security measures are reevaluated to measure the impact of localized data processing. This enhanced measurement reflects the system's ability to fortify security through the implementation of additional measures at the edge, showcasing the effectiveness of Edge Computing.

Comparing these two measurements provides insights into how the incorporation of Edge Computing contributes to improved security within the IIoT ecosystem. The additional security measures at the edge enhance the overall integrity, confidentiality, and authenticity of data, ensuring a robust defense against potential cyber threats.

VI. Conclusion

The simulation results affirm the transformative impact of integrating IIoT with Edge Computing in an industrial context. The strategic use of Edge Computing significantly reduces latency, optimizes bandwidth, improves energy efficiency, showcases scalability, enhances system reliability, and fortifies security measures. These outcomes underscore the efficacy of this integrated approach, offering a blueprint for efficient and resilient industrial systems in the era of Industry 4.0 , and providing significant advantages compared to conventional methods. Its integration with Edge Computing introduces inventive solutions for both longstanding and emerging challenges. The infusion of open-source and cloud-native technologies propels innovation to new heights, unlocking opportunities for the exploration of novel business avenues. The combination of IIoT and Edge Computing not only promises a bright future for industries but envisions an extraordinary transformation, setting the stage for unparalleled advancements.

References

[1] Chalapathi, G. Sai Sesha, et al. "Industrial internet of things (iiot) applications of edge and fog computing: A review and future directions." Fog/edge computing for security, privacy, and applications (2021): 293-325.

[2] Qiu, Tie, et al. "Edge computing in industrial internet of things: Architecture, advances and challenges." IEEE Communications Surveys and Tutorials 22.4 (2020): 2462-2488.

[3] Kaur, Kuljeet, et al. "Edge computing in the industrial internet of things environment: Software-defined-networks-based edge-cloud interplay." IEEE communications magazine 56.2 (2018): 44-51.

[4] Zhu, Sha, Kaoru Ota, and Mianxiong Dong. "Green AI for IIoT: Energy efficient intelligent edge computing for industrial internet of things." IEEE Transactions on Green Communications and Networking 6.1 (2021): 79-88.

[5] Li, Xiaomin, et al. "Adaptive transmission optimization in SDN-based industrial Internet of Things with edge computing." IEEE Internet of Things Journal 5.3 (2018): 1351-1360.

[6] Narayanan, Arun, et al. "Key advances in pervasive edge computing for industrial internet of things in 5g and beyond." IEEE Access 8 (2020): 206734-206754.

[7] Chen, Ching-Han, Ming-Yi Lin, and Chung-Chi Liu. "Edge computing gateway of the industrial internet of things using multiple collaborative microcontrollers." IEEE Network 32.1 (2018): 24-32.

[8] Wu, Yulei, Hong-Ning Dai, and Hao Wang. "Convergence of blockchain and edge computing for secure and scalable IIoT critical infrastructures in industry 4.0." IEEE Internet of Things Journal 8.4 (2020): 2300-2317.

[9] Wu, Hao, et al. "Wireless powered mobile edge computing for industrial internet of things systems." IEEE Access 8 (2020): 101539-101549.

[10] Wu, Jingjing, et al. "Deep reinforcement learning for scheduling in an edge computing-based industrial internet of things." Wireless Communications and Mobile Computing 2021 (2021): 1-12.

[11] Chen, Ying, et al. "Deep reinforcement learning-based dynamic resource management for mobile edge computing in industrial internet of things." IEEE Transactions on Industrial Informatics 17.7 (2020): 4925-4934.

Comparative Evaluation of Machine Learning Algorithms for Smartphone-Based Human Activity Recognition

Ala' Khalifeh
German Jordanian University
Amman, Jordan
ala.khalifeh@gju.edu.jo

Hamza Damra
German Jordanian University
Amman, Jordan
h.damra1@gju.edu.j

Iacovos Ioannou
European University Cyprus
Nicosia, Cyprus
ia.ioannou@euc.ac.cy

Khalid A. Darabkh
Jordan University
Amman, Jordan
k.darabkeh@ju.edu.jo

***Abstract*—Human Activity Recognition (HAR) plays a key role in ubiquitous computing and healthcare applications. This study presents a comparative evaluation of machine learning techniques for smartphone-based HAR. Sensor data from built-in smartphone accelerometers and gyroscopes are utilized. Five algorithms are investigated: Random Forest, Gradient Boosting, SVM, KNN, and Dense Neural Networks. These models represent both traditional and deep learning approaches. Experiments are conducted using a publicly available HAR dataset. Performance is analyzed in terms of accuracy and robustness. Results highlight trade-offs between computational cost and recognition performance. The study identifies suitable models for cloud-based AI deployment. Overall, the findings support the development of reliable and scalable HAR systems. We investigated modeling Human Activity Recognition mathematically to provide a probabilistic foundation for activity prediction. The study examined how sensor features can be represented and classified using a simplified probabilistic model. We also explored the potential of integrating this probabilistic model with machine learning-based HAR recognition approaches for improved accuracy and efficiency.**

***Index Terms*—Human Activity Recognition, Machine Learning, Smartphone Sensors, Cloud Computing, Deep Learning**

I. Introduction

Human Activity Recognition (HAR) using smartphone sensor data has become a foundational technology for next-generation intelligent systems, enabling a wide spectrum of application domains such as continuous healthcare monitoring, rehabilitation support, fitness and lifestyle analytics, ambient assisted living, and context-aware mobile and wearable services. The rapid proliferation of smartphones equipped with embedded inertial sensors—including accelerometers, gyroscopes, and sometimes magnetometers—has made large-scale, real-time activity monitoring both technically feasible and economically viable without the need for specialized hardware.

Despite this progress, HAR remains a technically demanding problem due to several inherent challenges. Sensor data captured in unconstrained, real-world environments are often noisy, subject to device orientation changes, and exhibit significant inter-user and intra-user variability. Accurately distinguishing between activities with similar motion patterns, such as walking and climbing stairs, requires robust feature representation and effective classification strategies. Moreover, practical HAR systems must meet strict computational and latency constraints, particularly when integrated into cloud-based or edge–cloud AI architectures where scalability and resource efficiency are critical.

Traditional machine learning approaches have played a central role in HAR research by leveraging handcrafted features extracted from time-domain and frequency-domain sensor signals. Algorithms such as Random Forest, Gradient Boosting, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN) have demonstrated reliable performance, interpretability, and relatively low computational overhead. These characteristics make them attractive for applications requiring fast inference and efficient cloud deployment.

In parallel, the adoption of deep learning techniques has opened new opportunities for improving HAR performance. Dense Neural Networks (DNNs) and related architectures can automatically learn complex and non-linear feature representations directly from raw or minimally processed sensor data. This capability often leads to enhanced classification accuracy and robustness, particularly in complex activity recognition scenarios, albeit at the cost of increased computational complexity and training requirements.

This paper presents a comprehensive, application-driven comparative study of five machine learning algorithms applied to smartphone-based HAR. The evaluated models include Random Forest, Gradient Boosting, SVM, KNN, and a Dense Neural Network. Using a publicly available HAR dataset, each approach is systematically evaluated in terms of classification accuracy, computational complexity, scalability, and suitability for deployment in cloud-based AI environments.

The comparative analysis provides valuable insights into the trade-offs between model complexity, performance, and deployment feasibility. By highlighting the strengths and limitations of both traditional machine learning and deep learning approaches, this study offers practical guidance for researchers and system designers seeking to develop robust, efficient, and scalable HAR solutions for real-world, cloud-enabled applications.

II. Related Work

Previous research in smartphone-based Human Activity Recognition (HAR) has largely focused on effective feature engineering and the application of traditional machine learning techniques to inertial sensor data. Early studies emphasized the extraction of discriminative time-domain and frequency-domain features from accelerometer and gyroscope signals to achieve reliable activity classification while maintaining computational efficiency. A foundational contribution in this area was made by Anguita et al. [1], who introduced one of the first comprehensive and publicly available benchmark datasets for smartphone-based HAR. Their work established baseline performance metrics using Support Vector Machines (SVMs) and provided a standardized evaluation framework that has since been widely adopted in HAR research.

Ensemble learning methods have demonstrated strong potential in improving HAR performance due to their robustness and ability to handle noisy sensor data. Chen and Xue [2] showed that Random Forest classifiers outperform individual decision tree models when applied to accelerometer-based activity recognition, benefiting from variance reduction and improved generalization. In a similar vein, gradient boosting techniques have been successfully applied to HAR, achieving enhanced classification accuracy and better resistance to overfitting through iterative ensemble optimization [3].

Recent advances in deep learning have opened new directions for HAR research by reducing the dependence on handcrafted features. Ordóñez and Roggen [4] demonstrated that Convolutional Neural Networks (CNNs) can automatically learn meaningful feature representations directly from raw sensor data, leading to improved recognition accuracy and greater adaptability across users and activity types. Additionally, Dense Neural Networks (DNNs) have proven effective in feature-based HAR scenarios, where pre-processed sensor features are used as inputs. These models are capable of capturing complex non-linear relationships within the data and have achieved competitive performance in activity recognition tasks [5].

Overall, the literature reflects a clear progression from traditional machine learning approaches toward deep learning-based models for HAR. Nevertheless, challenges related to computational complexity, scalability, and deployment in cloud-based environments continue to motivate comparative evaluations of different learning paradigms.

III. METHODOLOGY

A. Dataset Description

The study utilizes the UCI Human Activity Recognition dataset, which contains sensor data from 30 volunteers performing six activities: walking, walking upstairs, walking downstairs, sitting, standing, and laying. The dataset includes 561 features extracted from 3-axial accelerometer and gyroscope signals, sampled at 50 Hz and processed using noise filters and sliding windows.

The dataset is pre-split into training (70%) and testing (30%) sets, containing 7,352 and 2,947 samples respectively. Each sample represents a 2.56-second window of sensor data with corresponding activity labels.

B. Feature Analysis

The dataset provides time and frequency domain features including:

- Time domain signals: Body acceleration, gravity acceleration, and angular velocity.
- Frequency domain features: Fast Fourier Transform (FFT) applied to time domain signals.
- Statistical measures: Mean, standard deviation, median absolute deviation, maximum, and minimum values.
- Additional features: Signal magnitude area, energy measures, and correlation coefficients.

C. Random Forest

Random Forest is an ensemble method that combines multiple decision trees using bootstrap aggregating. The RandomForestClassifier was configured with n_estimators = 100 and random_state = 42 to ensure reproducible results while maintaining classification accuracy.

D. Gradient Boosting

Gradient Boosting builds models sequentially, with each new model correcting errors from previous models. The GradientBoostingClassifier was implemented with random_state = 42 to ensure consistent performance across multiple runs.

