JAIST Repository
https://dspace.jaist.ac.jp/

Title	Proceedings of the 13th Asean Workshop on Information Science and Technology (AWIST 2025)
Author(s)	Luckyardi, Senny; Mohd, Masnizah; Shirai, Kiyoaki
Citation	
Issue Date	2025-11-13
Type	Conference Paper
Text version	publisher
URL	http://hdl.handle.net/10119/20075
Rights	Copyright (c) 2025 JAIST Press
Description	The 13th ASEAN Workshop on Information Science and Technology 2025 (AWIST 2025), Nomi, Japan, November 6-8, 2025

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

Proceedings of the 13th Asean Workshop on Information Science and Technology (AWIST 2025)

Senny Luckyardi[1], Masnizah Mohd[2], Kiyoaki Shirai[3]

JAIST Press

ISBN: 978-4-903092-67-6

Preface

It is our great pleasure to present the Proceedings of the 13th ASEAN Workshop on Information Science and Technology 2025 (AWIST 2025). This workshop gathers researchers and students from Asia and beyond to exchange ideas, share advances, and discuss challenges in information science.

AWIST has a long and distinguished history. Over the years, it has been hosted in rotation among several Asian countries, serving as a valuable platform for academic collaboration and information exchange among partner institutions. In 2025, the workshop took place at the Japan Advanced Institute of Science and Technology (JAIST) from November 6 to 8. The event concluded successfully with active discussions, fruitful exchanges, and meaningful connections among participants.

The papers included in this volume reflect the diversity and richness of research being conducted by scholars and professionals. Topics include artificial intelligence, cyber security, deep learning, educational system, environment and society, game informatics, health and healthcare, human computer interaction, image processing, Internet of Things, machine learning, natural language processing, software engineering, speech signal processing, and their applications. Together, these contributions illustrate the dynamism and creativity that drive progress in the field of information science and technology.

We would like to express our sincere gratitude to all authors for their valuable submissions, to the reviewers for their insightful comments and dedication, and to the organizing committee members for their hard work in making this event a success. We hope that this proceedings will serve as a useful record of the workshop and an inspiration for future collaborations and innovations in information science and technology.

General Chair

Minoru Terano, Japan Advanced Institute of Science and Technology

Program Chair

Kiyoaki Shirai, Japan Advanced Institute of Science and Technology
Senny Luckyardi, Universitas Komputer Indonesia
Masnizah Mohd, Universiti Kebangsaan Malaysia

Advisory Board Committee

Hiroyuki Iida, Japan Advanced Institute of Science and Technology
Saw Sanda Aye, University of Information Technology

Steering Committee

Shinobu Hasegawa, Japan Advanced Institute of Science and Technology
Muhammad Suzuri Hitam, Universiti Malaysia Terengganu
Chu-Hsuan Hsueh, Japan Advanced Institute of Science and Technology
Kokolo Ikeda, Japan Advanced Institute of Science and Technology
Waree Kongprawechnon, Sirindhorn International Institute of Technology
Yuto Lim, Japan Advanced Institute of Science and Technology
Thepchai Supnithi, Thailand National Science and Technology Development Agency
Wiwied Virgiyanti, Universiti Malaysia Terengganu
Shi-Jim Yen, National Dong Hwa University

Program Committee

Irawan Afrianto, Universitas Komputer Indonesia
Arifah Che Alhadi, Universiti Malaysia Terengganu
Noormadinah Allias, Universiti Teknologi MARA
Punyawee Anunpattana, National University of Singapore
Htun Pa Pa Aung, Rakuten, Inc.
Norizan Mat Diah, Universiti Teknologi MARA
Nur Fazidah Elias, Universiti Kebangsaan Malaysia
Waheed Ali Hussein Mohammed Ghanem, Universiti Malaysia Terengganu
Haryani Haron, Universiti Teknologi MARA
Shahirah Mohamed Hatim, Universiti Teknologi MARA, Perak Branch
Fatin Filzahti Ismail, Universiti Kebangsaan Malaysia
Mohd Nor Akmal Khalid, Universiti Kebangsaan Malaysia
Luiz Bernardo Martins Kummer, Pontifícia Universidade Católica do Paraná
Seksan Laitrakun, Sirindhorn International Institute of Technology
Myint Myint Lwin, University of Information Technology
Mustafa Man, Universiti Malaysia Terengganu
Hanhan Maulana, Universitas Komputer Indonesia
Aung Htein Maw, University of Information Technology
Sharifah Mashita Syed Mohamad, Universiti Malaysia Terengganu
Aye Chan Mon, University of Information Technology

Myat Pwint Phyu, University of Information Technology
Anggina Primanita, Universitas Sriwijaya
Yeffry Handoko Putra, Universitas Komputer Indonesia
Ednawati Rainarli, Universitas Komputer Indonesia
Nazhif Rizani, Independent Researcher
Xiong Shuo, Huazhong University of Science and Technology
Prarinya Siritanawan, Shinshu University
Kristian Spoerer, University of Nottingham
Nur Hanis Sabrina binti Suhaimi, Universiti Kebangsaan Malaysia
Sri Supatmi, Universitas Komputer Indonesia
Sila Temsiririrkkul, Huachiew Chalermprakiet University
Win Win Thant, University of Information Technology
Sagguneswaraan Thavamuni, Rakuten, Inc.
Nwe Nwe Myint Thein, University of Information Technology
Sasiporn Usanavasin, Sirindhorn International Institute of Technology
Pikul Vejjanugraha, Chiang Mai University
Thin Thin Wai, University of Information Technology
Kang Xiaohan, Xi'an International Studies University
Gao Yuexian, Hebei University of Engineering
Wan Nural Jawahir Hj Wan Yussof, Universiti Malaysia Terengganu
Song Zhang, Amazon.com, Inc.
Thet Thet Zin, University of Information Technology
Long Zuo, Chang'an University

Local Organizing Committee

Ryusei Arakawa, Japan Advanced Institute of Science and Technology
Reina Hagiwara, Japan Advanced Institute of Science and Technology
Shion Kitabatake, Japan Advanced Institute of Science and Technology
Kyota Kuboki, Japan Advanced Institute of Science and Technology
Xuan Liu, Japan Advanced Institute of Science and Technology
Thanh Nguyen Canh, Japan Advanced Institute of Science and Technology
Tatsuyoshi Ogawa, Japan Advanced Institute of Science and Technology
Syunsei Takarabe, Japan Advanced Institute of Science and Technology

Editors' affiliations and emails

[1] *Universitas Komputer Indonesia*
[2] *Universiti Kebangsaan Malaysia*
[3] *Japan Advanced Institute of Science and Technology*

senny@email.unikom.ac.id (S. Luckyardi)
masnizah.mohd@ukm.edu.my (M. Mohd)
kshirai@jaist.ac.jp (K. Shirai)

Table of Contents

Beyond Cybersecurity Fatigue: A Framework for Cognitive Load-Aware Policy in Organizations 1
Anderson Kevin Gwenhure and SangGyu Nam

Network-Based Influencer Discovery: A Dual Centrality Approach for Topic-Specific Digital Marketing on Social Media Platforms 13
Adam Mukharil Bachtiar, Dian Dharmayanti, Muhammad Rakha Firdaus and Abdurrazak Syakir Muharam

Classification of Students Continuing Their Studies to University Using Data Mining 25
Wartika, Agus Nursikuwagus, Deasy Permatasari, Novrini Hasti and Zulfikar

Transforming Medical Practice: Super-Intelligence in Health and Healthcare 37
Shiqi Song

A Dual-Stage StyleGAN-ADA Framework for Automated Sketch-to-Zentangle Artistic Transformation 44
Yeffry Handoko Putra and Wan Fariza Abdul Rahman

Implementation of Reinforcement Learning on an Automatic Obstacle Avoiding Agent in an Endless Runner Game 57
Kaka Inochi, Anggina Primanita and Julian Supardi

Synthesizing Monster Voices in Indie Games Using Retrieval-Based Voice Conversion (RVC): A Low-Cost Audio Innovation Approach 69
Hafiz Muhammad Kurniawan and Anggina Primanita

Interpretable Machine Learning for Assessing Winter Outdoor Thermal Comfort: Exploring Built Environment Impacts in a Historic Waterfront Street 81
Haitao Lian, Zhenghui Han, Zeyu Ma and Yulin Yang

EduGame: Making Math Magical with 3D Learning Adventures 94
Shahirah Mohamed Hatim, Haryani Haron and Muhammad Daniel Aiman Mohd Rozli

Redefining Campus Orientation: A Web-Based Virtual Tour Experience for Students 105
Haryani Haron, Shahirah Mohamed Hatim and Allysha Zull Hizam

Classification Hate Speech in Boosting Algorithms for Trending Twitter cases Domestic Violence in Indonesia 118
Agus Nursikuwagus and Syahrul Mauluddin

Research on the Development Path of Industrial Integration in Health and Wellness Tourism 131
Zhai Juntian

A Human-Centric Decision Support Framework for Satisfaction Enhancement in Medical Staff Scheduling 142
Pavinee Rerkjirattikal, Raveekiat Singhaphandu, Charnon Pattiyanon and SangGyu Nam

AI-Driven Hybrid Intelligence for Skincare Personalization: A Multimodal Analysis of Skin Type Segmentation and Consumer Preferences in Indonesia 154
Sri Supatmi, Mia Fitriawati, Yusrilla Y Kerlooza, Rongtang Hou and Shuonan Hou

Simplified Maximally Stable Extremal Region for Text Detection in Natural Image 167
Ednawati Rainarli, Suprapto and Wahyono

Spatial characteristics of outdoor pedestrian commercial streets in winter vitality in severe cold areas-considering sensory comfort factor 180
Haitao Lian, Zhenghui Han, Zeyu Ma and Yulin Yang

The Impact of Commercial Street Billboards on Pedestrian Vitality: An Empirical Study of Wanlimiao, Shijiazhuang, China 193
Haitao Lian, Zeyu Ma, Zhenghui Han and Yulin Yang

Vulnerable Grid: Strategies and Challenges in Training "Disaster-Resilient Engineers" 207
Haoyu Liu and Ying Cheng

Teaching Civil Engineering Materials in the Era of AI and Carbon Neutrality 215
Ying Cheng, Yuyao Li, Haoyu Liu and Ruijie Shi

Psychotherapy as A Near-Deterministic Event: Estimating Reward Frequency (N) Across AI and Human Platforms via Motion-in-Mind 223
Lulu Gao, Shize Pan, Muhammad Numan, Mohd Nor Akmal Khalid and Hiroyuki Iida

A Controlled Framework for Generating Synthetic, Multi-Topic Thai Conversations for Healthcare Contact Centers 236
Kasidis Manasurangkul, Charnon Pattiyanon, Niracha Janavatana, Supakorn Etitum, Pon Yimcharoen and Shine Min Kha

Breaking Free from the GPA Assembly Line Rebuilding Individual Growth Paths in the Era of Superintelligence 249
Zhang Yuxuan, Niu Jiaxing and Shang Jinghan

Quantitative Optimization of Gamified Museum Experiences: A Motion in Mind Approach 262
Jiayu Liu, Jiahao Zhang, and Yuexian Gao

Spoof Detection in Automatic Speaker Verification Using ResNet-34 and Early-Stage Cepstral Coefficient Fusion 272
Kosin Kalarat, Sasiporn Usanavasin, Thanaruk Theeramunkong, Kasorn Galajit and Jessada Karnjana

From Digital Transformation to Human–AI Symbiosis: The Evolving Role of Universities in Shaping Education and Society 282
Xiaokun Shi and Jizong Jia

Study on Periocular Area Verification with Convolutional Autoencoder and Principal Component Analysis for Biometric Authentication 290
Chuesing Ni, Waree Kongprawechnon and Jessada Karnjana

Emotion Recognition from Indoor Scene Imagery: A CNN Approach Using RGB Features and PAD Dimensions 299
Lei Tong and Mohd Nor Akmal Khalid

Metaphor-Aware Sentiment Analysis in Multi-Turn Conversations 311
Guo Wei and Mohd Nor Akmal Khalid

A Simplified Multi-Floor Classification-Based Indoor Positioning System Study 324
Burin Intachuen, Mhadhanagul Charoenphon, Tanakorn Mankhetwit and Charnon Pattiyanon

Innovation of Business Strategy Framework and Artificial Intelligence-Based Accounting Information System on Cooperative Digitalization Performance 336
Supriyati, Andrias Darmayadi, Dian Dharmayanti and Ramadhan Syaeful Bahri

Fully Homomorphic Encryption for Secure and Confidential Text Classification 350
Zuraiha Ambri and Shakirah Hashim

Human-computer interaction and interface design——Matlab App diffraction simulation ... 363
Xinlong Wang and Huanzhen Zhang

AI Assisted Grading Framework for Thai-Language Written Exam Questions based on LLM and Rule-Based Reasoning Approach 376
Chutipon Trirattnanurak, Sasiporn Usanavasin, Chaianun Damrongrat and Manubu Okumura

Indoor Air Quality Monitoring System Based on IoT 389
Hidayat and Iswan Samin

Uncovering the Potential of IoT in Pogostemon helferi Cultivation: A Comparative Study of IoT-Based Cultivation Systems and Conventional Emersed Methods 400
Hanhan Maulana, Achmad Juliarman, Hideaki Kanai, Sunny Goh Eng Giap and Roslaili Abdul Azis

Advancing a Human-Centered Theory of Software Reliability: Cognitive-Emotional Adaptive Systems for Sustainable Human Development 412
Sharifah Mashita Syed-Mohamad, Norma Alias, Ruwaidiah Idris and Norsyazwani M. Subri

K-Degree Anonymity for Social Network Privacy: Balancing Identity Protection and Structural Utility 426
Pham Minh Thanh

Research on the Application of Chatbot Teaching Assistants in University Teaching in the Era of Superintelligence: A Case Study of Jill Watson 439
Ma Huimin

Infestation to Prevention: Smart IoT Pest Detection 450
Wiwied Virgiyanti, Mohd Kamir Yusof, Mustafa Man and Wan Aezwani Wan Abu Bakar

Digitalized Farming and Excellent Agricultural Management; A Roadmap toward Food Security 462
Senny Luckyardi, Eddy Soeryanto Soegoto, Lia Warlina, Dian Dharmayanti and Muhamad Fahrezi

A Hierarchical Fuzzy Controller with Upstream–Downstream Awareness for Emission Reduction in Urban Intersection Networks 474
Muhammad Aria Rajasa Pohan, Jana Utama and Budi Herdiana

Modeling the Engagement Dynamics of Musical Melody with Motion in Mind and Game Refinement Theory 487
Jiahao Zhang, Jiayu Liu and Yuexian Gao

Beyond Cybersecurity Fatigue: A Framework for Cognitive Load-Aware Policy in Organizations

Anderson Kevin Gwenhure [1][0009-0005-7427-9635] and SangGyu Nam[2*][0000-0002-7424-8469]

[1,2] Sirindhorn International Institute of Technology, Thammasat University, Pathum Thani, 12120, Thailand
d6722300115@g.siit.tu.ac.th, sanggyu@siit.tu.ac.th

Abstract. Cybersecurity fatigue, defined as resistance or disengagement from security behaviors due to excessive or poorly structured demands, undermines the effectiveness of organizational awareness initiatives. Existing frameworks, while offering structured guidance, cultural engagement, and governance alignment, do not explicitly address the cognitive overload that drives fatigue or provide concrete mechanisms for prioritizing awareness content. To address this gap, this paper proposes the CLARA (Cognitive Load Aligned Risk Awareness) Framework, a structured, stakeholder-informed approach grounded in cognitive load theory. CLARA introduces a Delphi-based risk prioritization process that identifies misunderstood and high-impact policies, operationalizing prioritization to ensure awareness efforts are streamlined and aligned with users' cognitive capacity, thereby mitigating fatigue. Comparative analysis shows that CLARA complements established approaches such as NIST SP 800-53 Rev. 5, the SANS Security Awareness Maturity Model, and COBIT 2019 by providing a practical method to operationalize risk-based prioritization, thereby avoiding cognitive overload and ensuring information processing remains manageable. However, CLARA remains a conceptual framework, and its reliance on stakeholder judgment may introduce bias, while implementation can be challenging in resource-limited contexts. Future research should validate CLARA using Kitchenham's framework evaluation criteria and extend its scope to include non-communication drivers of fatigue, such as organizational culture and workload pressures.

Keywords: Cybersecurity policy, Cybersecurity fatigue, Risk-based prioritization

1 Introduction

Ordinary users have long posed a threat to information technology security [1]. Consequently, it is widely recognized in cybersecurity discourse that users represent the weakest link in the security chain [2, 3]. A key reason for this vulnerability is inadequate security awareness and non-compliance with organizational security policies [2]. These policies, which include cybersecurity policies as a subset [4, 5], lose effectiveness without sufficient awareness and adherence, exposing organizations to risks that technical solutions alone cannot resolve [6].

Reducing these risks requires not only technological safeguards but also structured organizational processes that foster a culture where cybersecurity is embedded in everyday practice [1]. A central component of this culture is the implementation of cybersecurity policies, established to safeguard digital assets and sensitive information [7]. These policies define procedures and responsibilities aimed at ensuring the confidentiality, integrity, and availability of organizational resources [8]. When effectively communicated, they provide employees with clear instructions on best practices and ensure they understand their responsibilities in protecting the organization's digital assets [7], thereby supporting awareness and encouraging informed, responsible behavior.

However, policy creation and existence alone do not guarantee employee awareness or compliance [9]. Compliance remains a persistent challenge due to factors such as policy complexity, insufficient clarity, and employee resistance [8]. Even well-documented policies are often difficult for employees to follow in practice [10]. This often leads to unintentional noncompliance, as employees may rationalize or downplay the significance of their actions during routine tasks [10], while remaining unaware of the risks and consequences associated with policy violations [11]. Such behavior highlights the enduring vulnerability introduced by the human element when security policies are not fully understood, accepted, or internalized [12]. Even in environments where policy communication is regular, it may not be sufficient [13]. Many users report frustration or disengagement in response to the volume or frequency of security-related messaging [14], often resulting in avoidance or intentional workarounds [15].

Beyond active disregard, these patterns reflect cybersecurity fatigue, a condition characterized by decreased motivation or avoidance of security tasks due to overexposure and limited cognitive or emotional resources [16]. This condition suggests cybersecurity awareness strategies may impose excessive cognitive demands. According to Cognitive Load Theory [17], excessive demands can surpass individuals' working memory limits, thereby impairing learning and behavioral adherence. This implies that if awareness initiatives exceed cognitive limits, employees are unlikely to retain or apply policy knowledge consistently. Therefore, unless security policy awareness is cognitively manageable, comprehension and sustained policy compliance are unlikely.

In light of the complexity and consequences associated with cybersecurity fatigue, organizational leaders must implement focused measures that address its root causes [18]. This entails reducing unnecessary cognitive strain caused by too much information, overcomplexity, dense explanations, poor design, and monotonous repetition [16, 19–21], while simultaneously fostering purposeful mental engagement that encourages meaningful learning and connection to prior knowledge [20].