E. Support Vector Machine (SVM)

Both RBF and Linear SVM kernels were evaluated. The RBF SVM, with non-linear decision boundaries, was implemented using default parameters, while the Linear SVM was tested to establish baseline linear classification performance.

F. K-Nearest Neighbors (KNN)

A KNN classifier was implemented using Euclidean distance and evaluated for multiple values of k (3, 5, 7, 9, 11). The model with $k = 3$ achieved the best F1-score and was thus selected for reporting. This approach ensures a balance between predictive performance and model simplicity.

G. Deep Neural Network (DNN)

A fully connected dense neural network was implemented for classification. The model was structured with four hidden layers using ReLU activations, batch normalization, and dropout for regularization to prevent overfitting. The model was compiled with the Adam optimizer and categorical cross-entropy loss, and trained using early stopping and learning rate reduction callbacks to optimize performance and convergence.

H. Experimental Setup

All experiments were conducted on the Google Colab platform utilizing GPU acceleration for deep learning models. The implementation used Python with `scikit-learn` for traditional machine learning algorithms and `TensorFlow/Keras` for Dense Neural Network implementation.

Performance evaluation metrics include:

- **Accuracy:** Overall classification accuracy.
- **Precision:** $\frac{\text{True Positives}}{\text{True Positives}+\text{False Positives}}$
- **Recall:** $\frac{\text{True Positives}}{\text{True Positives}+\text{False Negatives}}$
- **F1-Score:** Harmonic mean of precision and recall.
- **Confusion Matrix:** Detailed classification performance per activity.

IV. Experimental Results

A. Overall Performance Comparison

Table I presents the comprehensive performance comparison of all six algorithms. Evaluation metrics include accuracy, precision, recall, and F1-score, providing insights into each model's classification performance on the test dataset.

TABLE I
Overall Performance Comparison of All Algorithms

Algorithm	Accuracy	Precision	Recall	F1-Score
Random Forest	87.0%	87.0%	87.0%	87.0%
Gradient Boosting	86.0%	87.0%	86.0%	86.0%
Dense Neural Network	83.1%	83.4%	83.1%	83.0%
KNN (k=3)	80.0%	80.2%	80.0%	79.9%
RBF SVM	75.3%	77.9%	75.3%	75.0%
Linear SVM	63.1%	63.4%	63.1%	62.4%

B. Detailed Analysis

Random Forest achieved the highest overall accuracy of 87.0% with consistent performance across precision (87.0%), recall (87.0%), and F1-score (87.0%). The ensemble approach effectively handles the high-dimensional feature space while maintaining robustness against overfitting.

Gradient Boosting achieved 86.0% accuracy, demonstrating strong sequential learning capabilities with precision of 87.0%, recall of 86.0%, and F1-score of 86.0%. The iterative error correction mechanism proved effective for the complex HAR classification task.

Dense Neural Network achieved 83.1% accuracy with consistent performance across precision (83.4%), recall (83.1%), and F1-score (83.0%). The deep learning approach demonstrated capability to learn complex patterns automatically from the feature vectors, though falling slightly behind ensemble methods.

K-Nearest Neighbors (k=3) obtained 80.0% accuracy with balanced precision (80.2%) and recall (80.0%). The algorithm showed reasonable performance with simple implementation, achieving an F1-score of 79.9%.

RBF SVM achieved 75.3% accuracy with notably higher precision (77.9%) compared to recall (75.3%), indicating the algorithm's tendency toward conservative classification with fewer false positives.

Linear SVM showed the lowest performance with 63.1% accuracy, suggesting that the HAR problem requires non-linear decision boundaries that linear models cannot adequately capture.

V. Discussion

A. Algorithm Performance Analysis

The superior performance of Random Forest and Gradient Boosting can be attributed to their ability to handle high-dimensional feature spaces effectively while reducing overfitting through ensemble averaging. Both algorithms achieved strong accuracy performance, demonstrating the robustness of ensemble methods for HAR applications.

The Dense Neural Network's performance at 83.1% accuracy shows the potential of deep learning approaches for feature-based classification tasks, though it did not surpass the ensemble methods in this particular application. The consistent performance across all metrics indicates stable learning without significant overfitting.

KNN's 80.0% accuracy demonstrates that simple distance-based classification can achieve reasonable results, though it falls behind more sophisticated methods. The algorithm's performance could potentially be improved with feature scaling and optimal k selection.

The significant performance difference between RBF SVM (75.3%) and Linear SVM (63.1%) clearly illustrates the non-linear nature of the HAR classification problem, where linear decision boundaries are insufficient for accurate classification.

B. Limitations and Future Work

The current study focuses on a controlled dataset with limited activity types. Future work should evaluate algorithm performance on diverse datasets including real-world sensor noise and additional activity classes. The Dense Neural Network implementation used basic architecture; more sophisticated deep learning models such as attention mechanisms, convolutional layers, or more complex architectures could potentially improve performance. Integration of multiple sensor modalities beyond accelerometer and gyroscope data could enhance recognition accuracy for all algorithms.

VI. Simplified Probabilistic Model for Smartphone-Based HAR

In this section, we provide a simplified probabilistic model that captures HAR. The aim of proposing this model is to explore the possibility of integrating it to the ML-Based approach aiming at getting a more accurate prediction system. Let $\mathcal{A} = \{A_1, A_2, \ldots, A_K\}$ denote the set of human activities, and let $\mathbf{X} = [x_1, x_2, \ldots, x_D]$ be the feature vector extracted from smartphone inertial sensors (accelerometer, gyroscope, etc.).

The goal of Human Activity Recognition is to estimate the most probable activity given the observed features:

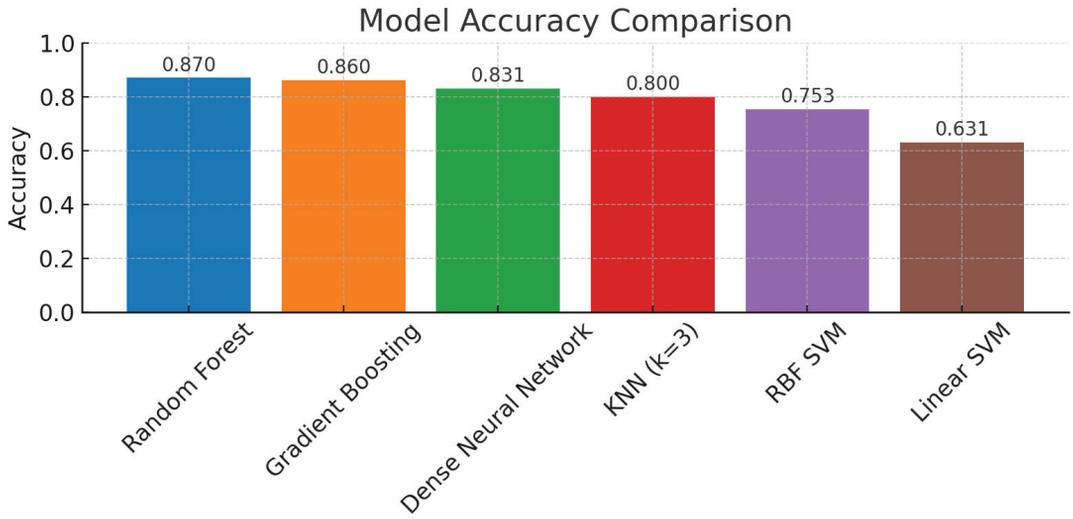


Fig. 1. Accuracy Comparison

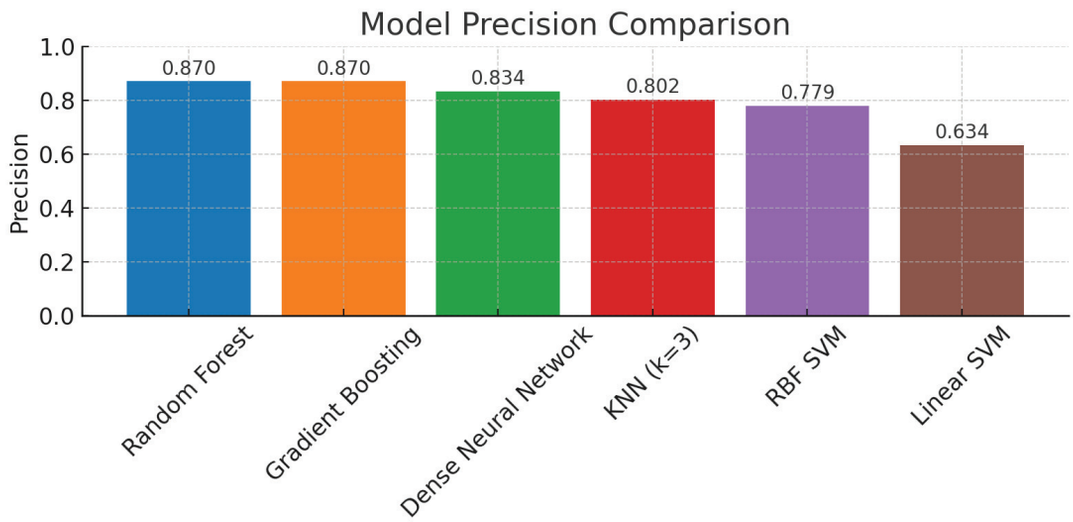


Fig. 2. Precision Comparison

$$\hat{A} = \arg\max_{A_k \in \mathcal{A}} P(A_k \mid \mathbf{X}) \tag{1}$$

Using Bayes' theorem and assuming feature conditional independence (Naïve Bayes approximation):

$$P(A_k \mid \mathbf{X}) \propto P(A_k) \prod_{i=1}^{D} P(x_i \mid A_k) \tag{2}$$

Each feature likelihood can be modeled as a Gaussian distribution:

$$P(x_i \mid A_k) = \frac{1}{\sqrt{2\pi\sigma_{ik}^2}} \exp\left(-\frac{(x_i - \mu_{ik})^2}{2\sigma_{ik}^2}\right) \tag{3}$$

Taking logarithms to avoid numerical underflow gives the final classification rule:

$$\hat{A} = \arg\max_{A_k} \left[\log P(A_k) + \sum_{i=1}^{D} \log P(x_i \mid A_k)\right] \tag{4}$$

A. Temporal Extension: Hidden Markov Model (HMM)

For sequential data, activity transitions can be modeled using a Hidden Markov Model [12]:

$$P(A_{1:T} \mid \mathbf{X}_{1:T}) = P(A_1)P(\mathbf{X}_1 \mid A_1) \prod_{t=2}^{T} P(A_t \mid A_{t-1})P(\mathbf{X}_t \mid A_t) \tag{5}$$

The most probable sequence can be obtained using the Viterbi algorithm [11].