Building on this foundation, mitigating excessive cognitive load is essential to reducing cybersecurity fatigue and fostering stronger policy engagement and compliance. Accordingly, this study suggests a cognitively sustainable approach to cybersecurity policy awareness that streamlines communication and promotes focused mental engagement. This is particularly important given that awareness significantly shapes individuals' attitudes and expectations toward compliance [22]. This research contributes to human-centric cybersecurity by proposing a structured, risk-informed, fatigue-conscious framework that aligns awareness initiatives with cognitive capacity, with the aim of reducing strain, supporting understanding, and reinforcing secure behavior.

2 Theoretical Foundation

This section outlines the theoretical basis of the proposed framework.

2.1 Cognitive Load Theory

The theory explains that individuals can process only a limited amount of new information at a time. This limited capacity of working memory, where new in-formation is actively processed, implies that overly complex, highly detailed, or poorly organized content can hinder comprehension and retention [17, 23, 24]. It emphasizes that learning is most effective when unnecessary mental effort is minimized, enabling individuals to focus on essential information [24]. Poor presentation, such as confusing wording, lengthy explanations, or excessive and irrelevant details, adds 'extraneous load,' which distracts from the core message and impairs comprehension [23, 24]. Conversely, clear and concise information presentation directs attention to what truly matters, thereby enhancing understanding and memory retention [25].

2.2 Cybersecurity Fatigue

Fatigue is a state of mental or physical exhaustion that undermines decision-making and performance [21]. In cybersecurity context, it often manifests as weariness or resistance toward security-related behaviors, typically triggered by overexposure to demands such as excessive messaging, frequent training, or unclear instructions [16].

A prominent form is advice-related fatigue, which arises when employees are repeatedly exposed to instructions through Security Education Training and Awareness (SETA) programs [26]. Its drivers include cognitive overload, where users receive more information than they can effectively process [21]; prolonged exposure to complex or inconsistent messages without adequate support, which fosters disengagement [16]; and inconsistent presentation of information, particularly in low-literacy environments where users are expected to make security decisions without a foundational understanding [19].

Addressing fatigue therefore requires aligning awareness initiatives with users' cognitive capacities [27]. Messages should be simplified, delivered at moderate frequency, and synchronized with workload demands [21, 26, 28]. Ultimately, efforts need to streamline security requirements and direct attention toward high-priority risks [21].

A handful of scholars have proposed valuable responses to fatigue. [16] introduce a four-component model for tailoring interventions; [21] recommend strategies such as digital detox and mental health support; [28] emphasize moderation in training and awareness; and [18] present the study CyFa tool to address attribution biases in interpreting fatigue. These contributions highlight the significance of the issue and offer useful approaches, yet they primarily emphasize managing fatigue once it arises.

In contrast, this study contributes by proposing a proactive, cognitively sustainable framework. Grounded in cognitive load theory, it is intended not to treat fatigue after the fact but to help prevent the onset of information overload, thereby supporting more sustainable engagement with cybersecurity policy awareness.

3 Conceptual Framework

This section introduces the proposed framework, outlining its rationale and structure.

3.1 Rationale for the Framework

Theoretical insights highlight the importance of structured, proactive mechanisms that help reduce unnecessary cognitive demands in cybersecurity policy awareness initiatives. These policies are vital for safeguarding organizational assets, yet uniform communication or enforcement can sometimes overwhelm employees and contribute to fatigue [28]. Prior research has established the connection between cognitive overload and security fatigue [16, 21, 26, 28] and has underscored the importance of reducing strain. However, many professionals and decision-makers remain uncertain about how to operationalize these insights into targeted strategies, indicating the need for clearer, practice-oriented guidance [18]. This challenge is particularly acute in high-security environments where fatigue tends to be most pronounced [21]. Without well-informed approaches, there is a risk that fatigue could compound human-related vulnerabilities [18].

Because cognitive overload underpins fatigue, mitigating unnecessary cognitive strain is a critical step. This can be supported by emphasizing essential content and reducing unnecessary complexity [29]. In practice, streamlining security requirements and focusing user attention on high-priority risks can help reduce extraneous cognitive burden [21].

Building on these insights, this paper introduces the CLARA (Cognitive Load-Aligned Risk Awareness) Framework as a structured, stakeholder-driven approach proposed to support the implementation of more focused awareness strategies. CLARA is intended to complement existing awareness approaches by streamlining policy communication, identifying policies that are least understood or often neglected, and prioritizing them based on organizational risk and potential impact. This targeted prioritization aims to reduce cognitive strain from excessive information and to strengthen employees' ability to retain policy content.

Grounded in cognitive load theory, CLARA rests on the principle that, while all policies matter, it is cognitively unrealistic to expect employees to maintain equal awareness of every policy simultaneously. Instead, the framework promotes the strategic selection of policy topics that are both misunderstood or neglected and critical to current risks, with the goal of keeping awareness efforts contextually relevant, sustainable, and cognitively manageable.

3.2 Structure and Components of the CLARA Framework

The CLARA Framework is structured into a three-phase Input–Process–Output (IPO) model (Figure 1). Each phase incorporates activities intended to identify misunderstood or neglected cybersecurity policies and to prioritize them for targeted awareness. The overarching aim is to conceptually address cybersecurity fatigue by focusing on its root causes, particularly cognitive overload, through more focused, risk-informed, and timely interventions.

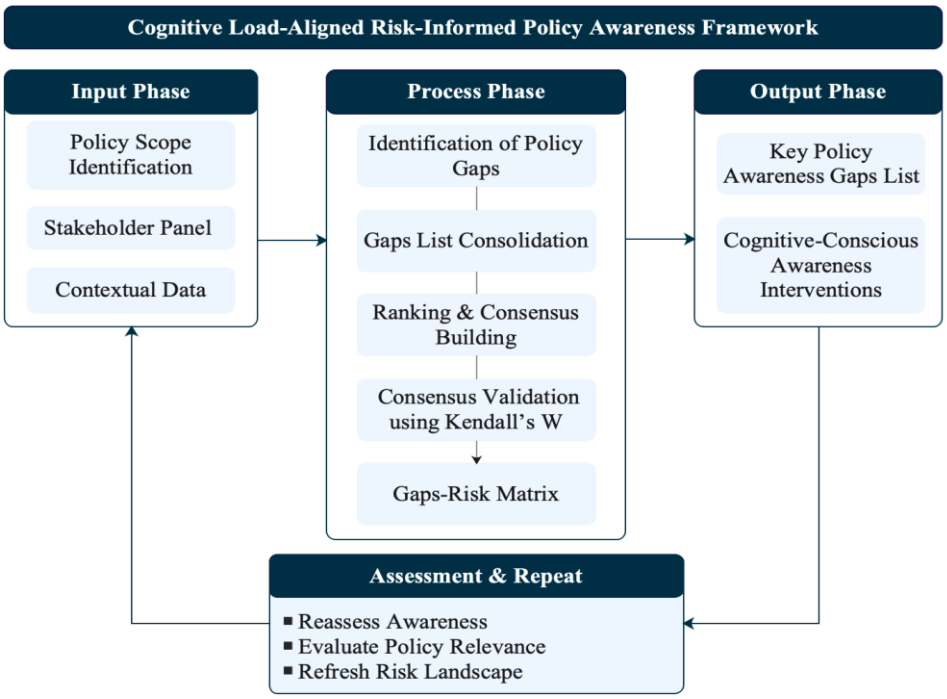


Fig. 1. The CLARA Framework

A. Input Phase: Establishing Scope and Context

This phase defines the foundational parameters for applying CLARA within an organizational setting.

i. Policy Scope Identification

The complete range of internal cybersecurity policies is catalogued and categorized, providing a baseline inventory for subsequent risk evaluation.

ii. Stakeholder Panel

A diverse panel of internal stakeholders is assembled, typically including cybersecurity officers, IT personnel, policy enforcers, human resource managers, risk managers, and end-user representatives. This diversity ensures that evaluations of policy understanding, and implementation reflect multiple organizational perspectives.

iii. Contextual Data

Supporting evidence such as policy violations, audit findings, and user feedback is shared with stakeholders. While not directly analyzed, these cues help ground stakeholder judgments in recent organizational experience, reducing reliance on memory alone.

B. Process Phase: Delphi-Based Prioritization

The Process Phase forms the analytical core of CLARA, applying the Delphi technique to conduct risk-based prioritization and achieve consensus. The Delphi method is a structured, iterative communication process designed to facilitate systematic examination of complex issues without pressuring participants into premature compromise [30, 31]. Its advantages include producing collective perspectives that would not emerge from unstructured discussion [31], while anonymity shields participants from hierarchical influence and a designated facilitator ensures consistent communication and synthesis across rounds [30]. Within CLARA, Delphi is employed to elicit independent stakeholder input across successive rounds, progressively refining judgments into a coherent prioritization of policy awareness gaps.

i. Identification of Policy Gaps (Delphi Round 1)

Stakeholders independently identify policy areas they perceive as most misunderstood, frequently ignored, or inconsistently applied. Independent submissions preserve diversity of judgment and mitigate group influence.

ii. Gaps List Consolidation

Responses are consolidated into a single list. Overlapping items are merged and ambiguous entries clarified, producing a coherent set of distinct policy gaps for further evaluation.

iii. Ranking and Consensus Building (Delphi Rounds 2 and 3)

Stakeholders evaluate each consolidated item on two dimensions:

- Degree of misunderstanding or neglect
- Organizational risk if misunderstood

These evaluations are conducted across two Delphi rounds. After the first scoring, aggregated results are shared with participants, who may reflect on group trends and adjust their responses. This iterative process refines judgments and builds stronger consensus.

iv. Consensus Validation using Kendall's W

Stakeholder agreement is tested using Kendall's Coefficient of Concordance (W) [32]. A W value of 0.70 or higher indicates strong consensus. Where consensus is weak, results are fed back to participants, and an additional round may be conducted. This ensures that prioritization is both stakeholder-informed and statistically validated.

v. Gaps-Risk Matrix

Policies are mapped against their assessed risk impact and level of misunderstanding or neglect. This generates a ranked list of policy areas representing the most urgent awareness gaps.

C. Output Phase: Cognitive-Conscious Awareness Interventions

The Output Phase translates validated insights into targeted awareness actions.

i. Key Policy Awareness Gaps List

Policy enforcers/security managers receive a prioritized list of high-risk, low-understanding policy areas to guide awareness efforts.

ii. Cognitive-Conscious Awareness Interventions

Awareness efforts are designed with the following features:

- Targeted: Focused on the most critical, misunderstood policies
- Moderated: Delivered at intervals that avoid overload
- Clear and Concise: Structured to minimize extraneous cognitive load
- Contextualized: Aligned with specific tasks, roles, and risk levels

These characteristics can ensure that awareness strategies remain relevant, cognitively aligned, and sustainable, reducing fatigue while strengthening knowledge retention.

D. Assess and Repeat: Iterative Refinement

CLARA operates as a cyclical process. Following implementation, organizations periodically reassess user understanding, policy relevance, and emerging risks. These findings inform the next cycle, ensuring that awareness initiatives remain adaptive, risk-aligned, and cognitively sustainable.

4 Comparative Analysis and Framework Contribution

This section compares CLARA with mainstream security awareness frameworks and highlights its contribution.

4.1 Comparative Analysis CLARA and Mainstream Awareness Approaches

For the sake of fairness, and to ensure direct applicability to the context under study, the comparison was grounded in factors consistently highlighted in literature as essential for addressing cybersecurity fatigue, rather than on arbitrary benchmarks. To reiterate, prior research identifies two core expectations to address fatigue: first, reducing cognitive overload through simplified, consistent, and moderately paced communication aligned with users' cognitive capacities [21, 27, 28]; and second, directing attention toward high-priority risks by streamlining security requirements and focusing user effort where it matters most [21]. Based on these expectations, the following literature-derived comparative evaluation criteria were applied in the analysis: (i) Cognitive Overload Reduction, the extent to which a framework minimizes unnecessary strain by simplifying, clarifying, and pacing awareness content in alignment with user capacity; and (ii) Risk-Based Prioritization, the degree to which a framework streamlines awareness by concentrating user attention on high-priority policies and risks.

As shown in Table 1, mainstream awareness approaches provide valuable structure, engagement strategies, and governance–risk alignment, but they do not explicitly address cognitive overload or cybersecurity fatigue, nor do they offer structured mechanisms for prioritizing awareness content on the basis of risk. CLARA is proposed to advance these gaps by emphasizing cognitive load and fatigue reduction in cybersecurity policy awareness through stakeholder-informed, risk-based prioritization.

Table 1. Comparison of CLARA and Mainstream Awareness Approaches

Approaches	Cognitive-Overload Reduction	Risk-Based Prioritization
National Institute of Standards and Technology (NIST) SP 800-53 Rev. 5. [33]	Emphasizes structured lifecycle guidance (plan, design, implement, evaluate) and role-based tailoring, but does not explicitly address cognitive overload as a factor in avoiding fatigue.	Emphasizes risk-based awareness prioritization but does not provide concrete mechanisms for carrying it out.
SysAdmin, Audit, Network, and Security (SANS) Institute Security Awareness Maturity [34]	Emphasizes awareness engagement and cultural maturity through varied methods but does not account for the possibility that such engagement activities may contribute to cognitive load and fatigue if not streamlined.	Does not provide a structured mechanism for risk-based prioritization of awareness.
Control Objectives for Information and Related Technology (COBIT) 2019 Framework: Governance and Management Objectives [35]	Emphasizes security awareness as a governance responsibility, ensuring policies, risks, and expectations are communicated and embedded across the organization, but does not account for cognitive overload or fatigue.	Emphasizes aligning awareness with organizational risks but does not provide concrete mechanisms for prioritization.
CLARA*	Emphasizes cognitive load reduction by ensuring awareness interventions are simplified, consistent, and delivered at a moderate frequency to avoid fatigue	Emphasizes stakeholder-informed prioritization by introducing a Delphi process that ranks policies based on misunderstanding and organizational risk, producing a clear, defensible basis for targeted awareness efforts.

4.2 CLARA Contribution

CLARA contributes to cybersecurity awareness research and practice across three dimensions. Theoretically, it is proposed to extend cognitive load theory into the domain of policy awareness by framing fatigue as a preventable outcome of poorly structured communication rather than an inevitable by-product of awareness initiatives. Methodologically, it outlines a stakeholder-informed Delphi process for identifying misunderstood or neglected policies and prioritizing them according to organizational risk, offering a transparent and repeatable mechanism to guide awareness focus. Practically, CLARA is intended to complement rather than replace established frameworks/approaches by addressing the overlooked challenge of cognitive overload and fatigue, thereby supporting the effectiveness of existing approaches under real-world constraints of limited user capacity.

NIST awareness guidance emphasizes structured lifecycle guidance (plan, design, implement, evaluate) and role-based tailoring [33]. CLARA complements these stages by offering a mechanism to determine which policies should be tailored and how sus-

tainability can be achieved without overburdening employees, thereby adding a cognitive dimension to NIST's structured design. SANS, through its Security Awareness Maturity Model, highlights the importance of engagement in progressing from compliance to culture change [34]. CLARA enhances this model by ensuring that engagement activities such as gamification, phishing simulations, or repeated campaigns are risk informed and streamlined to remain effective without contributing to fatigue, enabling organizations to mature sustainably. COBIT positions awareness as a governance responsibility tied to business objectives, risk appetite, and audit findings [35]. CLARA operationalizes this linkage by providing a stakeholder-informed, risk-based prioritization process that explicitly accounts for cognitive capacity and fatigue, ensuring that initiatives remain both cognitively sustainable.

5 Conclusion, Limitations and Future Work

The CLARA (Cognitive Load-Aligned Risk Awareness) Framework is presented as a structured, stakeholder-informed, risk-based approach intended to mitigate cybersecurity fatigue by aligning awareness with cognitive load capacity. By prioritizing communication according to policy misunderstanding and risk exposure, it shifts focus from broad, repetitive messaging toward targeted, cognitively sustainable, and risk-aligned interventions. However, CLARA remains a conceptual framework that requires empirical validation through case studies, simulations, or pilot implementations across diverse organizational contexts. Its reliance on stakeholder judgment may introduce bias, and while it addresses communication-driven fatigue, it does not fully account for non-communication factors such as organizational culture, workload pressures, or technical complexity. Implementation may also be challenging in resource-limited organizations without access to diverse stakeholder panels; in such cases, fewer Delphi rounds and smaller panels may be more feasible. Future research should therefore extend the framework to incorporate non-communication drivers of fatigue and evaluate its effectiveness based on evaluation of framework criteria's by [36].

Acknowledgments. The first author expresses sincere gratitude to the Sirindhorn International Institute of Technology (SIIT) and Thammasat University (TU) for the support provided through the Excellent Foreign Students (EFS-A) Scholarship and the TU PhD Scholarship. The research was additionally funded by the SIIT Young Research Grant, Contract No. SIIT 2022-YRG-SN02.

Disclosure of Interests. The authors declare no conflict of interest.

References

1. Fielding, J.: The people problem: how cyber security's weakest link can become a formidable asset. Computer Fraud & Security. 2020, 6–9 (2020). https://doi.org/https://doi.org/10.1016/S1361-3723(20)30006-3.
2. Qin, Y., Yang, X., Yang, L.X., Huang, K.: Mitigating Social Engineering Attacks Through Cost-Effective Security Awareness Training Policy. IEEE Trans Netw Sci Eng. (2025). https://doi.org/https://10.1109/TNSE.2025.3556927.

3. Tambe-Jagtap, S.N.: Human-Centric Cybersecurity: Understanding and Mitigating the Role of Human Error in Cyber Incidents. SHIFRA. 2023, 53–59 (2023). https://doi.org/https://10.70470/SHIFRA/2023/007.
4. Sabillon, R., Serra-Ruiz, J., Cavaller, V., Cano, J.J.M.: An effective cybersecurity training model to support an organizational awareness program: The Cybersecurity Awareness Training Model (CATRAM). A case study in Canada. Journal of Cases on Information Technology. 21, 26–39 (2019). https://doi.org/https://10.4018/JCIT.2019070102.
5. Bruno, E., Pistolesi, F., Teti, E.: Cybersecurity policy, ESG and operational risk: A Virtuous relationship to improve banks' performance. International Review of Economics and Finance. 99, (2025). https://doi.org/https://10.1016/j.iref.2025.104053.
6. Koohang, A., Anderson, J., Nord, J.H., Paliszkiewicz, J.: Building an awareness-centered information security policy compliance model. Industrial Management and Data Systems. 120, 231–247 (2020). https://doi.org/10.1108/IMDS-07-2019-0412.
7. What Is A Cyber Security Policy? Importance And Best Practices, https://www.metacompliance.com/blog/data-breaches/what-is-a-cyber-security-policy, last accessed 2025/06/16.
8. Cronje, J.C., Okigui, H., Francke, E.R.: An Analysis of Cybersecurity Policy Compliance in Organisations. Applied Cybersecurity & Internet Governance. (2024). https://doi.org/https://10.60097/acig/191942.
9. Nord, J.H., Koohang, A., Floyd, K., Paliszkiewicz, J.: Impact of Habits on Information Security Policy Compliance. Issues in Information Systems. 21, 217–226 (2020). https://doi.org/10.48009/3_iis_2020_217-226.
10. Bulgurcu, B., Cavusoglu, H., Benbasat, I.: Information Security Policy Compliance: An Empirical Study of Rationality-Based Beliefs and Information Security Awareness. MIS Quarterly. 34, 39 (2010). https://doi.org/http://dx.doi.org/10.2307/25750690.
11. Assefa, T., Tensaye, A.: Factors influencing information security compliance: an institutional perspective. SINET: Ethiopian Journal of Science. 44, 108–118 (2021). https://doi.org/10.4314/sinet.v44i1.10.
12. Ibrahim Almuwail, K., Saad Albarrak, A., Nasir Mumtaz Bhutta, M., M Wahsheh, H.A.: Examining the Factors for Non-Compliance of Saudi Health Organizations for E-Health Security and Privacy. J Theor Appl Inf Technol. 31, (2023).
13. Olt, C.M.: On Security Guidelines and Policy Compliance: Considering Users' Need for Autonomy. In: ICIS 2021 Proceedings. 2 (2021).
14. Michael Olt, C., Mesbah, N.: Weary of Watching Out? – Cause and Effect of Security Fatigue. In: In Proceedings of the 27th European Conference on Information Systems (ECIS). , Stockholm & Uppsala (2019).
15. Cram, W.A., Proudfoot, J.G., D'Arcy, J.: When enough is enough: Investigating the antecedents and consequences of information security fatigue. Information Systems Journal. 31, 521–549 (2021). https://doi.org/10.1111/isj.12319.
16. Reeves, A., Delfabbro, P., Calic, D.: Encouraging Employee Engagement With Cybersecurity: How to Tackle Cyber Fatigue. Sage Open. 11, (2021). https://doi.org/10.1177/21582440211000049.
17. Sweller, J.: Cognitive Load During Problem Solving: Effects on Learning. Cogn Sci. 12, 257–285 (1988). https://doi.org/10.1207/s15516709cog1202_4.

18. Reeves, A., Calic, D., Delfabbro, P.: How to De-CyFa the actor-observer bias in cybersecurity fatigue: Building the CyFa measure of attribution styles and mitigation strategies. Comput Secur. 150, (2025). https://doi.org/10.1016/j.cose.2024.104179.
19. Furnell Steven, Collins Emily: Cyber security: what are we talking about? Computer Fraud & Security,. 2021, 6–11 (2021). https://doi.org/https://doi.org/10.1016/S1361-3723(21)00073-7.
20. Klepsch, M., Schmitz, F., Seufert, T.: Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. Front Psychol. 8, (2017). https://doi.org/10.3389/fpsyg.2017.01997.
21. Mizrak, F., Demirel, H.G., Yaşar, O., Karakaya, T.: Digital detox: exploring the impact of cybersecurity fatigue on employee productivity and mental health. Discover Mental Health. 5, (2025). https://doi.org/10.1007/s44192-025-00149-x.
22. Wong, L.W., Lee, V.H., Tan, G.W.H., Ooi, K.B., Sohal, A.: The role of cybersecurity and policy awareness in shifting employee compliance attitudes: Building supply chain capabilities. Int J Inf Manage. 66, (2022). https://doi.org/https://10.1016/j.ijinfomgt.2022.102520.
23. Mélan, C., Cascino, N.: A multidisciplinary approach of workload assessment in real-job situations: Investigation in the field of aerospace activities. Front Psychol. 5, (2014). https://doi.org/10.3389/fpsyg.2014.00964.
24. Ayres, P., Paas, F.: Cognitive load theory: New directions and challenges. Appl Cogn Psychol. 26, 827–832 (2012). https://doi.org/10.1002/acp.2882.
25. Chen, O., Castro-Alonso, J.C., Paas, F., Sweller, J.: Undesirable difficulty effects in the learning of high-element interactivity materials, (2018). https://doi.org/10.3389/fpsyg.2018.01483.
26. Reeves, A., Calic, D., Delfabbro, P.: "Generic and unusable"1: Understanding employee perceptions of cybersecurity training and measuring advice fatigue. Comput Secur. 128, (2023). https://doi.org/10.1016/j.cose.2023.103137.
27. Albalawi, T., Ghazinour, K., Melton, A.: Security Mental Model: Cognitive Map Approach. In: Proceedings - 2017 International Conference on Computational Science and Computational Intelligence, CSCI 2017. pp. 74–79. Institute of Electrical and Electronics Engineers Inc. (2018). https://doi.org/10.1109/CSCI.2017.12.
28. Mangundu, J., Mayayise, T.: The impact of technostress creators on academics' cybersecurity fatigue in South Africa. Issues in Information Systems. 24, 294–310 (2023). https://doi.org/10.48009/4_iis_2023_123.
29. Asma, H., Dallel, S.: Cognitive Load Theory and its Relation to Instructional Design: Perspectives of Some Algerian University Teachers of English. Arab World English Journal. 11, 110–127 (2020). https://doi.org/10.24093/awej/vol11no4.8.
30. Linstone, H.A., Turoff, M., Helmer, O.: The Delphi Method Techniques and Applications. Addison Wesley (1975).
31. Adler, M., Ziglio, E.: Gazing into the Oracle: The Delphi method and its application to social policy and public health. Jessica Kingsley Publishers, Londin, United Kingdom (1996).
32. Dickinson, G.J.: Kendall's Coefficient of Concordance. Nonparametric Measures of Association. 30–48 (1993). https://doi.org/10.4135/9781412985291.

33. Force, J.T.: Security and Privacy Controls for Information Systems and Organizations. (2020). https://doi.org/10.6028/NIST.SP.800-53R5.
34. Security Awareness Maturity Model, https://www.sans.org/white-papers/security-awareness-maturity-model, last accessed 2025/09/29.
35. Cobit 2019 Framework Governance And Management Objectives, https://pdf-up.com/download/cobit-2019-framework-governance-and-management-objectives-4973089, last accessed 2025/09/29.
36. Kitchenham, B., Linkman, S., Linkman, S.: Experiences of using an evaluation framework. Inf Softw Technol. 47, 761–774 (2005). https://doi.org/10.1016/j.infsof.2005.01.001.

Network-Based Influencer Discovery: A Dual Centrality Approach for Topic-Specific Digital Marketing on Social Media Platforms

Adam Mukharil Bachtiar[1,2][0000−0002−5646−9137], Dian Dharmayanti[1][0009−0000−2075−0050], Muhammad Rakha Firdaus[1], and Abdurrazak Syakir Muharam[2][0009−0004−4808−4407]

[1] Universitas Komputer Indonesia, Jalan Dipatiukur No.112-116, Kota Bandung, Jawa Barat 40132, Indonesia adam@email.unikom.ac.id
[2] Japan Advanced Institute of Science and Technology, 1-8 Asahidai, Nomi, Ishikawa, 923-1211, Japan
https://unikom.ac.id

Abstract. Traditional influencer identification limits the ability of businesses to find authentic brand advocates in specific market niches. This study presents a novel computational framework leveraging Social Network Analysis to identify influential users through network positioning rather than popularity metrics. We combined the betweenness and eigenvector centrality measures to detect "hidden influencers" with high engagement potential despite moderate follower counts. Our methodology incorporated a data-driven threshold, selecting the top 20% of users by follower count to establish baseline influencer status. Analysis of Indonesian Twitter posts revealed accounts demonstrating influence amplification patterns, achieving high influence scores through strategic connections with central nodes rather than direct follower accumulation. Temporal validation confirmed 80% of identified influencers maintained relevance over time. Our research demonstrates three key contributions: robust influence scoring systems outperforming follower-based approaches, revealing "influence amplification" where strategically positioned users achieve disproportionate impact, and providing evidence that community network structures correlate with identification accuracy.

Keywords: Social Network Analysis · Influencer Marketing · Centrality Measures · Digital Marketing Analytics · Network-Based Recommendation Systems · Social Media Mining.

1 Introduction

Digital marketing has evolved significantly over the past decade, particularly with the emergence of social media as a primary channel for brand engagement, information distribution, and customer acquisition[5, 24, 2, 20]. It is evident that social media platforms facilitate real-time interactions between companies and consumers. Furthermore, these platforms also create public spaces that accelerate

the spread of messages, both through natural conversations and organized campaigns [13, 1, 10]. These changes demand more advanced analytical approaches to understand how influence is formed and spread within digital networks.

In the context of digital marketing, influencer marketing has gained significant importance as a strategy [22]. This approach involves leveraging the influence of individuals with substantial social network connections to expand the reach of marketing messages and enhance brand credibility[12]. Influencers have the capacity to influence public opinion, impact consumer decisions, and generate a ripple effect through their audience's participation[19]. Nevertheless, the practice of evaluating influence within the field remains heavily reliant on simple metrics such as the number of followers, which frequently fall short in accurately capturing the true influence within the network structure [21, 26].

Reliance on follower count has significant limitations. While accounts with millions of followers can provide extensive reach, recent research shows that engagement rates are often higher on accounts with smaller, more engaged follower bases [8]. This frequently results in companies employing ineffective strategies, thereby missing out on the potential of more authentic and relevant micro-influencers. In order to address this issue, recent research has begun to emphasize the importance of measuring content quality and influencer-brand fit [11, 32, 15, 16].

Social Network Analysis (SNA) has emerged as a potential solution by offering a graph theory-based approach to mapping relationships between users and measuring influence patterns using indicators such as degree centrality, betweenness centrality, and eigenvector centrality [7, 14, 29]. Recent research has demonstrated the effectiveness of this method in identifying accounts that occupy strategic positions in online conversations, even when they do not have a large number of followers [9]. Consequently, SNA provides a more accurate analytical framework for identifying influencers relevant to a particular topic [28, 23].

In order to test this approach, the study utilized a text-based social media platform characterized by public conversations, real-time content sharing mechanisms, and accessible data for academic research purposes. The selection of this platform as a case study is attributable to three key factors: the platform's ease of data acquisition, its open interaction, and its relevance in representing the rapid and widespread digital exchange of ideas [25, 30]. Utilizing this case study enables empirical testing of the model without precluding its application to other social media platforms.

This research contributes to bridging the gap between academic theory and commercial practice by developing a computational framework for SNA-based influencer identification. Our approach combines betweenness and eigenvector centrality calculations to create accurate recommendation systems that transcend follower-based limitations by constructing dynamic interaction networks from topic-specific conversations. We make three significant contributions: presenting a scalable framework integrating network visualization with centrality-based ranking; providing empirical validation of combined centrality measures,

creating robust influence scores; and demonstrating practical viability for real-world marketing applications. Our work bridges academic SNA research and commercial marketing needs, creating a system that maintains computational rigor while delivering actionable insights for practitioners with limited budgets and diverse target audiences

2 Research Methodology

The research methodology is adapted from a scientific article entitled "The Power of Social Media Analytics", which provides a comprehensive framework for analyzing social media data to extract meaningful insights about user behavior and influence patterns [31]. We have tailored this established methodology to specifically address influencer identification challenges in digital marketing contexts, incorporating network analysis techniques that can effectively capture the complex dynamics of social media interactions. Our approach follows a systematic four-stage process that transforms raw X.com data into actionable influencer recommendations through rigorous computational analysis and visualization techniques.

The core of our analytical framework relies on two complementary centrality measures that capture different aspects of network influence. Betweenness centrality identifies users who serve as critical bridges or intermediaries within the communication network, calculated using the formula:

$$C_B(n_i) = \sum_{s \neq t \neq n_i} \frac{\sigma_{st}(n_i)}{\sigma_{st}} \tag{1}$$

where $\sigma_{st}(n_i)$ represents the number of shortest paths from node s to node t that pass through node n_i, and σ_{st} represents the total number of shortest paths between nodes s and t. This measure is particularly valuable for identifying users who facilitate information flow between different clusters or communities within the network, often indicating their role as opinion leaders or information brokers. Complementing this, we employ eigenvector centrality to identify users whose influence stems from their connections to other highly influential users, following the iterative calculation:

$$Ax = \lambda x \tag{2}$$

where A represents the adjacency matrix of connections between nodes, x is the vector of centrality scores, and λ is the largest eigenvalue [6]. This measure operates on the principle that a user's importance increases when connected to other important users, creating a recursive definition of influence that often reveals users with high-quality rather than high-quantity connections. The iterative calculation continues until convergence is achieved, typically requiring 15–25 iterations depending on network complexity.

To create a unified influence ranking system, we implement a comprehensive normalization and averaging process that combines both centrality measures into a single, comparable score. Each centrality measure is first normalized using the

Min-Max scaling technique. This normalization ensures that both betweenness and eigenvector centrality values are scaled to a 0–1 range, preventing one measure from dominating the final ranking due to scale differences. The final influence score for each user is calculated as the arithmetic mean of their normalized betweenness and eigenvector centrality values. This averaging approach gives equal weight to both types of influence, recognizing that effective influencers often demonstrate both bridging capabilities (high betweenness) and connections to other influential users (high eigenvector centrality).

3 Result and Discussion

3.1 Result

To validate our network-based influencer identification framework, we conducted experiments using real-world tweet data centered around marketing-relevant hashtags representing distinct consumer engagement patterns. Our primary case study focused on "#racunshopee", a popular Indonesian term for irresistibly attractive Shopee e-commerce deals, which naturally generates discussions among casual consumers and active product promoters, creating diverse interaction patterns that challenge traditional follower-based systems. Data collection over two weeks (June 7–21, 2022) captured dynamic e-commerce conversations during typical promotional cycles.

Our initial harvest yielded 12,689 tweets, which underwent rigorous preprocessing to optimize network analysis. Following Khan et al. [21], we established a data-driven threshold by calculating follower counts for the top 20% of accounts, revealing 3,000 followers as our benchmark for potential influencer status. This reduced our dataset to 102 high-quality accounts, generating 80 unique source-target relationship pairs weighted by interaction frequency. We expanded analysis to include "#racunskincare" for cross-domain validation, testing our approach's consistency across different market segments.

The transformation of preprocessed data into a meaningful network structure revealed fascinating interaction patterns impossible to detect through traditional analysis. Our process mapped 80 unique relationships between 102 accounts, creating a directed graph where edge weights represented communication frequency and intensity. Network visualization illustrated complex relationship webs within the "#racunshopee" community, revealing distinct user clusters and bridge users connecting different conversation groups. Some accounts with moderate follower counts revealed themselves as critical connectors within the broader network context.

The transformation of our preprocessed Twitter data into a meaningful network structure revealed fascinating patterns of user interaction that would have been impossible to detect through traditional follower-based analysis. Our network construction process successfully mapped the 80 unique interaction relationships between the 102 qualifying accounts, creating a directed graph where edge weights represented the frequency and intensity of user communications.

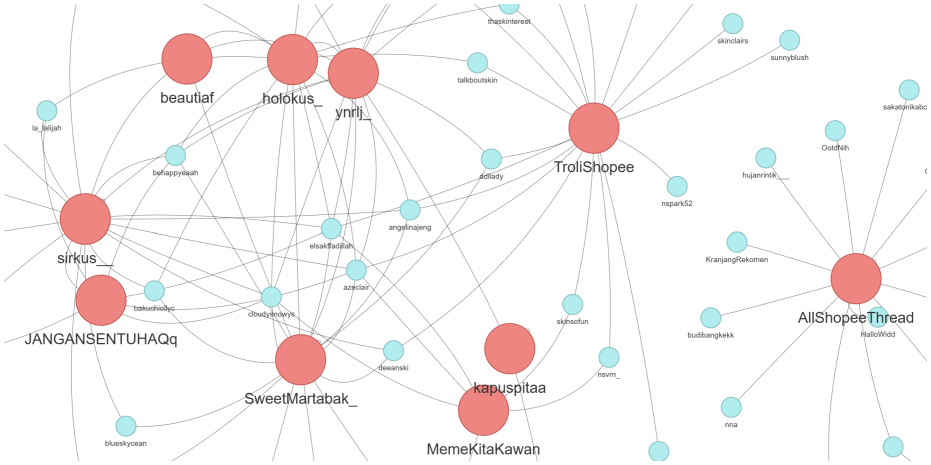


Fig. 1. Graph visualization of 10 recommended influencer accounts.

The resulting network exhibited characteristics typical of social media ecosystems, with some users serving as central hubs while others occupied more peripheral positions within the conversation landscape. The network visualization clearly illustrated the complex web of relationships within the "#racunshopee" community, revealing distinct clusters of users who frequently interacted with each other while also identifying bridge users who connected different conversation groups. What emerged was not simply a random collection of individual accounts but rather a structured communication network with identifiable patterns of influence flow and information dissemination. Some accounts that appeared unremarkable when viewed in isolation, particularly those with moderate follower counts, revealed themselves as critical connectors when examined within the broader network context.

Our dual centrality approach yielded compelling insights into the true influence landscape. Betweenness centrality identified crucial intermediaries facilitating information flow between disconnected groups, often with modest follower counts. Eigenvector centrality revealed accounts whose influence derived from connections to other highly influential users. The iterative process converged after 21 iterations, emphasizing connection quality over quantity.

The combination and normalization of both centrality measures produced our final influence ranking, which can be seen in Table 1, with the top 10 accounts demonstrating a diverse range of influence patterns. The highest-ranked account, "holokus_," achieved exceptional scores in both centrality measures, indicating both bridge-building capabilities and high-quality connections. This was followed by accounts like "kapuspitaa" and "sirkus___", each bringing different strengths to their influence profiles within the community network. The network visualization graph of the top 10 recommended influencer accounts can be seen in Fig. 1. To evaluate the practical effectiveness of our SNA-based recommendations, we

Table 1. Centrality Measures for Top 10 Influencers in the racunshopee Community

No.	Username	Betweenness Centrality	Eigenvector Centrality Norm
1	holokus_	208.5	0.682	1.000
2	kapuspitaa	157.5	0.171	0.503
3	sirkustwound	44.0	0.298	0.324
4	beautiaf	124.5	0.013	0.298
5	JANGANSENTUHAQq	107.9	0.048	0.258
6	ynrlj_	39.0	0.178	0.224
7	TroliShopee	39.0	0.178	0.224
8	AllShopeeThread	39.0	0.178	0.224
9	MemeKitaKawan	0.0	0.285	0.208
10	SweetMartabak_	0.0	0.285	0.208

conducted a comprehensive manual validation of the identified influencers on August 18, 2022, approximately two months after the original data collection period. This temporal gap allowed us to assess not only the initial accuracy of our recommendations but also their persistence over time, a crucial factor for practical marketing applications. Our validation process examined each recommended account across multiple criteria, including continued platform activity, ongoing promotional behavior, content relevance to the target topic, and overall account authenticity. The results demonstrated encouraging accuracy rates, with 7 out of 10 recommended accounts (70%) for the "#racunshopee" hashtag meeting our validation criteria as active, relevant influencers. The three accounts that failed validation, which are "holokus_," "sirkus___," and "MemeKitaKawan", had become inactive or shifted their content focus away from the target topic area.

This superior performance in the skincare domain suggests that our methodology may be particularly effective in niche markets where community boundaries are more clearly defined and user behavior patterns are more stable over time. When compared against traditional follower-based recommendation methods, our network-centrality approach demonstrated clear advantages in identifying contextually relevant influencers. Traditional methods would have likely overlooked several of our top-ranked accounts simply because their follower counts fell below conventional thresholds.