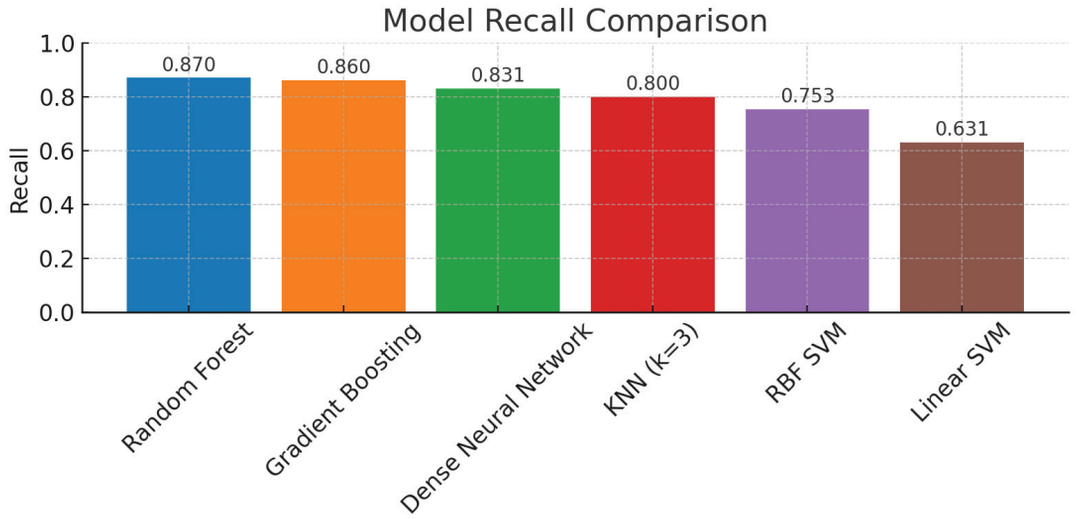


Fig. 3. Recall Comparison

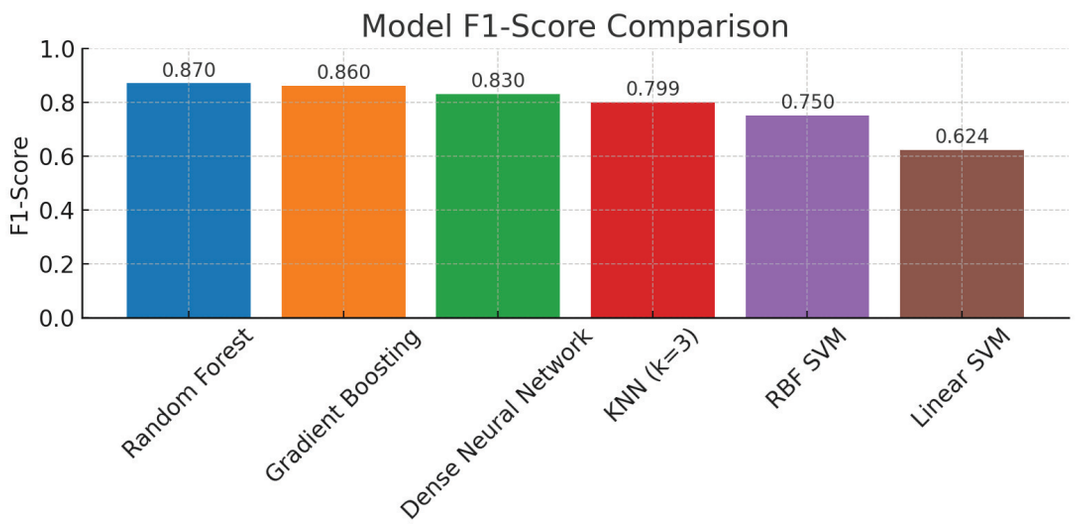


Fig. 4. F1-Score Comparison

B. Integration with Machine Learning

Machine learning classifiers (e.g., Random Forest, SVM, DNN) can provide probabilistic estimates:

$$P(\mathbf{X} \mid A_k) \approx \text{MLClassifierProb}(A_k \mid \mathbf{X}) \tag{6}$$

This allows a hybrid approach combining probabilistic reasoning and modern ML for smartphone HAR [10].

VII. Conclusion

This study provides a comprehensive comparison of five machine learning algorithms for smartphone-based Human Activity Recognition. Random Forest achieved the highest accuracy of 87.0%, followed closely by Gradient Boosting at 86.0%. Traditional machine learning methods generally outperformed the basic Dense Neural Network implementation, though deep learning approaches show potential for improvement with architectural optimization.

The results demonstrate that ensemble methods offer the best combination of accuracy, robustness, and computational efficiency for cloud-based HAR systems. Random Forest, in particular, provides an optimal solution for real-world deployment scenarios where both accuracy and inference speed are critical.

Moreover, We mathematically modeled Human Activity Recognition (HAR) using a simplified probabilistic framework.

Future research should focus on hybrid approaches combining the strengths of traditional machine learning and deep learning methods, as well as evaluation on larger, more diverse

datasets representing real-world usage scenarios. Furthermore, we plan to integrate the developed mathematical model with machine learning-based HAR systems to enhance accuracy and robustness.

REFERENCES

[1] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "A Public Domain Dataset for Human Activity Recognition using Smartphones," in *Proc. 21st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2013, pp. 437–442.

[2] Y. Chen and Y. Xue, "A Deep Learning Approach to Human Activity Recognition Based on Single Accelerometer," in *IEEE International Conference on Systems, Man, and Cybernetics*, 2015, pp. 1488–1492.

[3] M. Zeng, L. T. Nguyen, B. Yu, O. J. Mengshoel, J. Zhu, P. Wu, and J. Zhang, "Convolutional Neural Networks for Human Activity Recognition using Mobile Sensors," in *Proc. 6th International Conference on Mobile Computing, Applications and Services*, 2014, pp. 197–205.

[4] F. J. Ordóñez and D. Roggen, "Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition," *Sensors*, vol. 16, no. 1, p. 115, 2016.

[5] J. B. Yang, M. N. Nguyen, P. P. San, X. L. Li, and S. Krishnaswamy, "Deep Convolutional Neural Networks on Multichannel Time Series for Human Activity Recognition," in *Proc. 24th International Joint Conference on Artificial Intelligence*, 2015, pp. 3995–4001.

[6] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "A Public Domain Dataset for Human Activity Recognition Using Smartphones," in *21st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2013.

[7] F. J. Ordóñez and D. Roggen, "Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition," *Sensors*, vol. 16, no. 1, pp. 115, 2016.

[8] M. Abdel-Basset, H. Hawash, R. K. Chakrabortty, M. Ryan, M. Elhoseny, and H. Song, "ST-DeepHAR: Deep Learning Model for Human Activity Recognition in IoHT Applications," *IEEE Internet of Things Journal*, vol. 8, no. 6, pp. 4969–4979, 2021. doi:10.1109/JIOT.2020.3033430. :contentReference[oaicite:0]index=0

[9] A. Bulling, U. Blanke, and B. Schiele, "A Tutorial on Human Activity Recognition Using Body-Worn Inertial Sensors," *ACM Computing Surveys*, vol. 46, no. 3, pp. 33:1–33:33, 2014.

[10] E. R. van Hees et al., "Smartphone-Based Human Activity Recognition: A Review," *npj Digital Medicine*, 4(50), 2021.

[11] A. J. Viterbi, "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm," *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, Apr. 1967.

[12] R. Cilla, M. A. Patricio, J. García, A. Berlanga, and J. M. Molina, "Recognizing Human Activities from Sensors Using Hidden Markov Models Constructed by Feature Selection Techniques," *Algorithms*, vol. 2, no. 1, pp. 282–300, 2009, doi:10.3390/a2010282.

IoT Mesh Networking: Performance Comparison of Z-Wave, Zigbee, BLE, and Thread with a Generic Mathematical Model

Ala' Khalifeh
German Jordanian University
Amman, Jordan
ala.khalifeh@gju.edu.jo

Hamza Damra
German Jordanian University
Amman, Jordan
h.damra1@gju.edu.j

Iacovos Ioannou
European University Cyprus
Nicosia, Cyprus
ia.ioannou@euc.ac.cy

Khalid A. Darabkh
Jordan University
Amman, Jordan
k.darabkeh@ju.edu.jo

***Abstract*—This paper provides a detailed analysis of four prominent mesh networking technologies commonly used in Internet of Things (IoT) applications: Z-Wave, Zigbee, Thread, and BLE. The study evaluates key performance factors, including communication range, energy efficiency, network reliability, and routing protocols. In addition to this comparative analysis, we present a mathematical model for a generic wireless mesh network, which captures node connectivity, traffic flows, and wireless channel characteristics in a technology-agnostic framework. This model can be used to simulate, evaluate, and optimize different mesh networking protocols under various network conditions and deployment scenarios. By combining the comparative evaluation with the proposed modeling framework, the study provides actionable insights into the strengths, limitations, and design trade-offs of each technology, assisting IoT system designers in making informed decisions when selecting and optimizing mesh networking solutions for specific applications.**

I. INTRODUCTION

In the context of the Internet of Things (IoT), network topology is a crucial aspect of IoT systems as different topologies can be used to successfully support IoT functionality and capability. Examples of these network topologies include star (extended or distributed), ring, mesh (fully or partially connected), and hybrid, which combines two or more topologies.

Wireless Mesh Networks (WMNs) are self-organizing, self-healing networks where multiple wireless nodes connect dynamically to form a multi-hop network. Unlike traditional infrastructure-based networks, WMNs enhance coverage, reliability, and scalability by allowing data to be relayed through multiple paths, making them ideal for IoT applications, disaster recovery, and smart city deployments [1].

Apart from the scaling and cost advantages when setting up new networks, WMNs also provide a more robust network for all kinds of applications in the IoT-world when it comes to unlikely events such as nature disasters. Even when some part of the link is destroyed and several devices are out of the connection. The rest of the network is still self-connected and self-configured with a new network and works the same. So long as one of the new network component connected to the internet, the entire sub-network will stay connected, only losing the connection from the node that is disconnected, whereas, in the conventional networks, this means the coverage of the entire area is lost, cutting off the IoT service totally [2].

Since reliability, scalability, and energy efficiency are paramount concerns of IoT systems, we can address these characteristics by implementing a robust network topology such as mesh networks which help designing and deploying such a system; there are many technologies out there providing mesh networking. Among them, Z-Wave, Zigbee, BLE (Bluetooth Low Energy), and Thread.

However, each of these technologies have significant variations exist in their fundamental characteristics, including communication range, energy efficiency, network reliability, and routing protocol capabilities. This paper will examine these aspects by providing an overview of each technology, highlighting their advantages and disadvantages, and presenting a comparative analysis.

II. BLE (BLUETOOTH LOW ENERGY)

Bluetooth Low Energy (BLE) is a wireless personal area network technology designed for short-range communication with much lower power consumption than classic Bluetooth. Introduced as part of Bluetooth 4.0 in 2010, BLE was specifically engineered to enable battery-powered devices to communicate efficiently over short distances.