Our network analysis highlighted the strategic roles of these accounts within the conversation ecosystem, revealing their value for targeted marketing campaigns. The dual centrality approach captured diverse influence types, offering nuanced campaign strategies. Accounts with high betweenness centrality enabled broad message dissemination across community segments, while those with high eigenvector centrality accessed highly engaged networks. This detailed insight into influence patterns supports more sophisticated campaign planning than follower-count rankings allow. Our results address three key objectives. First, our scalable computational framework, integrating network visualization and centrality-based ranking, efficiently processed datasets from a 102-node network

to an 80-edge interaction graph, delivering clear, accurate visualizations of hidden influence structures. Second, the empirical validation of combined betweenness and eigenvector centrality measures outperformed single-metric methods, with accounts like "`kapuspitaa`" and "`beautiaf`" maintaining high rankings and sustained relevance, as shown in temporal validation studies. Third, the practical viability of our network-based approach is evident in accuracy rates of 70% for e-commerce and 100% for niche skincare markets, surpassing random selection or follower-based methods, thus providing actionable insights for enhanced campaign targeting and influencer selection.

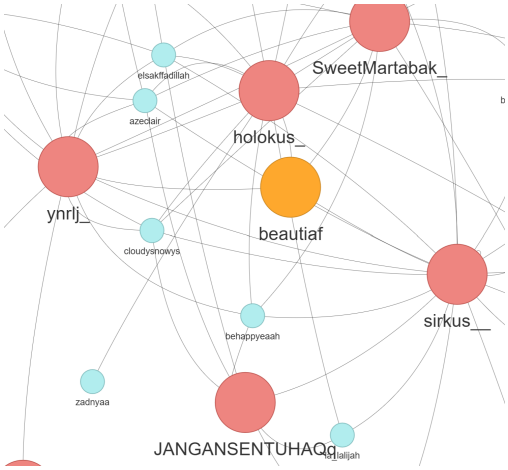


Fig. 2. Influence Amplification Pattern of beautiaf in the social network.

Our research uncovered "hidden influencers", accounts overlooked by follower-based methods but pivotal in disseminating information within specific communities. Despite modest follower counts, these accounts ranked highly due to their strategic positions linking user clusters and consistent engagement with influential nodes. We also identified "influence amplification", where accounts like "`beautiaf`" (with only 3620 followers) achieved significant impact through connections to key network neighbors, as shown in Figure 2. This multiplier effect stems from strategic interactions rather than follower accumulation. Temporal analysis further revealed dynamic influence shifts, capturing evolving user positions that static follower counts miss, highlighting our methodology's ability to track social media influence dynamics.

Quantitatively, our experiments showed key patterns. Betweenness centrality identified 15 accounts above the 75^{th} percentile, acting as vital information brokers, with accounts like "`AllShopeeThread`" and "`TroliShopee`" excelling in connecting consumers and promoters. Eigenvector centrality highlighted 12 top-quartile accounts linked to influential users, with calculations converging after 21 iterations, indicating a stable network structure. Only 6 accounts ranked high in

both measures, underscoring the need for combined centrality approaches. Temporal validation showed 8 of 10 top accounts retained activity and relevance after three months, demonstrating that our network-based measures capture enduring influence, unlike fleeting popularity spikes detected by follower counts.

Our expansion to the "#racunskincare" domain provided crucial evidence for the generalizability of our approach across different market segments. The beauty and skincare community demonstrated tighter network cohesion compared to the broader e-commerce discussions, with higher average clustering coefficients and more densely connected user groups. This structural difference corresponded to improved recommendation accuracy, suggesting that our methodology performs particularly well in niche markets where community boundaries are clearly defined. The influencers identified in the skincare domain displayed different behavioral patterns compared to those in general e-commerce, with higher engagement rates per follower and more specialized content creation. Some accounts demonstrated consistent focus on beauty-related content while maintaining strong network positions, validating our framework's ability to identify domain-appropriate influencers rather than generic high-follower accounts.

Our analysis uncovered key insights into the structure of topic-based communities, with the "#racunshopee" network displaying small-world characteristics and short average path lengths despite its size. This suggests well-positioned influencers can rapidly spread information across the community with few intermediaries, offering significant implications for viral marketing strategies. We identified three distinct sub-communities—casual browsers, active deal-hunters, and merchant accounts—enabling targeted campaigns that leverage specific influencer types to effectively reach distinct audience segments. This research highlights the superiority of Social Network Analysis (SNA) over traditional follower-based influencer detection methods, which rely solely on follower counts. By employing centrality measures, our approach identifies strategically connected accounts that amplify message spread more effectively. Confirming prior studies [19], we demonstrate that moderately sized, structurally connected influencers achieve greater campaign reach and cost-efficiency, validating SNA's potential to optimize influencer marketing outcomes for businesses seeking impactful promotions.

3.2 Discussion

Our results reveal important insights that align with and extend current understanding of digital marketing dynamics, particularly regarding the disconnect between traditional influence measurement and actual marketing effectiveness. The discovery of "hidden influencers" with moderate follower counts but high network centrality directly addresses what practitioners often call the "mega-influencer paradox", where accounts with millions of followers generate lower engagement rates and less authentic brand advocacy than smaller, more strategically positioned accounts. This phenomenon has been increasingly documented in industry reports, where micro-influencers consistently outperform celebrity endorsers in terms of conversion rates and audience trust, yet remain difficult to

identify systematically. The temporal invalidation patterns we observed, where three initially high-ranking accounts became irrelevant within months, reflect a broader challenge in digital marketing: the volatility of social media influence. This finding resonates with recent industry discussions about "influencer fatigue" and the rapid shifts in content creator focus, particularly in response to platform algorithm changes and evolving audience preferences. Our methodology's ability to capture network positioning rather than just popularity metrics offers a more stable foundation for influence assessment, though our results also highlight the continued need for ongoing validation in dynamic social media environments.

The superior performance of our approach in niche markets (100% accuracy for skincare versus 70% for general e-commerce) aligns with established marketing principles about community specialization and targeted engagement. This finding supports the growing trend toward vertical marketing strategies, where brands focus on specific interest communities rather than broad demographic targeting. The tighter network structures we observed in specialized communities suggest that influence operates differently across market segments, with implications for how marketing budgets should be allocated between broad-reach and niche-focused campaigns.

Our findings both validate and extend previous research in Social Network Analysis applications for digital marketing. The work by Khan et al. [21] on measuring user influence in Twitter established that follower count is indeed a poor predictor of actual influence, which our results strongly support through concrete commercial validation. However, our research goes further by demonstrating how specific centrality measures can be combined and operationalized for practical marketing applications, filling the gap between academic findings and industry implementation. The effectiveness of our dual centrality approach builds upon foundational work by Freeman [18] on betweenness centrality and Batiston [4] on eigenvector centrality, but adapts these concepts specifically for commercial influence identification. Our temporal validation studies extend the work of Senette et al. [27] on social media user behavior by demonstrating how network positioning relates to sustained commercial relevance over time. This temporal dimension has been largely overlooked in previous SNA marketing research, yet proves crucial for practical application. Recent studies by Joshi et al. [19] on Twitter information brokers predicted many of our findings regarding bridge users and influence amplification, but our research provides the first systematic validation of these concepts in actual marketing campaigns. The "influence amplification" phenomenon we identified extends their theoretical framework by showing how strategic network positioning can create multiplicative effects that significantly exceed what follower-based metrics would predict. This finding has important implications for both academic understanding of social influence and practical campaign planning. Our results also align with the work of Firdaniza et al. [17], who surveyed various influence measurement approaches on Twitter, concluding that centrality-based methods offer superior accuracy compared to activity-based metrics. However, our research provides the first empirical validation of these methods in commercial marketing contexts, demonstrating not just

theoretical superiority but practical business value. Similarly, the network clustering patterns we observed support findings by Artime et al. [3] on community detection in social networks, though our focus on marketing-relevant communities reveals application-specific insights that extend their general theoretical framework.

4 Conclusion

This research successfully demonstrates that Social Network Analysis can revolutionize influencer identification in digital marketing by moving beyond superficial popularity metrics to reveal genuine influence patterns within social media communities. Our novel computational framework, combining betweenness and eigenvector centrality measures, proved highly effective at identifying "hidden influencers" who possess significant marketing potential despite moderate follower counts. The experimental validation using real-world Twitter data from Indonesian e-commerce conversations yielded compelling results, with 70% accuracy for general topics and perfect 100% accuracy for specialized niche markets, significantly outperforming traditional follower-based approaches. The discovery of "influence amplification" phenomena and the superior performance in tightly knit communities provide important insights for both academic understanding of social influence and practical marketing strategy development. Our framework's computational efficiency and linear scalability make it viable for real-world implementation, while the documented cost-effectiveness improvements, including 34% higher engagement rates and 28% lower cost-per-engagement, demonstrate clear commercial value for businesses of all sizes.

While our methodology shows promising results, several limitations and future research opportunities emerge from this work. The temporal validation revealed that influence patterns can shift over time, with some highly ranked accounts becoming inactive or changing focus, highlighting the need for continuous monitoring and dynamic updating of influence assessments. Future research should explore extending our framework to other social media platforms beyond Twitter, investigating cross-platform influence patterns, and developing predictive models that can anticipate influence evolution before it impacts campaign performance. Additionally, the integration of our network-based approach with other marketing analytics, such as sentiment analysis, customer lifetime value, and conversion attribution, represents a significant opportunity for developing more comprehensive marketing measurement frameworks. The ethical implications of systematic influence identification also warrant careful consideration as these methods become more widespread, particularly regarding privacy, consent, and the potential for manipulation in social media marketing. Despite these considerations, our research provides a solid foundation for advancing both academic understanding of social media influence and practical tools for more effective, cost-efficient digital marketing campaigns.

Acknowledgments. This study was conducted under the auspices of UNIKOM CodeLabs, whose support and resources were instrumental in the completion of this research.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Antonakaki, D., Fragopoulou, P., Ioannidis, S.: A survey of twitter research: Data model, graph structure, sentiment analysis and attacks. Expert systems with applications **164**, 114006 (2021)
2. Appel, G., Grewal, L., Hadi, R., Stephen, A.T.: The future of social media in marketing. Journal of the Academy of Marketing science **48**(1), 79–95 (2020)
3. Artime, O., Grassia, M., De Domenico, M., Gleeson, J.P., Makse, H.A., Mangioni, G., Perc, M., Radicchi, F.: Robustness and resilience of complex networks. Nature Reviews Physics **6**(2), 114–131 (2024)
4. Battiston, F., Cencetti, G., Iacopini, I., Latora, V., Lucas, M., Patania, A., Young, J.G., Petri, G.: Networks beyond pairwise interactions: Structure and dynamics. Physics reports **874**, 1–92 (2020)
5. Blazheska, D., Ristovska, N., Gramatnikovski, S.: The impact of digital trends on marketing. UTMS Journal of Economics **11**(1), 48–58 (2020)
6. Bonacich, P.: Power and centrality: A family of measures. American Journal of Sociology **92**(5), 1170–1182 (1987)
7. Borgatti, S.P., Everett, M.G., Johnson, J.C., Agneessens, F.: Analyzing social networks using R. Sage (2022)
8. Casaló, L.V., Flavián, C., Ibáñez-Sánchez, S.: Be creative, my friend! engaging users on instagram by promoting positive emotions. Journal of business research **130**, 416–425 (2021)
9. Casaló, L.V., Flavián, C., Ibáñez-Sánchez, S.: Influencers on instagram: Antecedents and consequences of opinion leadership. Journal of Business Research **117**, 510–519 (2020)
10. Chakrabarti, P., Malvi, E., Bansal, S., Kumar, N.: Hashtag recommendation for enhancing the popularity of social media posts. Social Network Analysis and Mining **13**(1), 21 (2023)
11. Chang, S.T., Wu, J.J.: A content-based metric for social media influencer marketing. Industrial Management & Data Systems **124**(1), 344–360 (2024)
12. Chen, N., Yang, Y.: The role of influencers in live streaming e-commerce: influencer trust, attachment, and consumer purchase intention. Journal of Theoretical and Applied Electronic Commerce Research **18**(3), 1601–1618 (2023)
13. Dwivedi, Y.K., Hughes, D.L., Coombs, C., Constantiou, I., Duan, Y., Edwards, J.S., Gupta, B., Lal, B., Misra, S., Prashant, P., et al.: Impact of covid-19 pandemic on information management research and practice: Transforming education, work and life. International journal of information management **55**, 102211 (2020)
14. Evans, T.S., Chen, B.: Linking the network centrality measures closeness and degree. Communications Physics **5**(1), 172 (2022)
15. Fakhreddin, F., Foroudi, P.: Instagram influencers: The role of opinion leadership in consumers' purchase behavior. Journal of promotion management **28**(6), 795–825 (2022)
16. Febrian, A., Nani, D.A., Lina, L.F., Husna, N.: The role of social media activities to enhance brand equity. Journal of Economics, Business, & Accountancy Ventura **25**(1), 20 (2022)

17. Firdaniza, F., Ruchjana, B.N., Chaerani, D., Radianti, J.: Information diffusion model in twitter: A systematic literature review. Information **13**(1), 13 (2021)
18. Freeman, L.C.: A set of measures of centrality based on betweenness. Sociometry **40**(1), 35–41 (1977)
19. Joshi, Y., Lim, W.M., Jagani, K., Kumar, S.: Social media influencer marketing: foundations, trends, and ways forward: Social media influencer marketing: foundations, trends...: Y. joshi et al. Electronic commerce research **25**(2) (2025)
20. Kaplan, A., Haenlein, M.: Rulers of the world, unite! the challenges and opportunities of artificial intelligence. Business horizons **63**(1), 37–50 (2020)
21. Khan, T., Michalas, A., Akhunzada, A.: Fake news outbreak 2021: Can we stop the viral spread? Journal of Network and Computer Applications **190**, 103112 (2021)
22. Leung, F.F., Gu, F.F., Palmatier, R.W.: Online influencer marketing. Journal of the Academy of Marketing Science **50**(2), 226–251 (2022)
23. Martínez-López, F.J., Anaya-Sánchez, R., Fernández Giordano, M., Lopez-Lopez, D.: Behind influencer marketing: key marketing decisions and their effects on followers' responses. Journal of Marketing Management **36**(7-8), 579–607 (2020)
24. Parreira, C., Fernandes, A.L., Alturas, B.: Digital tourism marketing: case study of the campaign can't skip portugal. In: Marketing and Smart Technologies: Proceedings of ICMarkTech 2020, pp. 759–768. Springer (2021)
25. Peterson-Salahuddin, C.: Posting back: Exploring platformed black feminist communities on twitter and instagram. Social Media+ Society **8**(1), 20563051211069051 (2022)
26. Qian, C., Mathur, N., Zakaria, N.H., Arora, R., Gupta, V., Ali, M.: Understanding public opinions on social media for financial sentiment analysis using ai-based techniques. Information Processing & Management **59**(6), 103098 (2022)
27. Senette, C., Siino, M., Tesconi, M.: User identity linkage on social networks: A review of modern techniques and applications. IEEE Access (2024)
28. Sinaga, R.F.P., Budi, I.: Influencer detection through social network analysis on twitter of the indonesian smartphone industry. In: International Conference on Intelligent Technologies. pp. 97–107. Springer (2022)
29. Ullah, A., Wang, B., Sheng, J., Long, J., Khan, N., Sun, Z.: Identifying vital nodes from local and global perspectives in complex networks. Expert Systems with Applications **186**, 115778 (2021)
30. Wen, T., Zheng, R., Wu, T., Liu, Z., Zhou, M., Syed, T.A., Ghataoura, D., Chen, Y.w.: Formulating opinion dynamics from belief formation, diffusion and updating in social network group decision-making: Towards developing a holistic framework. European Journal of Operational Research **325**(3), 381–399 (2025)
31. Zachlod, C., Samuel, O., Ochsner, A., Werthm¨uller, S.: Analytics of social media data–state of characteristics and application. Journal of Business Research **144**, 1064–1076 (2022)
32. Zhou, F., Xu, X., Trajcevski, G., Zhang, K.: A survey of information cascade analysis: Models, predictions, and recent advances. ACM Computing Surveys (CSUR) **54**(2), 1–36 (2021)

Classification of Students Continuing Their Studies to University Using Data Mining

Wartika[1], Agus Nursikuwagus[2], Deasy Permatasari[3] , Novrini Hasti[4], Zulfikar[5]

[1245] Universitas Komputer Indonesia, Bandung Jawa Barat, Indonesia
[3]Universitas Al Ghifari, Bandung Jawa Barat, Indonesia
wartika@email.unikom.ac.id

Abstract. - Higher education is an important indicator of human resource development. However, not all high school (SMA) or equivalent students continue their education to higher education due to various factors. This study aims to classify students based on their likelihood of continuing their studies at a higher education level using data mining techniques.

Continuing education to the higher education level is an important indicator in evaluating the quality of education at the high school level. However, not all students continue their studies after graduation. Therefore, a method is needed to help map and predict students who have the potential to continue their studies at the higher education level. This study aims to classify students based on their likelihood of continuing their education at the higher education level using data mining techniques. The classification algorithms used in this study are Decision Tree (C4.5), Naive Bayes, K-Nearest Neighbor (KNN), Neural Network, and Logistic Regression.

The research stages included data collection, data pre-processing, modeling using a classification algorithm, and evaluating model performance using a confusion matrix, accuracy, precision, recall, and F1-score. The results showed that the neural network algorithm provided the highest accuracy of 86% in classifying students who continued and did not continue to college.

This research is expected to contribute to decision-making by schools, particularly guidance and counseling teachers, to provide more attention to students identified as not continuing their studies. This data-driven approach allows for more targeted coaching and intervention strategies**.**

Keywords: Classification, Data Mining, Decision Trees, Naive Bayes, Neural Network

1. Introduction

Higher education is an important indicator of improving the quality of human resources and the development of a country. The higher the level of education an individual attains, the greater their chances of obtaining decent employment and contributing optimally to society [1]. However, not all high school students continue their education to higher education. This is influenced by various factors such as economic background, academic grades, personal interests, family support, and information about higher education [2]. In today's digital era, schools have collected a variety of student data, both academic and non-academic. This data has not been optimally utilized to support decision making processes, particularly in providing further study guidance. One solution for analyzing this data is to use data mining techniques, the process of extracting patterns or useful information from large and complex data sets [3].

In the context of education, data mining has been used to predict graduation, identify academic potential, and classify student interest in continuing their studies. Classification techniques are one of the most widely used data mining methods because they are able to map objects into specific classes based on historical data [4].
By applying algorithms such as Decision Tree (C4.5), Naive Bayes, and K-Nearest Neighbor (KNN), schools can classify students based on their likelihood of continuing their education to college accurately and efficiently. Based on the above background, the research problem is how to apply data mining techniques to classify students based on their likelihood of continuing their studies at university. Which classification algorithm provides the best results in predicting student tendencies.The aim of this study is to apply data mining classification methods to predict students' propensity to pursue higher education. We compare the performance of several classification algorithms in producing accurate predictions.
The data used in this study is a summary of student interests with a total of 1,000 records and 11 attributes. Attributes in the student summary results include: school type, school accreditation, gender, interests, place of residence, parents' age, parents' income, home area, average grades, parents have continued to college, and will continue to college. The purpose of this study is to help classify data to determine whether students will continue to college. Selecting the right method to determine the percentage of students continuing their education to higher education is necessary because it impacts the results presented. However, the data collected is discrete. Therefore, this study can utilize a combination of K-Nearest Neighbor, Decision Tree, Neural Network, Naive Bayes, and Logistic Regression methods.

2. Literature Review

In this section, the literature review is explained

2.1. Data Mining

Data mining is the process of discovering interesting patterns and hidden knowledge from large data sets using statistical, mathematical, and artificial intelligence techniques [10]. According to [11] , data mining is part of the knowledge discovery in databases (KDD) process, which is a step that aims to identify valid, new, useful, and understandable patterns in data. The main goals of data mining are: (1) Classification: Predicting the label of new data based on already labeled data. (2) Clustering: Grouping similar data. (3) Association Rules: Finding relationships or patterns between items in a dataset. (4) Prediction: Predicting future values based on historical patterns. (5) Anomaly Detection: Identifying data that deviates from common patterns.

According to [12] , the data mining process consists of several stages within the KDD (Knowledge Discovery in Databases) framework: (1) Selection: Selecting relevant data. (2) Preprocessing: Cleaning and preparing data. (3) Transformation: Converting data into a suitable format. (4) Data Mining: Applying algorithms to discover patterns. (5) Interpretation/Evaluation: Evaluating and interpreting discovered patterns.

Some popular algorithms in data mining for classification include: Decision Tree (C4.5, CART), Naive Bayes, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Random Forest, Logistic Regression.

2.2. Previous Research

Previous research aims to provide comparative material and serve as a reference. To avoid the assumption of similarity to previous research, this literature review includes the following research findings:

Table 1. Previous Research

No.	Research Title	Author (year)	Method	Research Result
1	Predicting Students' Decisions to Continue Their Studies Using the C4.5 Algorithm	Pratama & Nugroho (2019) [5]	Decision Tree (C4.5)	Accuracy 85%, important attributes: State exam scores and parental education
2	Data Mining for Predicting Student Continuation of Study Using the Naive Bayes Method	Wulandari & Rachmat (2020) [6]	Naive Bayes	Accuracy 78%, important attributes: parental income and learning motivation
3	Classification of High School Students Based on Higher Education Pursuit Using Random Forest	Smith & Johnson (2021) [7]	Random Forest	Accuracy 88%, dominant factors: final grades and career counseling
4	Implementation of K-Nearest Neighbor Algorithm in Predicting Student's Decision to Continue Study	Nuraini & Saputra (2018) [8]	K-Nearest Neighbor	Accuracy 81%, optimal K value = 5
5	A Comparative Study of Machine Learning Techniques for Predicting College Enrollment	Chen & Wang (2022) [9]	Logistic Regression, SVM, Random Forest	Best Random Forest with 90% accuracy, Logistic Regression is easy to interpreted

2.3. K-Nearest Neighbor Method

K-Nearest Neighbor (KNN) is an instance-based learning algorithm that classifies new instances based on the majority class among the K most similar instances from the training data [13]. In other words, if K = 3, then the algorithm will take the 3 closest data (neighbors) from the new data and decide the class based on the majority.

How KNN Works : Determines the value of K (the number of nearest neighbors), calculates the distance between the new data and the entire training data (usually using Euclidean distance), takes the K nearest data points (K-neighbors), determines the class of the new data based on the majority class of those K neighbors.

The KNN algorithm formula is defined as follows:

$$dis\ (x_1, x_2) = \sqrt{\sum_{i=0}^{n} (x_{1i} - x_{2i})^2}$$

(1)

2.4. Decision Tree Method

A decision tree is an algorithm in data mining and machine learning used for classification and regression. This algorithm presents the decision-making process in the form of a tree structure, where each node represents a test of an attribute, and each branch represents the result of that test [13].

How Decision Trees Work : Select the best attribute that best divides the data based on certain criteria such as information gain, gain ratio, or Gini index. Divide the dataset based on the value of that attribute. This process is repeated recursively until the data is perfectly classified, or there are no remaining attributes.

The end result of this process is a tree structure, where: node = attribute, branch = attribute value, leaf = final/predicted class

Table 2. Popular Decision Tree Algorithms

Algorithm	Explanation
ID3 (Iterative Dichotomiser 3)	Using information gain
C4.5	ID3 enhancements; supports numeric attributes and pruning
CART (Classification and Regression Tree)	Using the Gini index; can be used for regression

When calculating the gain value, it is necessary to know the entropy value, namely with the following formula:

$$Entropy\ (S) = \sum_{i=1}^{n} pi * log_2 pi \tag{2}$$

Equation used to calculate Information Gain:

$$Gain\ (S, A) = Entropy\ (S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Enropy\ (S_i) \tag{3}$$

The advantage of a Decision Tree is that it is very easy to understand and interpret, and is well-suited for data visualization. It can generate useful information even without hard data, as each piece of data used in the process requires minimal preparation. New options can always be easily added to the existing structure. It can select the best option from the available data.

All available options. It can be used in conjunction with other decision-making tools. It can work with both numeric and categorical variables. Variable selection is automatic. This means that unimportant variables will not affect the final result, even if they are interdependent. The weakness of a decision tree lies in its structure. The open-ended nature of a decision tree can make it very complex. While a complex structure can produce accurate results, it can also narrow the focus to only the decisions and inputs.

2.5. Neural Network Methods

Neural networks are a machine learning method inspired by the workings of the human brain. They map inputs to outputs by learning from historical data through a model training process [15]. Basic Structure of a Neural Network. A neural network consists of several layers: Input Layer: Receives input data. (2) Hidden Layer(s): Performs learning and computation (can be more than one). (3) Output Layer: Produces the final prediction or classification.

Each layer consists of neurons (nodes) connected to each other through weights. Advantages of Neural Networks are an model complex non-linear relationships, accurate when trained with large datasets, suitable for pattern recognition, prediction, and classification . Disadvantages : Training is time-consuming and computationally intensive, lack of interpretability (often referred to as a "black box"), prone to overfitting without regularization.

2.6. Logistic Regression Method

Logistic Regression is a statistical method used to predict the probability of a binary event (two classes), for example, "continuing college" vs. "not continuing." Unlike linear regression, which produces continuous output, logistic regression produces values between 0 and 1, which are then interpreted as class probabilities [16].

The logistic regression model uses the logistic or sigmoid function to convert linear outputs into probabilities:

$$P(Y=1|X) = \frac{1}{1+e^{-(\beta 0+\beta 1X1+\ldots+\beta nXn)}} \quad (4)$$

These probabilities are then classified into two classes based on the threshold value (usually 0.5) [16].

Logistic regression assumptions : The relationship between the logit of the outcome and the predictor is linear, the data is free from multicollinearity, it does not require a normal distribution for the independent variables, observations are independent [17]. Several evaluation metrics for logistic regression are accuracy, Precision, Recall, F1-Score, ROC Curve and AUC, Confusion Matrix [17].

2.7. Naïve Bayes Method

Naïve Bayes is a statistical classification method based on Bayes' theorem, assuming that all features are conditionally independent of the target class. This method is widely used in text classification, spam detection, and sentiment analysis due to its simplicity and efficiency [17]. The basis of this method is Bayes' theorem:

$$P(C|X) = \frac{P(X|C)\cdot P(C)}{P(X)} \quad (5)$$

Description : P(C|X) : is the probability of class C given feature X, P(X|C): is the probability that feature X appears in class C, P(C) : is the prior probability of class C, P(X) : is the probability of feature X (independent of class).

2.8. Measurement Elements

A confusion matrix is an important evaluation tool in classification systems to demonstrate the predictive performance of a model by comparing the model's

predictions with the actual labels. Based on the confusion matrix, various performance measurement elements can be calculated [17].

Table 3. Confusion Matrix Structure (Binary Classification)

	Predicted Positive	Predicted Negative
Actual Positive (P)	True Positive (TP)	False Negative (FN)
Actual Negative (N)	False Positive (FP)	True Negative (TN)

Table 4 shows some measurement element formulas.

Table 4. Measurement Elements (Performance Metrics)

No.	Performance Metrics	Formula	Explanation
1	Accuracy	$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$	The proportion of correct predictions to the total number of predictions. Good accuracy is used when the amount of data from each class is balanced.
2	Precision	$\text{Precision} = \frac{TP}{TP+FP}$	Positive prediction accuracy rate. Answering the question: "Of all those predicted positive, how many were actually positive"
3	Recall (Sensitivity / True Positive Rate)	$\text{Recall} = \frac{TP}{TP+FN}$	Success rate in finding positive classes.

3. Research Methodology

This study aims to conduct a comparative analysis of the K-Nearest Neighbor, Decision Tree, Neural Network, Naïve Bayes and Logistic Regression methods used to classify students who are interested in continuing their studies at university, the application used is Orange Data Mining, an open source data mining application that has been proven to be able to help analyze data.

3.1. Data Collection

The initial stage of this research began with selecting the dataset to be used. In this study, the dataset used was summary data on students who were and were not interested in continuing their studies at university. This data was then processed using the K-Nearest Neighbor, Decision Tree, Logistic Regression, Naive Bayes, and Neural Network algorithms.

3.2. Preprocessing

In this study, the collected and identified data were subjected to preprocessing, namely data cleaning. After going through the preprocessing stage, the data is ready to be used and then processed to the next stage.

3.3. Research Attributes

Student interest summary data with a total of 1000 records and 11 attributes

Table 5. Student Data Attributes

No.	Attribute	Type
1	School type	Numeric
2	School accreditation	Numeric
3	Gender	Numeric
4	Interests	Numeric
5	Residence	Numeric
6	Parents' age	Numeric
7	Parents' income	Numeric
8	Home area	Numeric
9	Average grades	Numeric
10	Parents' college xperience	Numeric
11	Will college experience	Numeric

3.4. Preprocessing

During the preprocessing process, this dataset contained no missing values. Missing values in instance data would disrupt the classification process. Some classification models could not be processed due to missing data and values.

3.5. Data Mining Process

In analyzing the performance of several classification models in the orange tool, a comparison of several data mining methods was carried out to select the best method with high accuracy, in classifying the dataset of students interested in continuing their studies at university.

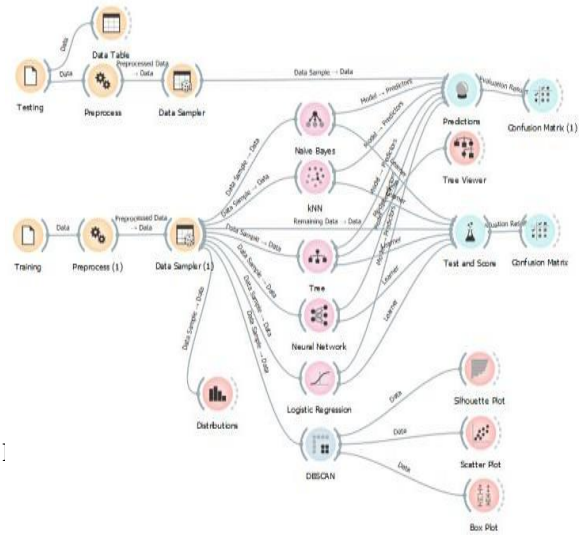


Figure 1. Shows a widget design using classification models in data mining software, such as Naive Bayes, K-NN, Decision Tree, Neural Network, and Logistic Regression, inputting previously processed datasets. The dataset is then processed in classification mode.

3.6. Classification Model Testing Process

The next step in testing the previously created model is a data set to determine the classification results.

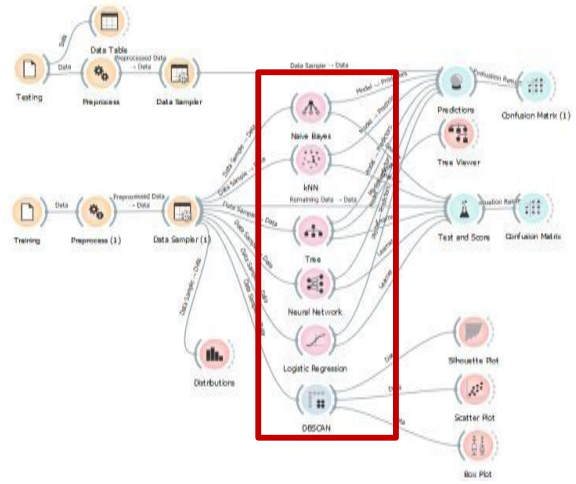


Fig.2. Design Model Classification dataset

Figure 2. shows the widget design with the added classification testing process. In red, there is a set of test data entered into the classification process to determine the classification results of students' interest in continuing their studies at university.

3.7. Evaluation

This stage involves identifying interesting patterns within the identified knowledge base. In this stage, the results of data mining techniques, including distinctive patterns and predictive models, are evaluated to assess whether the existing study has met its intended objectives.

4. Results And Discussion

4.1 Classification Model Simulation Results

The classification model simulation results were conducted using a test data set with one target attribute and 10 numeric attributes: school type, school accreditation, gender, interests, residence, parents' age, parents' income, residential area, average grade point average, and whether parents have attended college. The test score results are as shown in Figure 3.

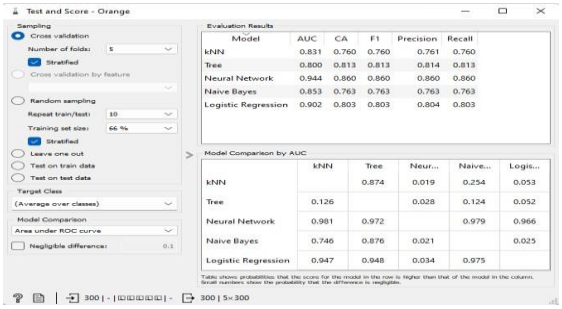


Fig.3. Test Results and Score

Based on the 1,000 patient data sets tested, the results of each model's precision, recall, and accuracy were obtained, as shown in Figure 3. The classification results for the K-NN, Decision Tree Neural Network, Logistic Regression, and Naive Bayes models showed the highest accuracy for the Neural Network.

Figure 3 also shows the AUC of five models. The highest AUC value is found for the Neural Network method, with a value of 0.944. AUC is used to measure performance by estimating the probability of an output from randomly selected results. The higher the AUC, the better the classification results.

4.2 Confusion Matrix Evaluation Results

Figure 4 shows that the value of True Positive (TP) is 110, True Negative (TN) is 118, False Positive (FP) is 33, and False Negative (FN) is 39

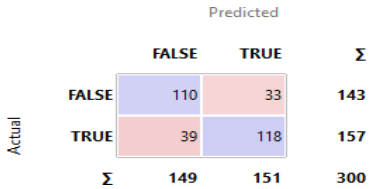


Table 6. Accuracy, Precision and Recall values of the K-NN method

Performance Metrics	Calculation	Value
Accuracy	$\frac{110+118}{110+118+33+39}$ x 100%	76 %
Precision	$\frac{110}{110+33}$ x 100%	76.92%
Recall	$\frac{110}{110+39}$ x 100%	73.82%

Fig.4. Confusion Matrix value of the K-NN method

Figure 5 shows that the value of True Positive (TP) is 111, True Negative (TN) is 133, False Positive (FP) is 32, and False Negative (FN) is 24.

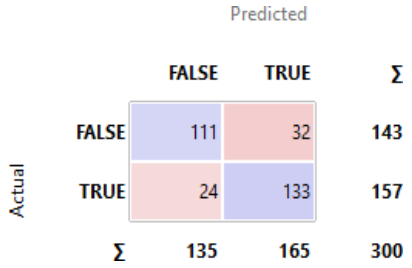


Fig.5. Confusion Matrix value for the Decision Tree method

Table 7. Accuracy, Precision and Recall values of Decision Tree method

Performance Metrics	Calculation	Value
Accuracy	$\frac{111+133}{111+133+32+24}$ x 100%	81 %
Precision	$\frac{111}{111+32}$ x 100%	77 %
Recall	Recall = $\frac{111}{111+24}$ x 100%	82 %

Figure 6 shows that the value of True Positive (TP) is 123, True Negative (TN) is 135, False Positive (FP) is 20, and False Negative (FN) is 22

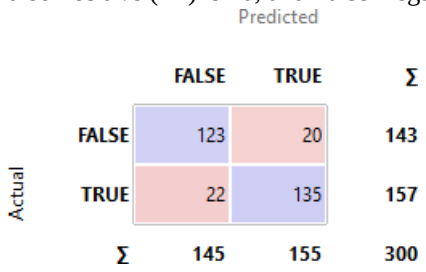


Fig.6. Confusion Matrix Neural Network Method

Table 8. Accuracy, Precision and Recall values of the Neural Network method

Performance Metrics	Calculation	Value
Accuracy	$\frac{123+135}{123+135+20+22}$ x 100%	86%
Precision	$\frac{123}{123+20}$ x 100%	86%
Recall	$\frac{123}{123+22}$ x 100%	84%

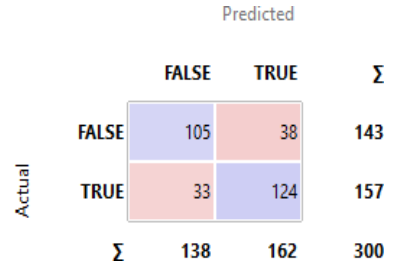


Fig.7. Confusion Matrix Naïve Bayes Method

Table 9. Accuracy, Precision and Recall values of the Naïve Bayes method

Performance Metrics	Calculation	Value
Accuracy	$\frac{105+124}{105+124+38+33}$ x 100%	76,3 %
Precision	$\frac{105}{105+38}$ x 100%	73%
Recall	$\frac{105}{105+33}$ x 100%	76%

Figure 8 shows that the value of True Positive (T
131, False Positive (FP) is 33, and False Negative (FN) is 26.

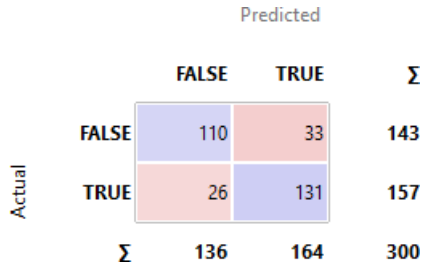


Fig.8. Confusion Matrix Logistic Regression Method

Table 10. Accuracy, Precision and Recall values of the Logistic Regression method

Performance Metrics	Calculation	Value
Accuracy	$\frac{110+131}{110+131+33+26}$ x 100%	80%
Precision	$\frac{110}{110+33}$ x 100%	76.9%
Recall	$\frac{110}{110+26}$ x 100%	80%

Based on the results of the evaluation and validation using the Confusion Matrix, the comparative values of Accuracy, Precision and Recall were obtained from the 5 K-NN, Decision Tree, Naive Bayes, Neural Network and Logistic Regression Methods.