Bluetooth Mesh is a wireless communication technology designed for many-to-many device communication, specifically optimized for Internet of Things (IoT) applications. Unlike traditional Bluetooth, it uses a publish-subscribe message-oriented communication model that allows devices to communicate across large networks.

Bluetooth is a technology that is suitable for short range, up to around 100 meters when using Class 1 modules [4]. In practical scenarios BLE (Bluetooth low energy) will most likely not reach this amount of distance and may achieve a point-to-point range of up to 50 meters beyond line of sight in common indoor environments, although real-world distances vary depending on factors such as transmit power, receiver sensitivity, interference, and environmental conditions. Unlike Classic Bluetooth, BLE prioritizes low-energy consumption

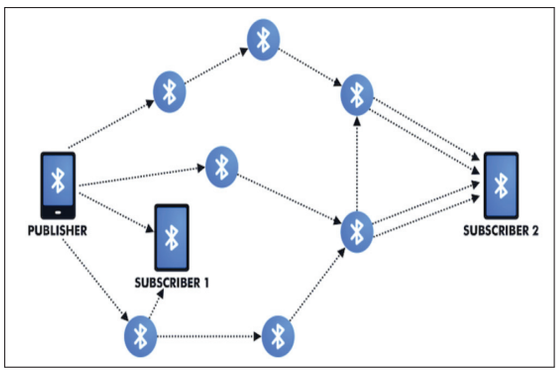


Fig. 1. An example of a BLE mesh network [3].

over extended range, making it ideal for short-range IoT applications.

When it comes to power efficiency, BLE is known for its ultra-low power consumption which is even more efficient in this aspect more than rivals protocols like ZigBee [5].

BLE routing protocols may be implemented using static and dynamic routing approaches, as suggested by academic routing-based solutions for BLE mesh networks explained in detail in [7].

III. Thread

Thread is open standard of mesh network base at IPv6 and 6LoWPAN technology [6]. It introduce a convenient way to build low power networks with direct access to the Internet. Thread uses IEEE 802.15.4 communication standard which is designed for low rate, low power WPANs (Wireless Area Networks) [8].

Thread network supports a connectivity model where the end devices are only connected to their parent nodes while all the routers are connected, forming a mesh network. Devices in Thread play different roles based on their functionalities, as shown in Figure 4 [9].

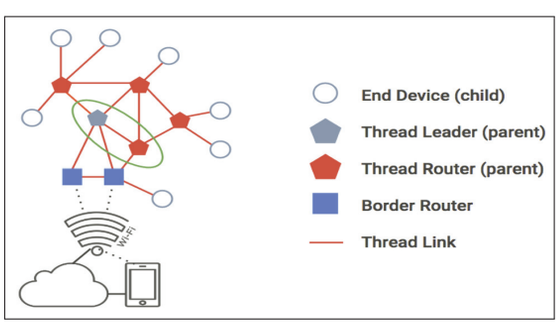


Fig. 2. Basic topology and device types of Thread network [10].

In terms of communication range in Thread, typical devices in conjunction with mesh networking provide more than enough range to cover a normal home [11]. Backbone Border Routers (BBRs) enable even larger IPv6 subnets composed of Thread devices, unifying multiple Thread networks into a single IPv6 subnet for the commercial market. Spread spectrum technology is used at the physical layer to provide good immunity to interference.

The Thread network uses 6LoWPAN "IPv6 over Low-Power Wireless Personal Area Networks" technology and IEEE 802.15.4 communication standard, both of which are specifically designed for low-data-rate, low-power Wireless Personal Area Networks (WPANs). Regarding power consumption, host devices can operate for several years on small batteries using suitable duty cycles [11].

From a reliability standpoint, Thread is a mesh network where each router node is connected to all available router nodes. If a connection breaks down between any two nodes, these nodes can still communicate as the network reroutes the traffic through another route. Another distinguishing feature of the Thread network is that it can support multiple border routers, these border routers provide connectivity from local mesh to the global Internet, so multiple border routers guarantee connectivity to the internet and make the network more reliable with no single point of failure [12]. While Thread supports the IPv6 architecture, DHCPv6 is used for router address assignment.

IV. ZigBee

ZigBee is a wireless communication protocol widely used in the Internet of Things (IoT), home automation, industrial control, and wireless sensor networks due to its energy efficiency and reliable mesh networking capabilities, ZigBee built on the IEEE 802.15.4 standard which has already been discussed above in this paper by emphasizing its beneficial capability which include, but are not limited to, low data rates and minimal power consumption.

The range in which Zigbee network can communicate can be up to 30 m indoors and up to 100 m Outdoors (line of sight) [15].

Mesh networking provides an efficient method for data routing, extending range by enabling data to hop between nodes. It also enhances reliability through self-healing capabilities, allowing the network to establish alternative paths if a node fails or a connection is disrupted. Nonetheless, according to the tests conducted in this study [16], Zigbee demonstrates excellent reliability and low latency (below the 200-millisecond).

A typical ZigBee mesh network consists of a coordinator and multiple routers and end device nodes, as shown in Figure 4.

ZigBee employs the Ad hoc On-Demand Distance Vector (AODV) routing protocol as part of its mesh networking capabilities, enabling efficient data transmission between nodes. AODV operates by dynamically discovering routes only when needed, reducing overhead and optimizing network performance [17].

V. Z-Wave

Z-Wave is a wireless communication protocol used primarily to integrate radio frequency sensors and actuators and provide smart home and office automation services. It is a

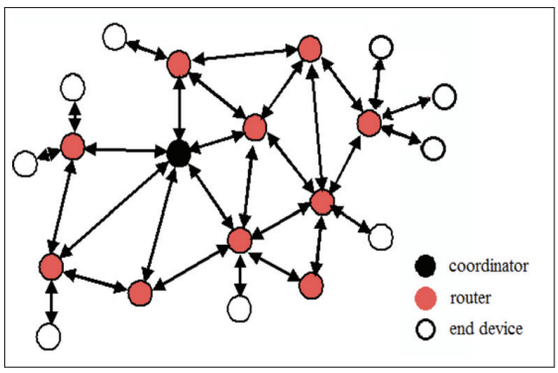


Fig. 3. Structure of typical ZigBee mesh network [14].

mesh network using low-energy radio waves to communicate from device to device, enabling wireless control of smart home devices [18].

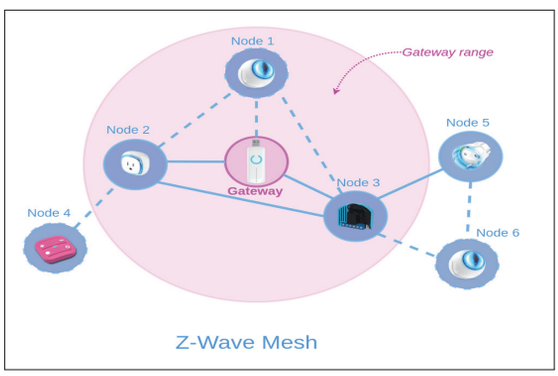


Fig. 4. Z-Wave mesh network [20].

Z-Wave has a communication range of up to 75 feet in a non-intrusive environment (i.e., with minimal interference), with an optimal range of 30 feet [19] (approximately 9–23 meters) .

Even though Z-Wave does not use IEEE 802.15.4 or 6LoWPAN (Unlike Zigbee and Thread) it still s a low powered mesh networking technology that has a battery which can last up to 10 years [21] and depend on its own proprietary MAC and PHY layers in addition it operates exclusively in the sub-GHz band (e.g., 908.42 MHz in the U.S., 868.42 MHz in Europe).

Z-Wave uses a source-routed mesh network topology and has one master (primary) controllers that control routing and security. Devices can communicate to another by using intermediate nodes to actively route around and circumvent household obstacles or radio dead spots that might occur [19].

VI. Evaluation

To provide a comprehensive comparison, Table 1 summarizes key performance characteristics of each protocol. The radar diagram in Figure 6 visually illustrates the trade-offs between these technologies, highlighting their strengths and limitations across different factors.

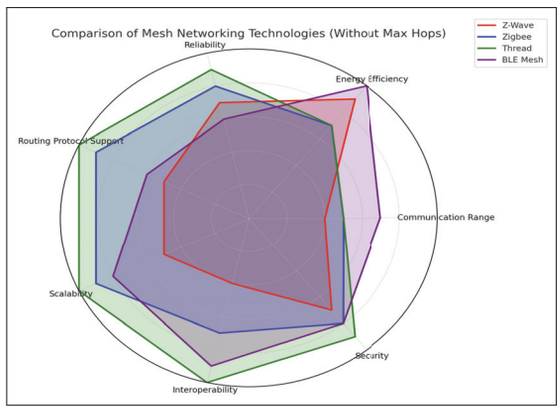


Fig. 5. Radar Diagram.

Generic Mathematical Model for Wireless Mesh Networks (WMN)

A wireless mesh network can be represented as a graph:

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}) \tag{1}$$

where $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}$ denotes the set of nodes and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ represents the set of links connecting nodes [21]. Each link (v_i, v_j) can be assigned a weight w_{ij} representing performance metrics such as SNR, bandwidth, path loss, and interference:

$$w_{ij} = f(\text{SNR}_{ij}, d_{ij}, B_{ij}, I_{ij}) \tag{2}$$

Topology is described via an adjacency matrix $A = [a_{ij}]$:

$$a_{ij} = \begin{cases} 1 & \text{if link exists between nodes } v_i \text{ and } v_j, \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

and the degree of node v_i is:

$$\deg(v_i) = \sum_{j=1}^{N} a_{ij} \tag{4}$$

Traffic flow $x_{ij}(t)$ between nodes must satisfy flow conservation for intermediate nodes:

$$\sum_{i:(v_i,v_k)\in\mathcal{E}} x_{ik}(t) = \sum_{j:(v_k,v_j)\in\mathcal{E}} x_{kj}(t), \quad \forall v_k \notin \{\text{sources, sinks}\} \tag{5}$$

with capacity constraint:

$$0 \le x_{ij}(t) \le C_{ij}, \quad \forall (v_i, v_j) \in \mathcal{E} \tag{6}$$

Wireless propagation effects such as path loss and fading are often modeled in terms of received power and a feasibility condition:

$$P_r = P_t G_t G_r L_p(d_{ij})^{-1}, \quad \text{SNR}_{ij} = \frac{P_r}{N_0} \ge \text{SNR}_{\text{min}} \tag{7}$$

Mesh network performance can be optimized by minimizing weighted flows or solving cross-layer mixed-integer programming formulations that jointly consider routing and scheduling [22]. For capacity analysis in multihop mesh networks, probabilistic throughput models and network graphs are widely used [23].

REFERENCES

VII. CONCLUSION

This paper presents a comparative analysis of four prominent mesh networking technologies for IoT applications: Z-Wave, Zigbee, Thread, and BLE, evaluating key factors such as communication range, energy efficiency, reliability, and routing protocols. In addition, we provide a generic mathematical model for wireless mesh networks, capturing node connectivity, traffic flow, and link performance in a technology-agnostic framework. As future work, we aim to explore how this model can be used to optimize the operation of existing mesh networks under the protocols investigated, providing insights for more efficient and reliable IoT deployments.

REFERENCES

[1] I. F. Akyildiz, X. Wang, and W. Wang, "Wireless mesh networks: A survey," Computer Networks, vol. 47, no. 4, pp. 445-487, 2005.

[2] Y. Liu and K.-F. Tong, "Wireless Mesh Networks in IoT Networks," Department of Electronic and Electrical Engineering, University College London, London, UK.

[3] A. Solovev and A. Petrova, "Bluetooth Mesh: Technology Overview, Examples, Alternatives, and First-Hand Experience," Integra Sources, Nov. 25, 2019. [Online]. Available: https://www.integrasources.com/blog/bluetooth-mesh-network-tutorial,

[4] M. Kähr, "Evaluating Bluetooth as a replacement for cables between sensor nodes," M.Sc. thesis, Dept. of Information Technology, Uppsala Univ., Uppsala, Sweden, 2012.

[5] M. E. Eltayeb, O. A. Saraereh, and M. M. Matalgah, "The Bluetooth Mesh Standard: An Overview and Experimental Evaluation," Sensors, vol. 18, no. 8, p. 2409, Aug. 2018. [Online]. Available: https://www.mdpi.com/1424-8220/18/8/2409

[6] M. E. Eltayeb, O. A. Saraereh, and M. M. Matalgah, "Bluetooth Low Energy Mesh Networks: Survey of Communication and Security Protocols," Sensors, vol. 20, no. 12, p. 3590, Jun. 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/12/3590

[7] M. S. J. Asad, M. A. Shah, M. A. Khattak, S. Khan, and M. Imran, "Bluetooth Low Energy Mesh Networks: A Survey," Sensors, vol. 17, no. 7, p. 1467, Jul. 2017. [Online]. Available: https://www.mdpi.com/1424-8220/17/7/1467

[8] S. H. Park, S. H. Park, and S. H. Park, "Performance Analysis of Thread Protocol in Wireless Mesh Networks," Proceedings of the 2015 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, Jan. 2015, pp. 169-170. [Online]. Available: https://people.computing.clemson.edu/ jmarty/projects/lowLatencyNetworking/papers/WiFi-Manets-Mesh/Mesh/ieee1504threadMeshPerf.pdf

[9] S. B. A. Khattak, M. M. Nasralla, H. Farman, and N. Choudhury, "Performance Evaluation of an IEEE 802.15.4-Based Thread Network for Efficient Internet of Things Communications in Smart Cities," Applied Sciences, vol. 13, no. 13, p. 7745, Jun. 2023. [Online]. Available: https://www.mdpi.com/2076-3417/13/13/7745

[10] Thread Network Fundamentals, White Paper. September 2022. Available online: https://www.threadgroup.org/support (accessed on 31 Jan 2025).

[11] Thread Network Fundamentals, Thread Group, May 2020. [Online]. Available: https://www.threadgroup.org/Portals/0/documents/support/Thread

[12] S. B. A. Khattak, M. M. Nasralla, H. Farman, and N. Choudhury, "Performance Evaluation of an IEEE 802.15.4-Based Thread Network for Efficient Internet of Things Communications in Smart Cities," Applied Sciences, vol. 13, no. 13, p. 7745, Jun. 2023. [Online]. Available: https://www.mdpi.com/2076-3417/13/13/7745

[13] Y. Qiang and F. Tian, "Design of network safety monitoring system based on ZigBee Mesh in mine," Journal of Xi'an University of Science and Technology, vol. 37, pp. 334-338, 2017.

[14] Y. Zhang, X. Li, and Z. Wang, "Design of Wireless Sensor Network Node System Based on ZigBee," Journal of Physics: Conference Series, vol. 1346, no. 1, p. 012046, Nov. 2019.

[15] An Introduction to ZigBee, Product Manual, Digi International, 2008.

[16] Silicon Labs, "AN1138: Zigbee Mesh Network Performance," Application Note, Rev. 0.3, pp. 1-11, Apr. 2019. [Online]. Available: https://www.silabs.com/documents/login/application-notes/an1138-zigbee-mesh-network-performance.pdf

[17] A. Hussein and A. Samara, "Coordinator Location Effects in AODV Routing Protocol in ZigBee Mesh Network," arXiv preprint arXiv:1712.04245, Dec. 2017.

[18] R. B. Luh, A. R. A. Ali, and M. M. Siraj, "Analysis and Security Enhancement of Z-Wave-Based IoT Networks," International Journal of Network Security and Its Applications (IJNSA), vol. 16, no. 6, pp. 27–41, 2024.

[19] Braincoke, "An Introduction to Z-Wave for the End User," Braincoke Blog, Dec. 2020. [Online]. Available: https://braincoke.fr/blog/2020/12/an-introduction-to-z-wave-for-the-end-user/#mesh-network.

[20] E. Zanaj, G. Caso, L. De Nardis, A. Mohammadpour, Ö. Alay, and M.-G. Di Benedetto, "Energy Efficiency in Short and Wide-Area IoT Technologies—A Survey," Technologies, vol. 9, no. 1, p. 22, Mar. 2021. [Online]. Available: https://www.mdpi.com/2227-7080/9/1/22.

[21] I. F. Akyildiz, X. Wang, and W. Wang, "Wireless mesh networks: a survey," *Computer Networks*, vol. 47, no. 4, pp. 445–487, 2005. :contentReference[oaicite:0]index=0

[22] L. Chen, Q. Zhang, and W. Shu, "Optimization models and methods for planning wireless mesh networks," *Computer Networks*, vol. 52, no. 11, pp. 2159–2171, 2008. :contentReference[oaicite:1]index=1

[23] P. Gupta and P. R. Kumar, "Modeling and computing throughput capacity of wireless multihop networks," *Computer Networks*, vol. 52, no. 1, pp. 116–129, 2008. :contentReference[oaicite:2]index=2

TABLE I
COMPARISON OF MESH NETWORKING TECHNOLOGIES FOR IOT (Z-WAVE, ZIGBEE, BLE,AND THREAD)

Feature	Z-Wave	Zigbee	Thread	BLE Mesh
Frequency Band	Sub-GHz (908.42 MHz, 868.42 MHz)	2.4 GHz ISM band	2.4 GHz ISM band	2.4 GHz ISM band
Topology	Source-routed mesh with controller-based routing [19]	Self-healing mesh using AODV routing [17]	IPv6-based self-healing mesh with border routers [8]	Flooding-based message relay with managed relay nodes
Routing Protocol	Source Routing [19]	AODV (Ad hoc On-Demand Distance Vector) [17]	6LoWPAN-based mesh routing with rerouting [8]	Flooding or managed flooding
Max Hops	4 hops	30+ hops	32 hops	No strict limit (depends on relays)
Communication Range	30–75 feet (9–23m optimal) indoors [19]	30m (indoor), 100m (outdoor) [15]	More than enough range to cover a normal home [11]	In ideal conditions, up to 100m (Class 1 BLE)[4]; in practical scenarios, its effective range is typically limited to around 50m.
Power Efficiency	Ultra-low power (up to 10 years battery life) [21]	Low power, optimized duty cycling	Low power, efficient sleep cycles	Ultra-low power consumption (more power efficient than ZigBee) [5]
Scalability	Up to 232 nodes per network	Up to 65,000+ nodes	Highly scalable (IPv6) [8]	Supports thousands of nodes
Interoperability	Proprietary standard, requires Z-Wave hub	IEEE 802.15.4, Zigbee-certified devices	IEEE 802.15.4, fully IP-compatible with IPv6	Fully Bluetooth-compatible
Security	AES-128 encryption, centralized key management	AES-128 encryption, Zigbee 3.0 security model	AES-128, IPv6 security mechanisms	AES-128 with secure key exchange
Reliability	Controller-based, limited hops	Self-healing mesh, moderate interference resistance	Multi-border router support [12], IPv6-based [8]	Flooding-based, higher risk of congestion
Best Use Cases	Smart homes, automation	Home automation, industrial IoT, smart sensors	Enterprise IoT, large-scale automation	Smart lighting, wearables, retail tracking

AI-Driven Sleep Apnea Screening Based on Wireless Abdominal Wearable Sensor

Thi Hang Dang *, Sukmin Lee†, Seungcheon Kim‡, Franklin Bien*†

* *Department of Electrical Engineering, Ulsan National Institute of Science and Technology, Ulsan, Republic of Korea*

† *SB Solutions Inc., Ulsan National Institue of Science and Technology, Ulsan, Republic of Korea*

‡ *Department of Convergence Security, Hansung University, Seoul, Republic of Korea*

Emails: *thihangdang@unist.ac.kr, smlee@sb-solutions.co.kr, kimsc@hansung.ac.kr, bien@unist.ac.kr*

Abstract **Obstructive sleep apnea (OSA) is prevalent among older adults and individuals with cardiovascular disease. Diagnosis is typically performed using polysomnography (PSG), which entails substantial costs, extensive resource utilization, and limited patient accessibility. Recently, wireless wearable technologies have emerged as promising tools for OSA screening and longitudinal monitoring. This study presented a novel approach for OSA detection using abdominal movement and acceleration signals acquired from a wireless abdomen-worn sensor (Soomirang). Two-hundred participants underwent simultaneous overnight monitoring with both PSG and Soomirang systems. Non-apnea and apnea events were classified using an MLP-Mixer based deep learning model applied to Soomirang data. The classified outcomes were used to calculate the apnea hypopnea index (AHI), with AHI ≥ 15 indicating moderate-to-severe OSA. Agreement between AHI derived from the Soomirang device and PSG was assessed using Pearson correlation and Bland–Altman analyses. The estimated AHI demonstrated good agreement with the reference, yielding a correlation coefficient of 0.92 and a mean difference of 3 events per hour. For moderate-to-severe OSA detection, the Soomirang system achieved an accuracy of 89%, sensitivity of 87.1%, specificity of 92.6%, and F1-score of 97.1%. Overall, the proposed algorithm based on a wireless abdomen-worn sensor demonstrated excellent performance in identifying moderate-to-severe OSA, highlighting the potential of a simple and accessible wearable device for OSA screening and follow-up.**

Keywords—obstructive sleep apnea, home sleep apnea test, abdomen-worn sensor

I. Introduction

Obstructive sleep apnea (OSA) is defined by recurrent episodes of partial or complete upper airway obstruction lasting at least 10 seconds, typically accompanied by cortical arousal or a reduction in blood oxygen saturation [1,2]. OSA severity is quantified using the apnea–hypopnea index (AHI, the average number of apnea and hypopnea events occurring per hour of sleep) and categorized as mild ($5 \leq AHI < 15$), moderate ($15 \leq AHI < 30$), or severe ($AHI \geq 30$) [1,2]. OSA is associated with an increased risk of cardiovascular disease [3] and affects approximately 20–60% of the adult population, with prevalence rising to nearly 75% among individuals aged 65 years and older [4,5]. These figures underscore the substantial public health burden of OSA and the urgent need for effective diagnostic and management strategies. Nonetheless, OSA diagnosis primarily depends on PSG, which is associated with high cost, substantial resource requirements, and limited patient convenience.