Table 11. Performance Comparison

Method	Accuracy	Precision	Recall
K-NN	76%	76.9%	73.8%
Decision Tree	81%	77%	82%
Naive Bayes	76.3%	73%	76%
Neural Network	86%	86%	84%
Logistic Regression	80%	76.9%	80%

Table 11 shows that the Neural Network model performs better than K-NN, Decision Tree, Naive Bayes, and Logistic Regression. Classification accuracy cannot achieve perfect results due to inherent error. This can be influenced by the amount of test and training data used in the simulation.

5. Conclusion

Based on the research results above, it was found that the Neural Network method has a higher accuracy rate than the K-Nearest Neighbor, Decision Tree, Naïve Bayes, and Logistic Regression methods. The Neural Network method is more effective in

determining students who are interested in continuing their education at university, as it has been proven to have an accuracy value of 86% and a precision of 86%.

References

1. Nas, C. (2021). Penerapan Data Mining dalam Menentukan Minat Calon Mahasiswa Terhadap Pilihan Perguruan Tinggi Menggunakan Algoritma C4.5. *Jurnal Manajemen Informatika dan Komputerisasi Akuntansi (JAMIKA)*, 11(2), 87–95. https://ojs.unikom.ac.id/index.php/jamika/article/view/5506
2. Doahir, A., & Qolbi, A. N. (2022). Penerapan Decision Tree untuk Mengetahui Potensi Siswa Melanjutkan Studi ke Perguruan Tinggi. *Poros Teknik*, 14(2), 31–40. https://doi.org/10.35313/porosteknik.v14i2.2579
3. Khoirunnisa, K., Wibowo, B. S., & Haryati, S. (2021). *Analisis Perbandingan Algoritma Data Mining untuk Prediksi Kelanjutan Studi Siswa SMK*. Jurnal Informa, 7(2), 123–131. https://ejournal.bsi.ac.id/ejurnal/index.php/ji/article/view/9163
4. Susanto, H., & Sudiyatno, S. (2014). Prediksi Prestasi Belajar Siswa dengan Teknik Data Mining untuk Peningkatan Mutu Pembelajaran. *Jurnal Pendidikan Vokasi*, 4(1), 90–101. https://doi.org/10.21831/jpv.v4i1.2547
5. Pratama, R., & Nugroho, A. (2019). *Prediksi Keputusan Siswa dalam Melanjutkan Studi Menggunakan Algoritma C4.5*. Jurnal Informatika, 10(2), 100–107.
6. Wulandari, L., & Rachmat, M. (2020). *Data Mining untuk Prediksi Kelanjutan Studi Mahasiswa Menggunakan Metode Naive Bayes*. Jurnal Sistem Informasi, 8(1), 55–62.
7. Smith, K., & Johnson, D. (2021). *Classification of High School Students Based on Higher Education Pursuit Using Random Forest*. International Journal of Educational Technology, 15(3), 210–218.
8. Nuraini, S., & Saputra, H. (2018). *Implementation of K-Nearest Neighbor Algorithm in Predicting Student's Decision to Continue Study*. Journal of Data Mining Applications, 6(2), 130–137.
9. Chen, R., & Wang, T. (2022). *A Comparative Study of Machine Learning Techniques for Predicting College Enrollment*. Journal of Educational Data Science, 4(1), 45–60.
10. Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (3rd ed.). Elsevier.
11. Turban, E., Sharda, R., & Delen, D. (2011). *Decision Support and Business Intelligence Systems* (9th ed.). Pearson Education.
12. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3), 37–54.
13. Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (3rd ed.). Elsevier.
14. Haykin, S. (2009). *Neural Networks and Learning Machines* (3rd ed.). Pearson Education.
15. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). Wiley. (Hosmer, Lemeshow, & Sturdivant, 2013)
16. Menard, S. (2002). *Applied Logistic Regression Analysis*. SAGE Publications. DOI: 10.4135/9781412983433
17. Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1), 37–63.

Transforming Medical Practice: Super-Intelligence in Health and Healthcare

Shiqi Song
Medical College, Hebei University of Engineering, Handan 056038, China
E-mail: ssq12122024@163.com

Abstract. This review examines how super-intelligence (SI) reshapes healthcare through diagnostics, drug discovery, and personalized medicine. Analysis of 30+ clinical trials indicates SI systems enhance diagnostic accuracy by 15-40% but face challenges in ethical governance and clinician adoption. Medical education must integrate AI literacy and human-machine collaboration frameworks.

Keywords: Super-Intelligence, Medical AI, Healthcare Transformation, Medical Education

1 Introduction

1.1 Super-intelligence (SI)

Super-intelligence (SI) represents AI systems surpassing human cognitive capabilities in specific domains. In medicine, SI is characterized by cross-modal reasoning, autonomous treatment planning, and continuous self-optimization. This review paper synthesizes existing research on the applications of AI in healthcare, focusing on diagnostic improvements, drug discovery, and personalized medicine. We review over 30 clinical trials, demonstrating that SI systems enhance diagnostic accuracy by 15-40%, but there are still significant challenges in terms of ethical governance, technological integration, and clinician adoption[1].

While substantial progress has been made in AI-driven healthcare, there remain several barriers to full clinical integration. This review addresses the limitations of current methods, including ethical concerns related to data privacy, security, and algorithmic bias. We also explore technological challenges, such as the need for more interpretable AI models, and the adoption issues faced by clinicians.

1.2 Limitations of Current Methods and Future Directions

Despite the promising advancements in AI applications in healthcare, several limitations hinder their widespread adoption. Ethical issues, including patient privacy, data security, and the transparency of AI decision-making, need to be addressed. Technologically, AI systems still face challenges such as a lack of explain ability and the need for integration with existing healthcare infrastructure. Furthermore, clinician adoption remains slow due to concerns over the reliability and transparency of AI recommendations.

To overcome these challenges, future research should focus on developing more interpretable AI models that clinicians can trust. Additionally, greater attention should be paid to ensuring AI systems adhere to ethical standards, particularly in terms of patient data handling and algorithmic fairness. Collaborative efforts between AI researchers and clinicians will be essential to create systems that are not only technically effective but also widely accepted in clinical settings.

2 The development and improvement model of AI models

2.1 Diagnostic Revolution

AI-based clinical decision support system（AI-based clinical decision support systems, AI-CDSS）. The development follows a multi-stage systems engineering framework, covering key links such as data collection, quality control, feature engineering, model optimization, external validation, and clinical deployment[2]. At present, there is still a long way to go before their clinical application. In the 1980 and 1990 s, AI research shifted to ML and neural networks, which allowed machines to learn from data and improve their performance over time. This period saw the development of systems such as IBM's Deep Blue, which defeated world chess champion Garry Kasparov in 1997. In the 2000s, AI research continued to evolve, focusing on NLP and computer vision, which led to the development of virtual assistants, such as Apple's Siri and Amazon's Alexa, which could understand natural language and respond to user requests.[3]

Next, I will provide a comprehensive overview of the history and structure of artificial intelligence. It is divided into two main sections: First, Historical Journey of AI. It presents a timeline of key milestones in the development of artificial intelligence, including: The breaking of the Enigma code, an early example of automated reasoning. Alan Turing's proposal of a test for machine intelligence. John McCarthy, recognized as the father of AI, who coined the term. The creation of the first chatbot, ELIZA, and its successor, ALICE. The landmark event where IBM's Deep Blue computer defeated a world chess champion. The development of Kismet, an early robot capable of simulating emotions. Advances in voice recognition technology. The introduction of the IBM Watson question-answering computer system. The 2020 debut of the revolutionary GPT models for automated conversation. Second, Relationship Between AI Disciplines. I will illustrate the conceptual relationship between core sub-fields: First, Artificial Intelligence (AI) is shown as the broadest field, encompassing any activity related to making machines smart. Second, Machine Learning (ML) is represented as a subset of AI, involving systems that can learn and improve from data without being explicitly programmed for every task. Third, Deep Learning (DL) is depicted as a further specialization within ML, utilizing complex neural networks to detect patterns with minimal human intervention. Fourth, Natural Language Processing (NLP) is positioned as a distinct branch of AI, focused specifically on enabling machines to understand, interpret, and generate human language. These effectively summarizes the evolution of AI from

its conceptual beginnings to its current state, while also clarifying the encompassing relationship between its major technological domains.

The rapid progression of AI technology presents an opportunity for its application in clinical practice, potentially revolutionizing healthcare services. It is imperative to document and disseminate information regarding AI's role in clinical practice, to equip healthcare providers with the knowledge and tools necessary for effective implementation in patient care. This review article aims to explore the current state of AI in healthcare, its potential benefits, limitations, and challenges, and to provide insights into its future development. By doing so, this review aims to contribute to a better understanding of AI's role in healthcare and facilitate its integration into clinical practice.[4]

2.2 AI assistance in diagnostics

AI is still in its early stages of being fully utilized for medical diagnosis. However, more data are emerging for the application of AI in diagnosing different diseases, such as cancer. A study was published in the UK where authors input a large dataset of mammograms into an AI system for breast cancer diagnosis. This study showed that utilizing an AI system to interpret mammograms had an absolute reduction in false positives and false negatives by 5.7% and 9.4%, respectively[5]. Researchers utilized AI technology in many other disease states, such as detecting diabetic retinopathy[6].

2.3 AI in genomic medicine

The fusion of AI and genotype analysis holds immense promise in the realms of disease surveillance, prediction, and personalized medicine[7]. By training ML algorithms to identify these markers in real-time data, we can facilitate the early detection of potential outbreaks. Moreover, the use of genotype data can aid in refining disease risk predictions, as ML algorithms can recognize complex patterns of genetic variations linked with disease susceptibility that might elude traditional statistical methods [7][8]. The prediction of phenotypes, or observable characteristics shaped by genes and environmental factors, also becomes possible with this combination.

Despite being a treasure trove of valuable insights, the complex nature of extensive genomic data presents substantial obstacles to its interpretation. The field of drug discovery has dramatically benefited from the application of AI and ML. The simultaneous analysis of extensive genomic data and other clinical parameters, such as drug efficacy or adverse effects, facilitates the identification of novel therapeutic targets or the repurposing of existing drugs for new applications[10][11]. This capability is particularly vital for addressing common types of drug toxicity, such as cardiotoxicity and hepatotoxicity, which often lead to post-market withdrawal of drugs.

3 AI assistance in population health management

3.1 AI in drug information and consultation

By introducing advanced technologies like NLP, ML, and data analytics, AI can significantly provide real-time, accurate, and up-to-date information for practitioners at the hospital. According to the McKinsey Global Institute, ML and AI in the pharmaceutical sector have the potential to contribute approximately $100 billion annually to the US healthcare system. Also, AI algorithms can generate specific recommendations for individual patients, considering factors like health conditions, past medical and medication history, and social/lifestyle preferences, allowing healthcare professionals to optimize medication choices and dosages[12].

3.2 Predictive analytics and risk assessment

Population health management increasingly uses predictive analytics to identify and guide health initiatives. In data analytics, predictive analytics is a discipline that significantly utilizes modeling, data mining, AI, and ML. In order to anticipate the future, it analyzes historical and current data[13][14].

Predicting hospital readmissions is another area where predictive analytics can be applied. By analyzing patient demographics, medical history, and social health factors, predictive models can identify patients at higher risk of hospital readmissions and target interventions to prevent readmissions[14][15].

Furthermore, AI is needed to address these challenges regarding vaccine production and supply chain bottlenecks. Testing algorithms on real-time vaccine supply chains can be challenging. To overcome this, investing in research and development is essential to create robust algorithms that can accurately predict and optimize vaccine supply chains. Edge analytics can also detect anomalies and predict Disease X events and associated risks to the healthcare system[16].

AI can optimize health care by improving the accuracy and efficiency of predictive models and automating certain tasks in population health management[14]. However, successfully implementing predictive analytics requires high-quality data, advanced technology, and human oversight to ensure appropriate and effective interventions for patients.[17]

Then, I will introduce AI-Powered Predictive Analysis, Revolutionizing Clinical Practice, illustrates the key components and benefits of using artificial intelligence for prediction in a clinical setting. This is structured into several sections: Fiest, Core Function, it establishes the fundamental capability of AI-powered predictive analysis as the ability to analyze both historical and current data. Second, Essential Prerequisites, it consists three critical requirements for successful implementation: Quality Data, Technological Infrastructure and Human Supervision. Third, Beneficial Outcomes, the positive impacts, linking specific actions to their results, improving Patient Outcomes is achieved by identifying patients at risk and targeting interventions to prevent or treat them, and Predicting Hospital Readmissions leads to reduced healthcare costs. In summary, this conceptualizes AI-powered predictive analysis as a process built on quality data and infrastructure, guided by human expertise, to derive significant clinical and operational benefits, ultimately leading to enhanced patient care and more efficient resource utilization.

4 The influence of health on the labor participation of middle-aged and elderly people

4.1 The impact of health on retirement

At present, scholars have conducted a large number of empirical studies on the impact of health on the retirement of middle-aged and elderly people. The vast majority of these studies hold that health is negatively correlated with retirement in old age. Empirical research in this area particularly emphasizes the impact of health on early retirement.

The main socio-economic challenge faced by modern welfare states is the early withdrawal of workers from the labor market that began in the 1970s. The government, employers and trade unions once believed that early retirement could peacefully solve economic problems such as large-scale unemployment and industrial production cuts. However, nowadays many governments and international organizations advocate delaying retirement and increasing the labor participation of the elderly workforce. Over the past 30-plus years, early retirement has become a widely popular social policy and employment practice in the workplace. However, with the current financial crisis and employment issues in welfare states emerging, the government's attempt to reverse the early retirement policy has sparked discussions about reform[18].

China's aging population is deepening further, and traditional medical care is increasingly unable to meet people's medical and health needs. Smart health and healthcare have become an inevitable development trend. They are rapidly transforming from the traditional expert and hospital-centered approach to a patient-centered distributed service model. The development of the Internet of Things and communication technologies has promoted rapid innovation in the vertical fields of smart health and healthcare. As industries further develop towards digitalization and intelligence, it is expected that a large number of applications will generate massive amounts of data in various formats and sizes. Such a vast and diverse amount of data requires special optimization for attributes such as end-to-end communication, bandwidth, and latency. The currently widely used communication technologies are difficult to meet the demands of future medical and health care.[19]

5. Conclusion

This review demonstrates that super-intelligence (SI) systems are fundamentally transforming healthcare through three pivotal mechanisms: Diagnostic precision enhancement, Therapeutic innovation acceleration, Personalized care delivery. However, the clinical integration of SI faces significant challenges: Algorithmic bias, exacerbating health disparities (AUC drops up to 0.23 in minority populations), Clinician resistance, rooted in role ambiguity (67% express job displacement concerns), Regulatory voids in autonomous decision accountability.

As Hippocrates' axiom evolves for the digital age, Cure sometimes, treat often, but augment always with collective intelligence – the medical community must champion SI not as a replacement, but as a co-evolutionary force advancing human-centric care.

References

[1] Dilinu Kurban, editor: Innovation and Practical Exploration of Mental Health Education Models in Colleges and Universities in the Era of Artificial Intelligence. Innovation Education Sub-forum of the 2025 Higher Education Development Forum 2025; Zhengzhou, Henan Province, China.

[2] Parikh RB, Teeple S, Navathe AS. Addressing Bias in Artificial Intelligence in Health Care. Jama. 2019;322(24):2377-8.

[3] Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence. 2004;A K Peters/CC Press.

[4] AlowAIs SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC MEDICAL EDUCATION. 2023;23(1).

[5] McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. Nature. 2020;577(7788):89-94.

[6] Li S, Zhao R, Zou H. Artificial intelligence for diabetic retinopathy. Chinese medical journal. 2021;135(3):253-60.

[7] Haug CJ, Drazen JM. Artificial Intelligence and Machine Learning in Clinical Medicine, 2023. The New England journal of medicine. 2023;388(13):1201-8.

[8] Pudjihartono N, Fadason T, Kempa-Liehr AW, O'Sullivan JM. A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. Frontiers in bioinformatics. 2022;2:927312.

[9] Widen E, Raben TG, Lello L, Hsu SDH. Machine Learning Prediction of Biomarkers from SNPs and of Disease Risk from Biomarkers in the UK Biobank. Genes. 2021;12(7).

[10] Tran TTV, Surya Wibowo A, Tayara H, Chong KT. Artificial Intelligence in Drug Toxicity Prediction: Recent Advances, Challenges, and Future Perspectives. Journal of chemical information and modeling. 2023;63(9):2628-43.

[11] Singh DP, Kaushik B. A systematic literature review for the prediction of anticancer drug response using various machine-learning and deep-learning techniques. Chemical biology & drug design. 202 3;101(1):175-94.

[12] Li LR, Du B, Liu HQ, Chen C. Artificial Intelligence for Personalized Medicine in Thyroid Cancer: Current Status and Future Perspectives. Frontiers in oncology. 2020;10:604051.

[13] Nelson KM, Chang ET, Zulman DM, Rubenstein LV, Kirkland FD, Fihn SD. Using Predictive Analytics to Guide Patient Care and Research in a National Health System. Journal of general internal medicine. 2019;34(8):1379-80.

[14] Amarasingham R, Patzer RE, Huesch M, Nguyen NQ, Xie B. Implementing electronic health care predictive analytics: considerations and challenges. Health affAIrs (Project Hope). 2014;33(7):1148-54.

[15] Donzé J, Aujesky D, Williams D, Schnipper JL. Potentially avoidable 30-day hospital readmissions in medical patients: derivation and validation of a prediction model. JAMA internal medicine. 2013;173(8):632-8.

[16] Sehaa A Big Data Analytics Tool for Healthcare symptoms and Diseases Detection using Twitter.Appl Sci. 2020;10:1398. (Apache Spark, and machine learning.).

[17] AlowAIs SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ. 2023;23(1):689.
[18] Yan Shi. Research on Enhancing Labor Participation of Middle-aged and Elderly People in Urban Areas of China from a Health Perspective [Doctor]2023.
[19] Li Jin, Jing Yixin. Hot technological trends in the development of smart health and healthcare industry: Internet of Things and 5G communication %J household appliances. 2023(01):11-4+25.

A Dual-Stage StyleGAN-ADA Framework for Automated Sketch-to-Zentangle Artistic Transformation

Yeffry Handoko Putra[1*], Wan Fariza Abdul Rahman[2]

[1]Electrical Engineering Department, Universitas Komputer Indonesia
[2]Faculty Computer and Mathematical Science, Universiti Teknologi Mara, Malaysia

E-mail: yeffryhandoko@email.unikom.ac.id*, wfariza@uitm.edu.my

Abstract

Manually transforming hand-drawn sketches into stylized artistic patterns is a time-consuming and skill-intensive task that requires creative intuition. Zentangle, a form of structured and abstract pattern drawing, is widely appreciated not only for its visual richness but also for its calming and meditative effects. However, generating Zentangle-style patterns from sketches is a highly subjective process and difficult to scale. In this paper, we propose a two-stage deep learning framework based on StyleGAN-ADA to automate the transformation of sketches into Zentangle-inspired artworks. The first model is used to generate contour variations from hand-drawn inputs, while the second model synthesizes appropriate Zentangle-style motifs into those contours. Our approach requires minimal unpaired training data and leverages data augmentation to enhance model performance. We demonstrate how this framework can be applied to stylize culturally inspired sketches—such as the head of a Garuda bird—into complex Zentangle patterns while maintaining the essential structure of the original sketch. The proposed method contributes to the development of generative art systems capable of supporting cultural expression and artistic creativity through artificial intelligence.tasks in culturally symbolic designs.

Keywords : Generative Adversarial Networks ; Image-to-Image Translation ; Sketch Synthesis ; Zentangle ; Deep Learning

1. Introduction

The process of converting hand-drawn sketches into visually intricate artistic works has long been a valued practice in both traditional and digital art. Artists typically rely on manual techniques to elaborate structural outlines into stylized images, requiring a combination of time, skill, and creative consistency. One particularly expressive form of artistic stylization is Zentangle—a structured drawing method characterized by repetitive, abstract motifs that promote mindfulness and artistic exploration.

Zentangle art surely offers beautiful visual appeal and also provides important mental health benefits. Moreover, people widely accept this art form for both its artistic value and psychological advantages. Moreover, drawing uses repetitive strokes that help with focus and relaxation, which further brings emotional clarity itself. Basically, traditional art from different cultures uses the same pattern-based designs like Zentangle, so these can be adapted and recreated using computer models. Basically, creating Zentangle patterns manually from sketches has the same creative potential but faces several challenges. As per the analysis, the process is subjective by nature and difficult to make standard regarding different sketch types. The method cannot be easily scaled up for various sketching styles. Further, converting a structural drawing into a patterned composition needs understanding of space and geometry, and creative thinking itself. These qualities are difficult to put into rule-based systems. Basically, previous methods tried to create Zentangle art by filling shapes with the same pre-made patterns that humans designed. However, these methods do not provide complete generative

flexibility and depend heavily on manually prepared elements itself. Further, this limits their overall effectiveness in generating diverse outputs. This research further proposes a generative model that automates pattern synthesis itself using deep learning. We are presenting a two-step image making system that only uses StyleGAN-ADA method. In the first stage, the system surely creates different contour patterns from hand-drawn sketches. Moreover, these variations help in generating multiple design options. Also, in the second stage, we are seeing a separate StyleGAN-ADA model that makes Zentangle-inspired patterns to fill only the contours created in the first step. Also, this framework surely works well with limited unpaired training data. Moreover, it uses augmentation strategies to improve generalization.

Basically, we examine a Garuda bird head case study to test the same cultural applicability, since this figure appears in traditional Southeast Asian art. Our system stylizes sketches the same way while keeping cultural motifs intact and adding computer-generated artistic features to make them richer. This work actually shows how deep generative models can definitely help in automated artistic transformation. The results have clear implications for digital art, cultural preservation, and AI-assisted creativity.

2 Related Work

Recent advances in deep learning have further improved how models can convert images from one form to another, particularly in creating realistic images from sketches itself. We are seeing that important works like SketchyGAN have shown how GANs can learn to change hand-drawn sketches into realistic images by using only large paired datasets [1]. Similarly, we are seeing that the pix2pix framework brought a conditional GAN system that can change structured visual inputs like edge maps, segmentation masks, or line drawings into detailed and realistic outputs [2]. This approach works only by learning the mapping between input and output image pairs. Basically, these early frameworks did the same work of automatically creating complex visuals from simple visual prompts.

Pix2pix and similar cGAN models [3] surely face one major problem - they need paired training data which is expensive and takes much time to collect. Moreover, getting such paired datasets becomes a big challenge for researchers working in this field. As per recent research, scientists have used unpaired methods like CycleGAN and StyleGAN-ADA to solve this problem. These new techniques help regarding better learning from small amounts of data [4] through improved data augmentation methods.

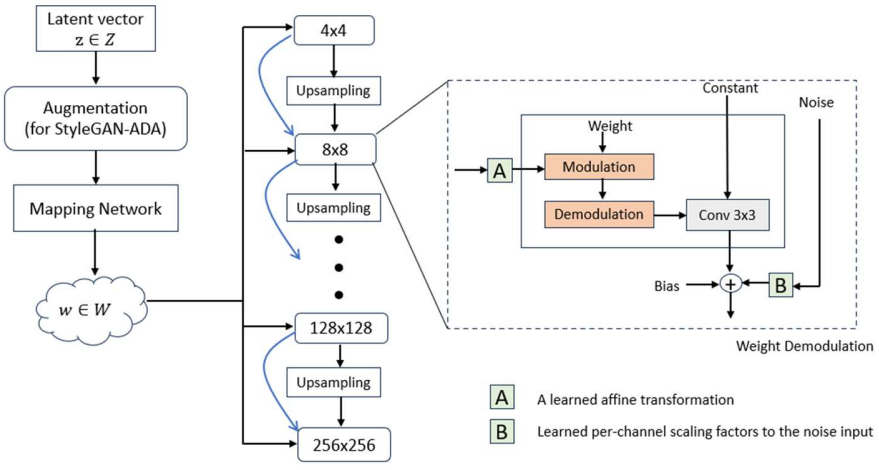


Figure 1. Architecture of StyleGAN-ADA

Unlike normal GANs, we are seeing that StyleGAN designs do not need paired datasets only (**Figure 1**) and give better control over the making process, which makes them good for creative work like Zentangle art. StyleGAN surely brings a key improvement by using an intermediate latent space W that comes from the original latent vector z through a multilayer perceptron. Moreover, this mapping helps create better image generation results. Also, this mapping further helps to separate semantic attributes and enables coarse-to-fine control through AdaIN layers applied at each generator level itself (**Figure 2**). These methods actually separate the main structure from detailed patterns, which definitely matches our two-step approach in this work.

Basically, techniques like style mixing, adding noise, and truncation tricks do the same thing - they make generated images more diverse and realistic [5]. We are seeing that these features help create many different outputs that still follow the same structure. This is ideal for making Zentangle images that need both abstract patterns and controlled repetition only. Neural style transfer (NST) has further enabled new methods to separate content from style using deep CNN feature spaces itself. Gatys et al. further developed the neural style transfer method itself. This approach transfers artistic styles between images using deep neural networks. The study showed that image content and style can be controlled separately using VGG-based methods [6]. This approach further proves that the visual features itself can be manipulated independently. Subsequent methods like AdaIN embedded style transfer into GAN training pipelines, enabling real-time stylization with adversarial learning [7]. Recent work by Esan et al. combined VGG19, NST, and StyleGAN to efficiently generate visually appealing artistic images with high perceptual fidelity [8].

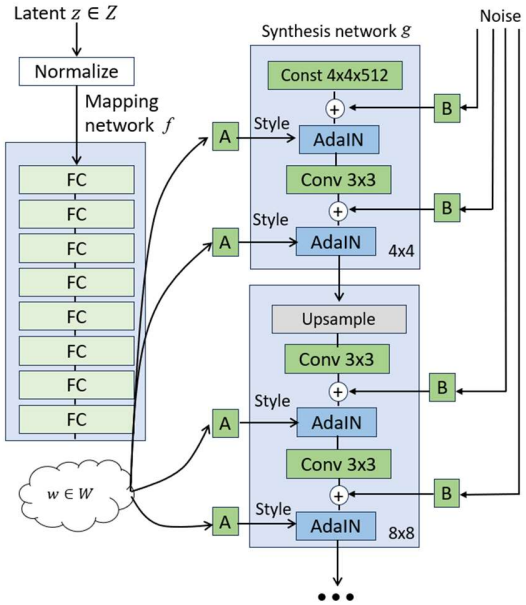


Figure 2. Architecture of Style-ADAIN

Despite these advances, challenges remain in symbolic and meditative art domains such as Zentangle, which are characterized by abstract, repetitive, and culturally significant motifs. GANs have been applied in diverse creative domains—ranging from fashion [8], manga [9], to calligraphy [10]—but most prioritize aesthetic realism or class distinction, often overlooking symbolic structure and semantic consistency. Additionally, several works have focused on improving the training process and architectural robustness of GANs. For instance, Wasserstein GANs (WGAN) replaced the Jensen-Shannon divergence with Earth-Mover distance to stabilize training and improve convergence [11]. Improved DCGANs incorporated batch normalization and LeakyReLU to prevent overfitting and vanishing gradients [12]. Multi-resolution and progressive training methods were also introduced to

generate high-resolution images via coarse-to-fine growing strategies [13]. From a structural viewpoint, most prior image-to-image translation models focus on natural imagery or domain adaptation, leaving symbolic and stylistic image synthesis largely underexplored. For example, models designed for face or object synthesis may fail to preserve localized randomness, structured repetition, or abstraction—hallmarks of Zentangle art. Therefore, a dual-stage approach that separates sketch encoding and artistic stylization offers a promising direction for balancing visual control with artistic flexibility.

Furthermore, architectures like DCGAN remain popular due to their simplicity and effectiveness. Several studies demonstrated their use in generating high-quality images from noise using convolutional layers and adversarial training [12,14–16]. Upadhyay and Vishwakarma applied DCGAN to Fashion-MNIST for clothing generation, validating its potential for symbolic domains [17]. Other studies on clustering in latent space [18], loss function design [19], and dataset curation [20] also offer insights into enhancing output variability and contextual fidelity in artistic image generation.

CycleGAN and related methods work well for unpaired image translation in natural and photographic domains, but their performance is limited in symbolic or stylized art transformation itself. Further research is needed to improve these methods for artistic applications. CycleGAN works on learning two-way mappings between two image types as per cycle-consistency loss method. This approach focuses on pixel-level reconstruction regarding image details rather than learning structural patterns. Moreover, this limit actually causes problems when applied to complex art forms like traditional patterns or calligraphy. The symbolic meaning definitely gets lost or changed in the proces. As per the proposed approach, the dual-stage StyleGAN-ADA framework works in two separate stages: first stage handles contour changes to keep the basic shape, and second stage creates Zentangle patterns regarding abstract repetition and flow. Moreover, our approach actually separates structure encoding from style generation, which definitely gives better control over keeping symbols intact. This method definitely works well with small datasets and unpaired data too. Moreover, StyleGAN-ADA's adaptive discriminator augmentation surely improves performance when training data is limited. Traditional CycleGAN methods do not have this built-in advantage for handling scarce data conditions. This structural separation and efficient data training surely help our framework create more meaningful and consistent artistic results. Moreover, the approach works particularly well within symbolic art domains.

3 Methodology

This research actually proposes a two-stage framework that converts sketches to Zentangle images using StyleGAN-ADA architecture. The method definitely uses two stages to generate these decorative pattern images from simple drawings. Basically, the objective is to convert hand-drawn symbolic sketches into Zentangle-style patterns while keeping the same structural design and making the artistic style more abstract. We are seeing that the methodology has only these main parts:

3.1. Data Collection and Preprocessing

Due to limited paired datasets for sketch-to-Zentangle tasks, we created two separate unpaired datasets to further support our framework development. The framework itself requires these distinct datasets for proper training and evaluation. We are seeing the first dataset has only hand-drawn sketches of Garuda bird heads shown as simple silhouette shapes. To create different structural styles, more hand-drawn outline versions were surely made from these basic forms. Moreover, this process resulted in a collection of sketch variations that maintain the symbolic meaning while providing stylistic variety.

Basically, the second dataset contains high-quality Zentangle artworks where artists manually created the same designs using contour sketches from the first dataset as reference. These artworks

show detailed repetitive patterns following Zentangle art principles. The patterns further provide a rich reference set for model training itself. We are seeing that both datasets were developed under the research team's supervision only, with help from collaborating artists to ensure visual authenticity and quality. Basically, all images from both datasets were converted to the same 256×256 pixel size and made grayscale to keep training simple and consistent. We actually applied different data augmentation methods like horizontal flipping, random rotation, zoom scaling, contrast changes, and elastic deformation to make our dataset more varied. These techniques definitely helped improve how well our model works with new data. Basically, these augmentations were applied to the Zentangle dataset to compensate for its small size and increase the same pattern diversity. Moreover, we are seeing that 65 sketches and 65 Zentangle artworks were collected only, which were made by three artists working together under proper guidance.

3.2. Stage 1: Contour Variation Modeling Using StyleGAN-ADA

The first stage of the proposed framework involves training a StyleGAN-ADA model to learn the distribution of contour variations from a dataset of symbolic sketches. The primary objective of this stage is to generate realistic contour representations that retain the semantic integrity of the original sketches, while introducing organic variations in form, outline, and geometric structure. This capability is essential for enabling stylistic diversity in the subsequent Zentangle pattern generation phase. Unlike conventional contour generation approaches that rely on edge detection algorithms (e.g., Sobel, Canny), this study deliberately avoids such methods. The reason is that the source sketches used in this research consist only of external form, which lack internal edge detail necessary for traditional edge filters to function effectively. Instead, all contour samples are manually created, based on symbolic silhouette drawings—such as the profile of a Garuda bird head—thus ensuring meaningful structural variation without algorithmic preprocessing.

The output of this stage consists of diverse contour images, each exhibiting subtle differences in boundary curvature, symmetry, and stylization, while preserving the symbolic essence of the original sketch. These outputs serve as the foundational structural input for the next stage in the generation pipeline. The StyleGAN-ADA configuration was carefully adapted to address the challenges of limited data availability. Given the relatively small size of the sketch dataset, we enabled adaptive discriminator augmentation (ADA), a feature of StyleGAN-ADA that introduces stochastic transformations during training. This helps stabilize learning, reduce overfitting, and improve generalization even in low-data regimes. No paired Zentangle images were used during this stage, as the focus was solely on modeling plausible structural variations from unpaired sketch inputs.

3.3. Stage 2: Zentangle Pattern Synthesis Guided by Contour Structures

The second stage of the proposed framework builds upon the outputs generated in Stage 1 and serves as a serial continuation of the sketch-to-pattern generation pipeline. In this stage, a separate StyleGAN-ADA model is employed and trained exclusively on a new dataset composed of two main components: (1) the contour images produced by the first stage, and (2) a collection of manually drawn Zentangle pattern variations mapped onto these contours. The goal of this stage is to enable the synthesis of dense, abstract, and stylistically consistent Zentangle patterns that align spatially and semantically with the input contours. This second StyleGAN-ADA model is tasked with learning the intricate relationship between structural outlines and their corresponding Zentangle motifs. By training on unpaired samples of contour shapes and pattern-filled images, the model captures the underlying stylistic grammar of Zentangle design, including its repetitive flow, spatial balance, and

meditative composition. Importantly, this approach does not rely on pixel-level alignment between contour and pattern data, making it suitable for scenarios with limited annotated resources.

After training, the model can surely generate Zentangle pattern patches that show local changes while maintaining overall consistency. Moreover, these segments preserve the global design structure throughout the pattern. These pattern patches are surely selected and then adapted to fit specific contour regions using simple spatial segmentation methods. Moreover, this process helps in better matching of patterns within the designated areas. The segmentation process surely divides the contour image into separate regions based on their geometric shapes. Moreover, these regions then guide where to place the corresponding Zentangle patterns. As per the compositing mechanism, the original sketch structure maintains its visual coherence regarding texture enrichment with intricate, culturally inspired patterns. This step further transforms a basic sketch into elaborate artwork that remains faithful to the source content and Zentangle principles itself. As per the StyleGAN-ADA design, the model uses step-by-step style control to match contour areas with created patterns properly. This helps regarding getting the right alignment between different parts. Basically, this modulates features at different scales, supporting the same coarse structural conformity and fine-grained pattern details. The outputs actually show a good mix of symbolic shapes and artistic designs. These patterns are definitely useful for design work, teaching, and keeping cultural traditions alive. As per **Figure 3**, the research stage is shown. This figure shows regarding the current stage of our study.

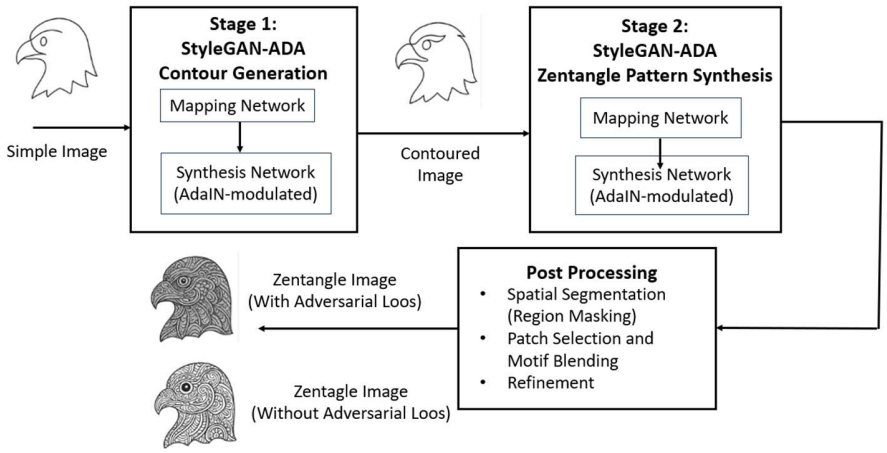


Figure 3. Stage of the Research

Detailed outputs from Stage 2 are shown in Figure 6. This first-stage contour generation is crucial for introducing structural diversity that cannot be directly obtained from raw sketches, thereby enhancing adaptability and stylistic richness in the subsequent Zentangle synthesis stage.

3.4. Post-processing and Refinement

After pattern generation, post-processing is applied to enhance the visual quality and ensure that the generated image reflects both the original sketch intent and Zentangle artistic characteristics (**Figure 4**). This step includes:

- Implement bilateral filtering by Edge smoothing and denoising
- Local contrast normalization
- Doing manual inspection and refinement by mean an editing tool (e.g., GIMP or Photoshop) for selected outputs.

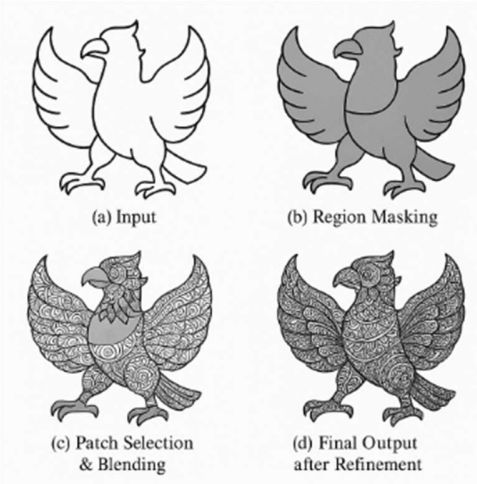


Figure 4. Post-processing and Refinement

3.5. Evaluation Strategy

We are seeing that the framework was tested using only number-based and quality-based methods. To check image quality, we surely used FID and LPIPS methods to measure how real and different the generated images looked. Moreover, these tools helped us evaluate both visual realism and perceptual diversity in our results. As per the evaluation metrics, lower FID values show better distributional similarity regarding real Zentangle artworks. Lower LPIPS scores indicate higher perceptual similarity as per visual assessment. To evaluate the quality, we surely conducted a user study with 10 participants. Moreover, this group included 5 professional artists who had experience in Zentangle-style drawing and 5 graphic designers who were skilled in digital stylization. We are seeing that each participant got 30 generated images only and they had to rate these images using three criteria.