AI-enabled wireless wearable devices have emerged as promising alternatives to address the limitations of PSG. This study presents an automated deep learning-based algorithm for detecting moderate-to-severe OSA using data collected by a wireless abdomen-worn sensor (Soomirang), which integrates a capacitive sensor for abdominal movement measurement and a 3-axis accelerometer.

II. Methods and Materials

A. Data collection

Two-hundred participants with suspected OSA underwent overnight in-laboratory PSG (MiniScreen, Lowenstein, Germany) alongside simultaneous data collection from an abdomen-worn sensor (SB Solutions Inc., Ulsan, Republic of Korea). The Soomirang device was attached to the belly between the navel and the chest to measure signals at a sampling rate of 5 Hz.

Respiratory events were manually scored by a certificated sleep clinician using PSG data in accordance with established American Academy of Sleep Medicine (AASM) guidelines [2]. An apnea event was defined as a reduction in airflow of at least 90% from baseline lasting for a minimum duration of 10 seconds. Hypopnea was defined as a reduction in airflow of at least 30% lasting at least 10 seconds and accompanied by either a ≥3% decrease in arterial oxygen saturation or an associated cortical arousal. The AHI was subsequently calculated as the total number of apnea and hypopnea events per hour of sleep (AHI_{PSG}), and OSA severity was classified accordingly, with moderate-to-severe OSA defined as an $AHI_{PSG} \geq 15$ events per hour.

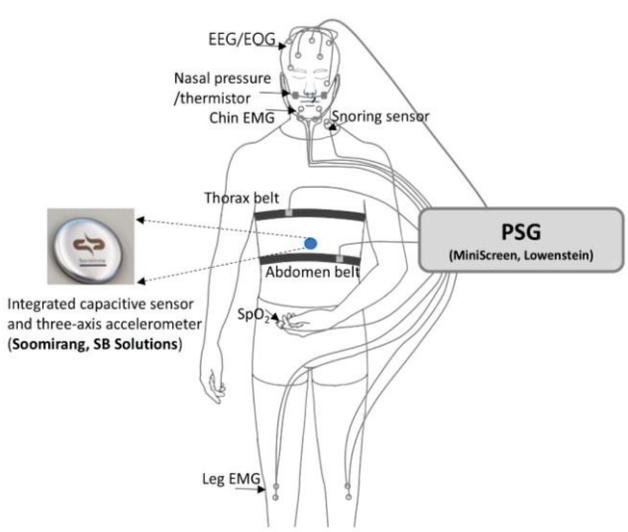


Fig. 1. Experimental setup for a full-night sleep monitoring using polysomnography and a wireless abdomen-worn sensor

B. Preprocessing

The data acquired from the PSG and Soomirang systems were temporally aligned prior to analysis. Breathing events were was categorized as either apnea events (obstructive, central, mixed apneas, and hypopneas) or non-apnea breathings. The onset and offset times of apnea events were extracted directly from PSG annotation files. Based on these annotations, Soomirang samples were assigned apnea labels if their timestamps overlapped with the annotated apnea event intervals; samples outside these intervals were labeled as non-apnea.

The capacitance signal recorded by the Soomirang device was denoised using a Savitzky–Golay smoothing filter with a window size of 11 samples and a polynomial order of 3, followed by signal normalization [6]. The choice of a third-order polynomial allowed the filter to adequately preserve both local maxima and minima in the respiratory waveform. The window length was determined according to the formulation proposed by Ross et al., assuming a nominal respiratory rate of 15 breaths per minute, while accounting for the broader range of respiratory rates observed during apnea episodes (0–30 breaths per minute) [7]. Within each sliding window, the capacitance signal was standardized by subtracting the local mean and scaling by the corresponding standard deviation. To characterize short-term signal variability on a per-second basis, a local amplitude feature (local gap) was computed as the difference between the maximum and minimum values within a five-sample moving window.

Sleep posture information was derived from the three-axis accelerometer data, and each time sample was classified into one of five sleep positions: upright, supine, prone, left lateral, or right lateral, based on gravitational acceleration components and tilt angles [8].

In addition, the magnitude of the acceleration vector was computed as the square root of the sum of squared acceleration components across all three axes. Fig. 2 illustrates an example of data recording from PSG and Soomirang systems, including airflow, respiratory effort, and oxygen saturation signals from the PSG, alongside capacitance and tri-axial acceleration signals from the Soomirang device. Apnea events identified by the PSG system are indicated by orange-shaded regions aligned with the respiratory flow signal.

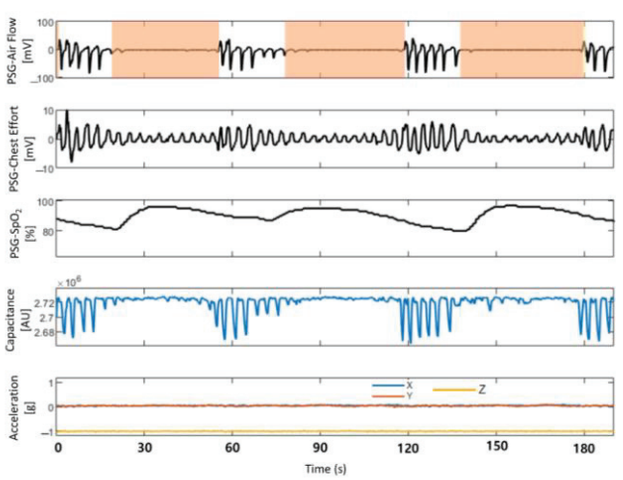


Fig. 2. Example of data recording from PSG and Soomirang. The orange areas represent apnea events detected by PSG

C. MLP-Mixer model

The MLP-Mixer model was employed for automated apnea/non-apnea classification (Fig.3). The MLP-Mixer architecture, originally proposed by Tolstikhin et al. for image classification, was modified in this work to accommodate time-series analysis and implemented using the PyTorch framework (version 2.5.1) [9]. The adapted model processes input data in fixed-length segments by applying a sliding window of 100 seconds across the entire overnight recording. Consequently, each model input is represented as a tensor of size 500 × 4, corresponding to 500 temporal patches and four feature channels.

The four input channels comprise features derived from the Soomirang device, as detailed in the preprocessing stage, namely normalized capacitance, acceleration vector magnitude, local gap, and sleep posture. Prior to feature mixing, the input tensor is transposed such that each temporal patch is treated as an independent token and passed through a shared multilayer perceptron for token-wise feature interaction (token mixing). Each multilayer perceptron block consists of two fully connected layers interleaved with dropout layers and employs the Gaussian Error Linear Unit (GELU) as the activation function [10]. Weight parameters are shared across all patches within the same feature channel.

Following token mixing, the tensor is reshaped back to its original configuration, and a second multilayer perceptron is applied uniformly across feature channels to model inter-channel dependencies (channel mixing). In the implemented architecture, token mixing is realized using one-dimensional convolutional layers with a kernel size of one, whereas channel mixing is achieved through fully connected linear layers. Residual (skip) connections are incorporated in both mixing stages to facilitate stable gradient propagation during training. A sigmoid activation function is applied at the output layer to produce probabilities for the classification task.

Model training was conducted using a five-fold cross-validation scheme. The training configuration included a batch size of 500, a window length of 500 samples, a hidden layer dimension of 128, and a total of 50,000 training iterations. Optimization was performed using the Adam optimizer with a learning rate of 0.001.

The classified outcomes were used to calculate the AHI as the ratio between the total number of apnea events and total recording time ($AHI_{Soomirang}$). The $AHI_{Soomirang} \geq 15$ indicating moderate-to-severe OSA.

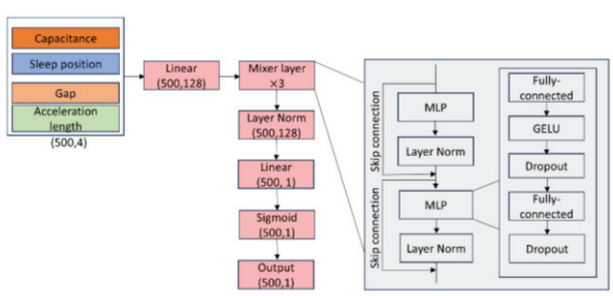


Fig. 3. MLP-Mixer modeal for apnea/non-apnea classification

D. Analysis

Model performance was evaluated using sensitivity, specificity, accuracy, and F1-score. Agreement between

AHI_{PSG} and $AHI_{Soomirang}$ was assessed through Pearson correlation and Bland-Altman analysis.

III. RESULT

Seven recordings were excluded due to low-quality data, leaving 100 recordings for training and validation, and 93 recordings for testing. Fig.4-5 show the Pearson correlation and Bland-Alterman plots in a comparison between estimated $AHI_{Soomirang}$ and AHI_{PSG} on the test set.

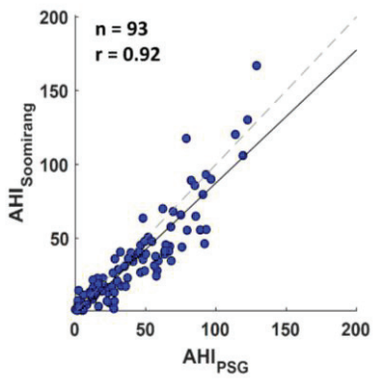


Fig. 4. Pearson correlation between AHI calculated by Soomirang device ($AHI_{Soomirang}$) and PSG (AHI_{PSG})

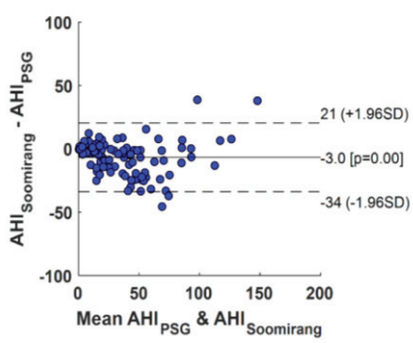


Fig. 5. Bland-Alterman plot shows the agreement of AHI calculated by Soomirang device ($AHI_{Soomirang}$) and PSG (AHI_{PSG})

The correlation coefficient (r) between AHI_{PSG} and $AHI_{Soomirang}$ was 0.92. Bland-Altman analysis indicated a mean difference between AHI_{PSG} and $AHI_{Soomirang}$ of 3 events per hour, with limits of agreement averaging 27.5 events per hour.

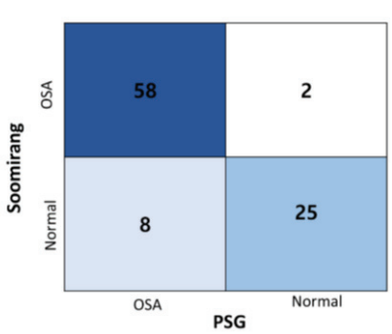


Fig. 6. Confusion matrix shows the performance for OSA prediction of Soomirang (OSA: AHI≥15, Normal: AHI<15).

The model achieved an accuracy of 89%, a sensitivity of 87.1%, a specificity of 92.6%, and an F1-score of 91.7% for detecting moderate-to-severe OSA.

IV. DISCUSSION

A. Main findings

This study presents an automated algorithm for detecting OSA using data collected by a wireless abdomen-worn sensor, which integrates a capacitive sensor for abdominal movement and a three-axis accelerometer body motion measurements. The new compact Soomirang device with an integrated novel automatic algorithm demonstrated excellent performance for OSA detection, especially in moderate-to-severe OSA patients. The results revealed the good performance of the MLP-Mixer model in an application for OSA detection.

B. Comparisons with existing devices

Gu et al. reported a strong agreement between AHI values derived from the Belun Ring and PSG, demonstrating a correlation coefficient of 0.894 and achieving a sensitivity of 87.5% and a specificity of 87% at an AHI threshold of ≥15 [11]. Similarly, Ghahjaverestan et al. observed a high correlation (r = 0.91) between PSG-derived AHI and measurements obtained using the BresoDX1 system, which integrates tracheal–sternal motion, tracheal sound, and oximetry, with corresponding sensitivity and specificity of 90% and 85.9%, respectively, for AHI ≥15 [12]. Papini et al. reported more modest performance for a smartwatch-based model trained on 250 recordings, yielding an AHI correlation of 0.61, sensitivity of 59%, and specificity of 90% [13]. Yeh et al. demonstrated that a smartwatch-based approach achieved an overall accuracy of 80.8%, with sensitivity and specificity of 93.1% and 73.5%, respectively [14]. Despite differences in datasets, sensing modalities, and scoring methodologies, these results suggest that the performance of the Soomirang device is comparable to that of previously reported wearable-based OSA screening systems.

C. Limitations

This study has several limitations. First, the AHI is conventionally defined as the number of apnea and hypopnea events per hour of sleep. In the present study, sleep time was not explicitly estimated; instead, $AHI_{Soomirang}$ was calculated using the total recording time, which may have led to an underestimation or overestimation of AHI and reduced its accuracy. Incorporating sleep–wake or sleep stage detection into the algorithm could improve the precision of AHI estimation. Second, data collection was performed under supervised conditions using simultaneous PSG measurements. Although this approach enabled reliable comparison, the experimental setting may differ from typical home sleeping environments. Additional studies conducted in unsupervised home settings would be valuable to further evaluate the system's performance and robustness during natural sleep.

V. CONCLUSION

This study proposes an automated algorithm for the detection of moderate to severe obstructive sleep apnea using signals acquired from a wireless abdomen-worn sensor. The Soomirang device integrates a capacitive sensor for monitoring abdominal respiratory motion with a tri-axial accelerometer for capturing body movement and posture. Owing to its suitability for home-based use, lightweight design, and cost-effective and repeatable measurements, the

system shows promise for large-scale OSA screening. The results indicate that the Soomirang system may serve as a practical and non-invasive screening tool for obstructive sleep apnea. Nevertheless, further studies involving larger populations and real-world home environments, as well as assessments of long-term usability and patient comfort, are required to validate its clinical applicability.

ACKNOWLEDGMENT

This work was supported by Innovative Human Resource Development for Local Intellectualization program through the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT)(IITP-2025-RS-2022-00156361), and the Technology Innovation Program(or Industrial Strategic Technology Development Program)(RS-2023-00236657, Development of implantable artificial intelligence convergence sensor module technology for diagnosis and prediction of cardiovascular disease) funded by the Ministry of Trade Industry & Energy(MOTIE, Korea).

REFERENCES

[1] D. J. Gottlieb and N. M. Punjabi, "Diagnosis and management of obstructive sleep apnea: A review," *JAMA*, vol. 323, no. 14, pp. 1389–1400, 2020.

[2] V. K. Kapur *et al*., "Clinical practice guideline for diagnostic testing for adult obstructive sleep apnea: An American Academy of Sleep Medicine clinical practice guideline," *J. Clin. Sleep Med.*, vol. 13, no. 3, pp. 479–504, 2017.

[3] Y. Yeghiazarians *et al*., "Obstructive sleep apnea and cardiovascular disease: A scientific statement from the American Heart Association," *Circulation*, vol. 144, no. 3, pp. e56–e67, 2021.

[4] R. S. Osorio, M. Á. Martínez-García, and D. M. Rapoport, "Sleep apnoea in the elderly: A great challenge for the future," *Eur. Respir. J.*, vol. 59, no. 4, 2022.

[5] P. Balagny *et al*., "Prevalence, treatment and determinants of obstructive sleep apnoea and its symptoms in a population-based French cohort," *ERJ Open Res.*, vol. 9, no. 3, 2023.

[6] N. B. Gallagher, "Savitzky–Golay smoothing and differentiation filter," Eigenvector Research Inc., Manson, WA, USA, Tech. Rep., 2020.

[7] R. Ross *et al*., "An adaptively parameterized algorithm estimating respiratory rate from a passive wearable RFID smart garment," in *Proc. IEEE 45th Annu. Comput., Softw., Appl. Conf. (COMPSAC)*, Jul. 2021, pp. 774–784.

[8] A. R. Friederich, M. L. Audu, and R. J. Triolo, "Trunk posture from randomly oriented accelerometers," *Sensors*, vol. 22, no. 19, Art. no. 7690, 2022.

[9] I. O. Tolstikhin *et al*., "MLP-Mixer: An all-MLP architecture for vision," in *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 24261–24272, 2021.

[10] M. Lee, "Mathematical analysis and performance evaluation of the GELU activation function in deep learning," *J. Math.*, vol. 2023, Art. no. 4229924, 2023.

[11] W. Gu *et al*., "Belun Ring Platform: A novel home sleep apnea testing system for assessment of obstructive sleep apnea," *J. Clin. Sleep Med.*, vol. 16, no. 9, pp. 1611–1617, 2020.

[12] N. Montazeri Ghahjaverestan *et al*., "Sleep apnea detection by tracheal motion and sound, and oximetry via application of deep neural networks," *Nat. Sci. Sleep*, pp. 423–432, 2023.

[13] G. B. Papini *et al*., "Wearable monitoring of sleep-disordered breathing: Estimation of the apnea–hypopnea index using wrist-worn reflective photoplethysmography," *Sci. Rep.*, vol. 10, no. 1, Art. no. 13512, 2020.

[14] E. Yeh *et al*., "Detection of obstructive sleep apnea using Belun Sleep Platform wearable with neural network-based algorithm and its combined use with STOP-Bang questionnaire," *PLoS One*, vol. 16, no. 10, Art. no. e0258040, 2021.

Lightweight AI-Generated Text Detection under Domain Ambiguity via Latent Lexical Domains

Jisoo Lee *, Hyemyung Lee †, Skyler Chun ‡, Sang-Yep Nam §
* *Department of computer Science& software Engineering. Korea University, Republic of Korea*
† *Department of Software. Kookje University, Pyeongtaek-si, Republic of Korea*
‡ *Ridgewood.H.S, United States of America*
§ *Hands On Technology, Republic of Korea*
Emails: *jisoo1129a@gmail.com, hml22@hanmail.net, sangminsky08@gmail.com, synam58@gmail.com*

***Abstract*—As large language models continue to advance their performance to a nearly human level, distinguishing AI-generated text from human writing has become a significant task to ensure content authenticity and clarify authorship across academic and journalistic productions. While existing detectors that use training and fine-tuning methodology often rely on explicitly defined domains such as news or social media, these rigid domains are misaligned with real-world content, which contains mixed or ambiguous characteristics. Therefore, we propose word distribution driven domains as selecting them that do not demarcate to predefined ones. Using unsupervised clustering over document-level lexical representations, we categorized clusters to distributionally close subspaces without relying on manual domain labels. This approach preserves the key advantage of Naive Bayes, the efficiency in computation, and relatively high and consistent accuracy.**

Keywords—Detection, Artificial Intelligence, Naïve bayes Multinomial, NLP

I. Introduction

The continued advancement of large language models(LLMs) has brought AI-generated text to a level of fluency comparable to human writing. As a result, reliably distinguishing AI-generated content from human-authored text has become important for ensuring content authenticity, academic integrity, and clear authorship attribution[4, 5, 6]. A wide range of detection approaches has been proposed, including perplexity-based statistical methods and large neural classifiers[7]. Many of these methods, however, implicitly assume that texts belong to clearly defined domains—such as news, social media, or academic writing—and that these domain boundaries are stable across datasets.

In real-world settings, this assumption often meets counterexamples. Texts within the same nominal domain often vary widely in lexical choice and stylistic structure, and many documents naturally combine features from multiple genres. Prior studies have shown that NLP models suffer notable performance degradation under domain shift, where discrepancies between training and evaluation distributions reduce robustness and generalization. These observations indicate that manually defined domain labels may not accurately reflect the underlying distributional structure of language, particularly in heterogeneous or evolving text collections.

We explore an alternative approach that addresses domain-related instability without relying on explicit domain annotations. Rather than treating domain as a fixed external label, we induce latent lexical domains directly from similarities in word distributions across documents. Through unsupervised clustering of document-level lexical representations, we partition the clusters into distributionally coherent subspaces that capture intrinsic patterns in the data. Within each latent domain, we apply Multinomial Naive Bayes to estimate class-conditional word probabilities for human-written and AI-generated text, enabling localized and context-sensitive detection.

By conditioning detection on distribution-induced subspaces instead of rigid, predefined domains, the proposed framework retains the computational efficiency and interpretability of Naive Bayes while alleviating the effects of domain ambiguity and distributional mismatch. Summarizing it, this work demonstrates that inducing latent lexical domains from word distributions offers a lightweight yet principled way to improve the stability of AI-generated text detection in heterogeneous environments.

II. Related work

Fernando et al. (2025) applied TF-IDF features to a Multinomial Naive Bayes classifier in comparison of human and ChatGPT and reported high enough detection accuracy while maintaining low computational requirements [1]. Additionally, performance in plagiarism detection and misinformation analysis highlighted the practical value of probabilistic models in handling unstructured textual data. While deep-learning based detectors such as RoBERTa or other transformer architectures often achieve higher accuracy under controlled conditions, they require considerable resources and lack transparency in decision-making. This trade-off draws attention to lightweight probabilistic approaches.

Bhattacharjee et al. (2023) examined AI-generated text detection under varying generation conditions and demonstrated that detectors trained on text from a particular source or generation setting often experience performance degradation when applied to data generated under different conditions [2]. Their study highlights how changes in underlying text distributions can function as implicit domain shifts, motivating methods that account for such variability during detection.

These findings suggest that manually defined domain labels do not always align with the complex distributional structure of textual data. Rather than indicating fundamental shortcomings of existing detection models, this line of research emphasizes the importance of incorporating distribution-aware mechanisms to improve robustness under domain ambiguity.

Building on prior work in lightweight probabilistic text classification and domain-sensitive detection, the present study explores a complementary direction. Instead of relying on explicit domain annotations or complex adaptation mechanisms, we induce latent lexical domains directly from word distribution similarities in the data and perform authorship inference within these induced subspaces. This distribution-driven conditioning enables Naive Bayes models to retain their computational efficiency and interpretability while addressing domain sensitivity through localized probabilistic modeling.

III. Methodology

A. Lexical Representation

Each document is first represented using a lexical feature vector that captures word distribution characteristics while maintaining computational efficiency. Following standard practice in lightweight text classification, we apply common preprocessing steps including lowercasing, punctuation removal, tokenization, and stop word filtering.

To represent documents in a fixed-dimensional space, we employ term-frequency–inverse document frequency (TF-IDF) features. This representation emphasizes discriminative lexical patterns while remaining compatible with probabilistic models such as Multinomial Naive Bayes. Importantly, no semantic embeddings or pretrained language models are used at this stage, ensuring that the representation remains lightweight and interpretable.

B. Induction of Latent Lexical Domains

Rather than relying on predefined domain labels, we induce latent lexical domains directly from the data based on similarities in word distribution. Given the TF-IDF representations of documents, we perform unsupervised clustering to partition the corpus into distributionally coherent subsets. Each cluster corresponds to a latent lexical domain, reflecting intrinsic regularities in vocabulary usage and stylistic patterns.

Crucially, the clustering process is performed without using authorship labels (human or AI), ensuring that domain induction remains independent of the detection task itself. By grouping documents that share similar lexical distributions, this step aims to reduce heterogeneity within each subset and to expose localized patterns that may be obscured in a global model.

C. Domain-Conditioned Naive Bayes Modeling

Within each latent lexical domain, we train a separate Multinomial Naive Bayes classifier to model the conditional word distributions of human-written and AI-generated text. For a given domain k, the model estimates class-conditional probabilities $P(w \mid c, k)$, where w denotes a word and $c \in \{\text{Human, AI}\}$.

This domain-conditioned formulation allows the classifier to capture distribution-specific lexical regularities while preserving the core advantages of Naive Bayes, including low computational cost and probabilistic interpretability. Compared to a single global model trained over the entire corpus, domain-conditioned modeling reduces the impact of distributional mismatch by restricting inference to locally homogeneous subsets.

D. Inference for Unseen Texts

For an unseen input document, the detection process proceeds in two steps. First, the document is mapped into the lexical feature space and assigned to its closest latent lexical domain based on distributional similarity. Second, the Naive Bayes model associated with the selected domain is used to compute the posterior probability of AI authorship.

Formally, given an input document xxx and its assigned domain $k*$, the detector estimates $P(\text{AI} \mid x, k*)$ using the domain-conditioned model. This two-stage inference enables localized probabilistic reasoning while maintaining a simple and efficient pipeline. Documents with low confidence scores may optionally be flagged as ambiguous, although the present study focuses on binary classification.

IV. Experimental Results and Disscussion

A. Experimental Setup

We evaluate the proposed distribution-aware Naive Bayes framework on a dataset consisting of human-written and AI-generated texts. All documents are preprocessed using standard normalization procedures, including lowercasing, tokenization, and stop word removal, and are represented using TF-IDF features.

We apply K-means clustering over TF-IDF document representations to induce latent lexical domains, using cosine similarity as the measure of distributional proximity. Within each induced latent lexical domain, a separate Multinomial Naive Bayes classifier is trained to estimate class-conditional word distributions for human-written and AI-generated text, using Laplace smoothing. For an unseen document, domain assignment is performed by selecting the nearest cluster centroid based on cosine similarity, after which inference is conducted using the corresponding domain-conditioned classifier.

For evaluation, we compare the proposed domain-conditioned Naive Bayes model with a global Multinomial Naive Bayes baseline, which is trained on the entire dataset without domain partitioning. Both models use identical lexical representations to ensure a fair comparison. Performance is measured using accuracy, F1-score, and false positive rates (FPRs). Results are reported as averages across multiple evaluation splits to reduce variance.

B. Detection Performance

The proposed model demonstrates more stable and consistent performance compared to the global baseline. The global Naive Bayes classifier achieves an average accuracy of 91.8% with an F1-score of 91.6%, reflecting the effectiveness of lightweight probabilistic modeling for AI-generated text detection. By conditioning inference on latent lexical domains, the proposed method improves overall accuracy to 93.2% and achieves an F1-score of 93.0%.

While the gain in accuracy is moderate, the improvement is consistent, indicating enhanced robustness in heterogeneous text collections. These results suggest that performance gains arise not from increased model complexity, but from improved alignment between probabilistic modeling assumptions and the local distributional structure of the data.

C. Discussion

We analyze detection behavior across induced domains and compare it directly with the Naive Bayes model to examine the impact of domain conditioning. The results indicate that latent domains characterized by higher lexical coherence exhibit clearer separation between human-written and AI-generated texts, leading to more confident and stable posterior estimates under the proposed framework. This suggests that partitioning the data into distributionally homogeneous subspaces allows lightweight probabilistic models to better align their assumptions with local lexical regularities.

In contrast, the Naive Bayes model demonstrates greater variability in prediction confidence. This variability results in higher false positive rates for human-written texts, which reflect the limitations of modeling under heterogeneous distributions; the false positive rate decreases from 8.2% under the baseline model to 6.1% when latent domain conditioning is applied, indicating a meaningful reduction in cases where human-authored content is incorrectly flagged as AI-generated.

In conclusion, the experimental results validate the central methodological premise of this work: lightweight probabilistic classifiers benefit substantially from distribution-aware conditioning under domain ambiguity. Rather than relying on explicitly defined domain labels or resource-intensive adaptation mechanisms, the proposed approach leverages intrinsic lexical distribution structure to guide localized inference.

By aligning probabilistic inference with locally coherent distributional subspaces, the proposed method demonstrates that Naive Bayes remains a viable and interpretable option for AI-generated text detection in heterogeneous environments. More broadly, these results support the use of latent lexical domains as a principled mechanism for improving robustness, reinforcing the value of distribution-aware probabilistic modeling as a lightweight alternative to more complex detection architectures.

V. Conclusion and Future Work

A. Conclusion

This work presented a lightweight, distribution-aware framework for AI-generated text detection that addresses a fundamental limitation of domain-dependent detection methods: the mismatch between manually defined domain labels and the intrinsic distributional structure of real-world text. Prior evidence has shown that domain shifts and heterogeneous lexical patterns undermine detection robustness. Motivated by the research, we reframed domain not as an external categorical attribute, but as an emergent property induced directly from word distribution similarities.

By combining unsupervised induction of latent lexical domains with domain-conditioned Multinomial Naive Bayes modeling, the proposed approach preserves the computational efficiency and interpretability of traditional probabilistic classifiers. At the same time, it improves detection stability under domain ambiguity. Importantly, the observed performance gains arise not from increased model complexity or richer representations, but from better alignment between probabilistic modeling assumptions and locally coherent lexical distributions. This alignment enables lightweight models to operate more reliably in heterogeneous environments where global assumptions are often violated.

The empirical results provide concrete support for this formulation. Conditioning inference on distribution-induced subspaces leads to a consistent reduction in false positives and a stabilization of prediction confidence across diverse textual patterns. These effects indicate improved calibration and robustness without sacrificing efficiency. The findings further suggest that limitations commonly attributed to lightweight detectors stem primarily from global modeling assumptions rather than insufficient representational capacity.

More broadly, this work positions latent lexical domains as a principled alternative to explicit domain labeling and resource-intensive adaptation strategies. Rather than attempting to enumerate or predict domain categories a priori, the proposed framework leverages intrinsic distributional structure to guide localized inference. This offers a practical and interpretable pathway toward robust detection in evolving and mixed-domain settings.

Several directions remain for future research. While latent lexical domains were induced via unsupervised clustering in this study, exploring alternative similarity measures and clustering strategies may further improve domain stability and interpretability. A systematic analysis of cluster quality and its relationship to detection performance would strengthen the theoretical grounding of distribution-induced conditioning. Extending the framework to other probabilistic or linear classifiers would also help disentangle the effects of domain conditioning from model-specific characteristics and clarify the generality of the approach.

Further experimental validation across a wider range of datasets, text lengths, and generation conditions would enhance confidence in real-world applicability. This is particularly important under explicit domain shift and mixed-style inputs. Finally, the distribution-aware structure introduced in this work naturally lends itself to uncertainty-aware inference. Integrating calibration methods or conformal prediction within latent lexical domains represents a promising direction for future research. It would enable principled confidence estimation while maintaining the efficiency and interpretability that motivate lightweight probabilistic models.

References

[1] Fernando, Rafel, et al. "NLP Implementation For AI Generated Text Detection (ChatGPT) Using Naive Bayes Method." J-INTECH 13.02 (2025): 292-302.

[2] Bhattacharjee, Amrita, et al. "Conda: Contrastive domain adaptation for ai-generated text detection." arXiv preprint arXiv:2309.03992 (2023).

[3] Calderon, Nitay, et al. "Measuring the robustness of NLP models to domain shifts." Findings of the Association for Computational Linguistics: EMNLP 2024. 2024.

[4] Opara, Chidimma. "Distinguishing AI-generated and human-written text through psycholinguistic analysis." International Conference on Artificial Intelligence in Education. Cham: Springer Nature Switzerland, 2025.

[5] Khan, Hikmat Ullah, et al. "Identifying artificial intelligence-generated content using the DistilBERT transformer and NLP techniques." Scientific Reports 15.1 (2025): 20366.

[6] Crothers, E. N., N. Japkowicz, and H. L. Viktor. "Machine-generated text: a comprehensive survey of threat models and detection methods. IEEE Access 11, 70977–71002 (2023)."

[7] Ghosal, Soumya Suvra, et al. "A survey on the possibilities & impossibilities of AI-generated text detection." Transactions on Machine Learning Research (2023)