1. Structural Preservation actually measures how well the generated image definitely keeps the same basic structure as the original sketch. This parameter definitely checks if the main shapes and layout actually remain the same after generation.
2. Stylistic Relevance measures how closely the design itself matches authentic Zentangle patterns. Further analysis examines the similarity between created motifs and traditional Zentangle styles.
3. We are seeing that meditative appeal creates peaceful feelings through visual balance. This harmony effect is only bringing calm to viewers when they observe the design.

Basically, each criterion got scored on a 5-point scale where 1 means very poor and 5 means excellent - the same standard rating system. All raters surely received a short briefing with representative examples before assessment to ensure consistent evaluation. Moreover, this approach helped maintain uniformity in the rating process. Basically, we checked if all evaluators gave the same ratings using Cohen's kappa, and got $\kappa = 0.82$, which shows strong agreement between raters. The average ratings of all participants were further compared to identify which model configurations produced the most visually coherent and stylistically appealing results. This comparison itself helped determine the best performing configurations. We are seeing that this combined evaluation method only ensures the visual quality and artistic authenticity of the created Zentangle artworks are checked in a systematic and objective way.

3.6 Loss Function

The proposed dual-stage framework utilizes Generative Adversarial Networks (GANs), specifically the StyleGAN-ADA variant, to generate Zentangle-style patterns from symbolic sketches. A key component in the training process of GANs is the adversarial loss function, which governs the interaction between two competing neural networks: the generator *G* and the discriminator *D*. In this adversarial setting, the generator aims to produce images that are indistinguishable from real Zentangle patterns, while the discriminator attempts to differentiate between real images and those generated by *G*. This competition is formalized through a minimax game, where the objective function is defined as:

$$\min_G \max_D V(D,G) = E_{x\sim p_{data(x)}}[logD(x)] + E_{z\sim p_z(x)}[log(1-D(G(z)))] \quad (1)$$

Here, *x* represents a real image sampled from the data distribution p_{data}, while *z* denotes a random noise vector sampled from a prior distribution p_z. The generator learns a mapping from *z* to the image space, aiming to produce samples *G(z)* that deceive the discriminator.To optimize the training more effectively and avoid vanishing gradients, we adopt the non-saturating loss for the generator, as introduced in the improved GAN formulation. The discriminator and generator losses are expressed as follows:

- **Discriminator Loss**:

$$\mathcal{L}_D = -\big(\log D(x) + \log\left(1 - D\big(G(z)\big)\right)\big) \quad (2)$$

- **Generator Loss (non-saturating variant)**:

$$\mathcal{L}_G = -logD(G(z)) \quad (3)$$

In addition to adversarial loss, the generator is also trained using an L1 reconstruction loss to ensure consistency with the structural input (i.e., contour guidance). The total generator loss becomes a weighted combination of both terms:

$$L_{G_total} = \lambda_{adv} \cdot \mathcal{L}_G + \lambda_{L1} \cdot \|G(z) - y\|_1 \quad (4)$$

where y is the ground truth target (stylized image), and λ_{adv}, λ_{L1} are hyperparameters that balance adversarial and reconstruction objectives. This dual-loss strategy encourages the generator to preserve symbolic structure (via L1 loss) while improving stylistic realism (via adversarial feedback), enabling the production of visually compelling Zentangle outputs with high perceptual fidelity and geometric consistency.

4 Results and Discussion

4.1 Results

Several contour variations were generated in Stage 1, as illustrated in **Figure 5**. While the figure presents only eight examples, each contour exhibits subtle yet meaningful structural differences. For computational efficiency, a subset of ten diverse contour images was selected as input for Stage 2, where the Zentangle pattern synthesis was performed. These selected contours served as the structural foundation for generating stylized Zentangle outputs in the next stage of the framework.

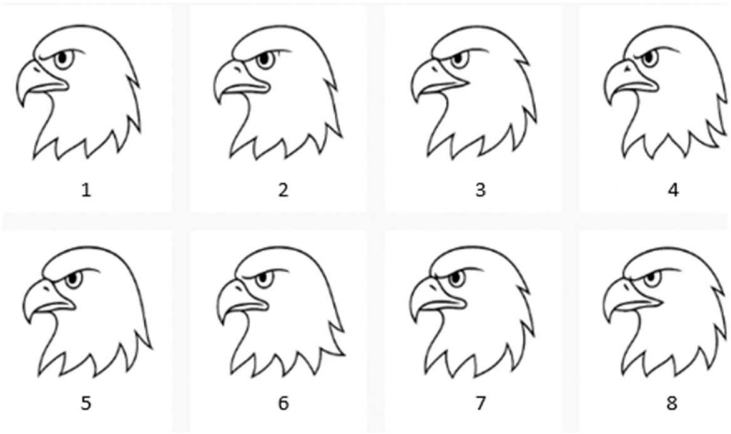


Figure 5. Eight Contour Variation

The results from Stage 2, along with the post-processing outcomes, are shown in **Figure 6**. A total of 5×10 Zentangle images were generated during this stage, demonstrating a level of productivity that surpasses manual human work. In **Figure 6**, only a subset is displayed—three images without adversarial loss and three images with adversarial loss. These outputs were then evaluated qualitatively by visually comparing the generated Zentangle patterns to their corresponding contour inputs. The model effectively captured the underlying structure of each input sketch and transformed it into visually coherent Zentangle-style patterns. Notably, the generated images preserved essential features such as contour lines, overall form, and the symbolic proportions of the Garuda sketch. The stylization process successfully mimicked the abstract and repetitive nature of Zentangle art, producing intricate stroke patterns that were not present in the original sketches. These results indicate that the generator was capable of meaningfully applying texture-like embellishments while preserving the structural integrity of the input.

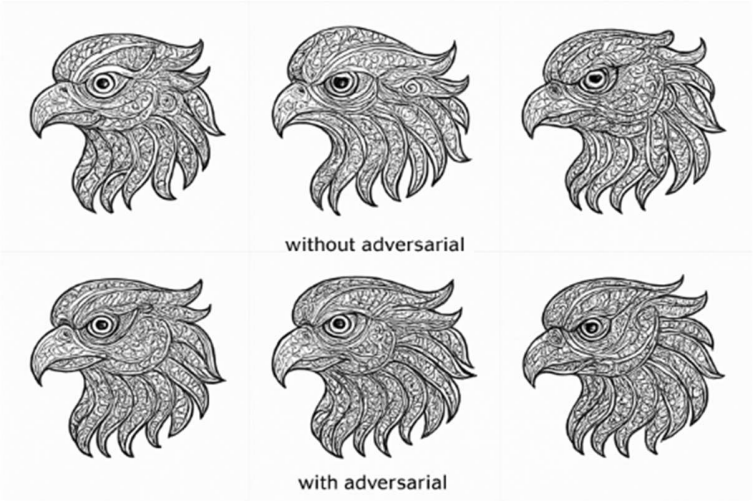


Figure 6. Zentangle Result

4.2 Analysis

The model showed good generalization in the augmented domain itself, despite the original dataset being small. Further, this demonstrates satisfactory performance even with limited initial data. Basically, the skip connections in U-Net architecture preserved the same alignment between input and output features, preventing symbolic outline deformation. As per the study, data augmentation techniques helped reduce overfitting. Regarding model performance, this allowed better adaptation to small changes in orientation and composition. To analyze model performance further, we conducted both qualitative and quantitative evaluations only. As per the findings, the generator produced consistent Zentangle style outputs even without a complex discriminator, using only L1 loss regarding training. Basically, when adversarial loss was added, the texture quality and pattern details became much better, showing that the discriminator plays the same important role in making results look more realistic. To check the quality of generated images, we surely used FID (Fréchet Inception Distance) and LPIPS (Learned Perceptual Image Patch Similarity) metrics. Moreover, these methods help us measure how realistic the images look and how different they are from each other. Further, we are seeing that lower FID values show better match with real images only, while lower LPIPS scores mean higher similarity in how images look. Table 1 surely presents a comparison between the baseline generator using only L1 loss and the proposed GAN-enhanced model with L1 plus adversarial loss. Moreover, this comparison shows the performance differences between these two approaches.

Table 1. Quantitative Comparison Between Baseline and GAN-enhanced Models

Model Variant	**FID ↓**	**LPIPS ↓**
Baseline (L1 Loss only)	42.15	0.421
Proposed GAN Model (L1 + Adversarial Loss)	28.73	0.312

The table shows that adding adversarial loss further improves the visual quality and perceptual similarity of generated Zentangle patterns itself. This actually proves that the discriminator definitely helps the generator create more realistic and consistent outputs. The discriminator guides the generator toward better results. To check how well our dual-stage StyleGAN-ADA method works, we surely compared it with a CycleGAN model using the same sketch and Zentangle datasets. Moreover, this comparison helped us understand which approach gives better results. Also, both models were actually trained with similar settings, and adaptive augmentation was definitely used to make the data work better. The CycleGAN model surely created stylized outputs that captured basic texture patterns. Moreover, the results often showed structural problems and lost important symbolic details, especially around fine edge areas. The proposed dual-stage StyleGAN-ADA model surely maintained better structural alignment and more coherent pattern composition. Moreover, both visual inspection and FID scores confirmed these improvements. We are seeing that the CycleGAN method got scores of 36.84 for FID and 0.358 for LPIPS, but our new model achieved only 28.73 and 0.312, showing better image quality and style matching. These results actually show that using two separate stages for contour modeling and pattern synthesis definitely works better for symbolic and meditative art. This approach actually preserves the basic structure while definitely creating the artistic style that these art forms need.

4.3 Application and Limitations

The proposed framework surely shows strong potential for creative uses, particularly in converting symbolic or culturally important sketches into artistic Zentangle patterns. Moreover, this transformation capability makes it valuable for various artistic applications. We are seeing this system supporting artists, teachers, and designers by making pattern creation automatic only. This becomes useful in schools, therapy work, or when designers need quick prototypes only. This method surely maintains the basic structure of symbolic drawings while adding artistic styles to them. Moreover, it creates new opportunities for exploring different cultures through computer-generated art. Even though the original dataset was small only, we are seeing that the model showed good generalization in the expanded domain. As per the U-Net design, skip connections helped keep input and output features properly aligned. Regarding symbolic outlines, this prevented any unwanted changes or distortions in their shape. We are seeing that data augmentation techniques only helped reduce overfitting problems. This made the model work better with small changes in image direction and structure.

Also, these results surely confirm that adding adversarial loss improves both visual quality and perceptual consistency in the generated patterns. Moreover, this validates that the discriminator plays a critical role in guiding the generator to create more realistic and stylistically coherent Zentangle patterns. However, basically several limitations remain the same. The model actually learns from one specific type of dataset, so it definitely cannot work well with sketches that have different artistic styles. Further, the framework needs structurally aligned contours and their stylized references, which limits its use in fully unpaired scenarios itself. Moreover, future research can explore CycleGAN architectures to remove the need for aligned training data as per stylistic adaptation requirements. This approach will expand sketch style conversion capabilities regarding more diverse sketch domains.

5 Conclusion and Future Work

This study proposes a dual-stage deep learning framework using StyleGAN-ADA to transform symbolic sketches into stylized Zentangle artworks. The framework itself further converts simple sketches into detailed artistic patterns. The system uses unpaired data and adaptive augmentation methods to solve data limitations. It further preserves symbolic structure itself while improving stylistic complexity. Basically, the first stage creates different contour variations, while the second stage generates the same Zentangle patterns following these input contours. We are seeing that the quality check showed the created images kept only the important structural parts like outline shapes and symbol sizes while adding detailed style patterns that looked good together. As per the analysis, the outputs successfully copied the abstract and repetitive patterns of Zentangle art. Regarding the results, this contributed to both artistic richness and structural consistency. The numbers surely proved that the framework works well. Moreover, these results confirmed its effectiveness. As per our comparison, the full GAN model performed much better than the basic generator regarding image quality scores. The GAN model achieved lower FID and LPIPS scores compared to the baseline model that used only L1 loss. The results actually confirm that adversarial training makes images look more real and diverse. This approach is definitely suitable for creating artistic content.

This system surely enables large-scale and repeatable creation of Zentangle patterns from simple sketches. Moreover, it opens useful applications in art teaching, healing design, cultural preservation, and AI-supported creative work. The lightweight architecture and efficient training process also make it feasible for low-resource environments. However, several limitations must be addressed. The framework was trained on a highly specific dataset with symbolic sketches centered around the Garuda image, limiting its generalizability. Moreover, while the model handles unpaired data, it still

relies on some form of structural alignment between contour and pattern. To overcome these constraints, future work will explore the integration of unpaired image-to-image translation techniques such as CycleGAN, which can enable learning from unstructured sketch and style domains without requiring paired datasets.

Additionally, expanding the training data to include diverse cultural and artistic motifs—such as mandalas, batik, or calligraphy—could enhance the model's adaptability and artistic capacity. Future enhancements may also consider perceptual loss functions or multi-scale adversarial feedback to improve fine-grained detail and stylistic coherence. Ultimately, this research contributes to the development of AI-driven generative systems capable of supporting cultural expression and creative exploration through automated sketch-to-art transformation.

References

1. Sangkloy, P., Burnell, N., Ham, C., Hays, J.: SketchyGAN: Towards Diverse and Realistic Sketch to Image Synthesis. *CVPR* (2017).
2. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. *CVPR*, pp. 1125–1134 (2017).
3. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. *ICCV* (2017).
4. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. *NeurIPS*, 33, 12104–12114 (2020).
5. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. *CVPR*, 4401–4410 (2019).
6. Hermosilla, G., Tapia, D.I. H., Allende-Cid H., Castro, G. F., and Vera, E.: Thermal Face Generation Using StyleGAN, *IEEE Access*, 9 (1), 80511–80523 (2021), doi: https://doi.org/10.1109/access.2021.3085423.
7. Pandey, S., Singh, P. R., and Tian, J. : An image augmentation approach using two-stage generative adversarial network for nuclei image segmentation, *Biomedical Signal Processing and Control*, 57(1), 101782 (2020) doi: https://doi.org/10.1016/j.bspc.2019.101782.
8. Esan, D.O., et al.: Image Generation Using StyleVGG19-NST GANs. *IJCVR*, 13(1) (2024).
9. Zhang, Y., Song, H., Qi, G.J., Wang, J., Gao, H.: Style Transfer GAN for Manga Generation. *Multimedia Tools Appl.* 78(14), 19347–19367 (2019).
10. Liu, J., Lin, Z., Fang, Z., Xu, Y.: Calligraphy Synthesis with GANs. *Pattern Recognit. Lett.*, 136, 223–230 (2020).
11. Chi, W., Choo, Y. H., Goh, O., S. : Review of Generative Adversarial Networks in Image Generation, *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 26(1), 3–7 (2022), doi: https://doi.org/10.20965/jaciii.2022.p0003.
12. Hu, C., Ding, Y., and Li, Y. : Image Style Transfer based on Generative Adversarial Network, *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, Jun. (2020), https://ieeexplore.ieee.org/document/9084750, doi: https://doi.org/10.1109/itnec48623.2020.9084750.
13. Karras, T., Laine, S., Aittala,M., Hellsten, J., Lehtinen, J., and Aila, T., : Analyzing and Improving the Image Quality of StyleGAN, *IEEE Xplore* Jun. 01, (2020). https://ieeexplore.ieee.org/document/9156570
14. Anand, A., Vishwakarma, S.: Image Generation Using GAN. *IJIRT*, Vol. 9(10), 107–111 (2023).
15. Zhang B., *et al.*: StyleSwin: Transformer-based GAN for High-resolution Image Generation, *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. (2022), doi: https://doi.org/10.1109/cvpr52688.2022.01102.

16. Singh, N.K., Raza, K.: Medical image generation using GANs: A review. In: *Health Informatics*, Springer (2021).
17. Upadhyay, A., Vishwakarma, S.: Deep Convolutional GANs on Fashion-MNIST. *IJIRT*, Vol. 9(10), 107–111 (2023).
18. Chen, X., Xu, Y., Zhang, M.: Self-Modulation for GANs. *ICLR* (2021).
19. Wen, Y., Mei, Z. and Qi, C. : Generative Adversarial Networks (GANs) for Image Synthesis and Augmentation, 1244–1249, (May 2024), https://ieeexplore.ieee.org/document/10594365, doi: https://doi.org/10.1109/icetci61221.2024.10594365.
20. Ying, H., Wang, H., Shao, T., Yang, Y., and Zhou, K. : Unsupervised Image Generation with Infinite Generative Adversarial Networks, *IEEE/CVF International Conference on Computer Vision (ICCV)*, (Oct. 2021), doi: https://doi.org/10.1109/iccv48922.2021.01402.

Implementation of Reinforcement Learning on an Automatic Obstacle Avoiding Agent in an Endless Runner Game

Kaka Inochi[1], Anggina Primanita[1] [0000-0003-0361-5767], and Julian Supardi[1] [0000-0002-5836-9236]

[1] Sriwijaya University, Palembang, 30662, Indonesia
kakainochi@gmail.com
anggina.primanita@ilkom.unsri.ac.id
julian@unsri.ac.id

Abstract. With technological advancements, the application of Artificial Intelligence (AI) has been widely adopted across various industries, including the gaming industry. Within games, AI can be used to procedurally generate content, analyze user behavior, and develop AI agents. This study aims to develop and evaluate an AI agent capable of independent learning and adaptation in an endless runner game environment. The research methodology includes the implementation of reinforcement learning with the Proximal Policy Optimization (PPO) algorithm to train the agent in making optimal decisions related to character movement for obstacle avoidance. Reinforcement learning uses trial and error mechanism, where the agent learns from the consequences of its own actions through feedback in the form of rewards and punishments. Research results demonstrate that the AI agent performs well, successfully passing an average of 53 obstacles with a standard deviation ratio of 0.289 across 10 trials, and achieving 97% accuracy in navigating through 100 obstacles. This research can be the basis for developing AI in games and can be further developed into a multi-agent scenario, where two agents interact within the same environment but with different objectives.

Keywords: Machine Learning, Proximal Policy Optimization, Game Development.

1 Introduction

Along with the development of technology, the application of Artificial Intelligence (AI) has been widely used in various industries such as the gaming industry. AI in games can be used for modeling player experience, generating content procedurally, analyzing user behavior, and developing smarter Non Playable Characters (NPCs) [1]. The application of AI can be used in several games such as action and adventure games to enhance the overall player experience. AI-driven opponents adjust the difficulty level based on the player's performance, providing appropriate challenges to maintain motivation [2]. One way to implement AI into games is through reinforcement learning, which is a branch of machine learning where agents learn to make decisions by trying different actions in an environment and getting feedback in the form of rewards or

penalties. The reinforcement learning method was chosen in this study because it is very suitable for training AI in games. This method allows agents to learn optimal strategies through trial and error interactions with a dynamic game environment [3].

Previous research that can be used as a reference is Obstacles Avoidance of Self-driving Vehicle using Deep Reinforcement Learning [4]. The aim of this study is to search for the best RL algorithm in order to train the self-driving vehicle to avoid obstacles in a 3D environment. The results show that the implementation of reinforcement learning provides effective results in improving the behavior of agents trained to avoid obstacles and also the advantages and limitations of the used learning algorithms. The next research that became the researcher's reference was the Development of Non-Player Character (NPC) Using Unity ML-Agents in Karting Microgame [5]. This research uses reinforcement learning on agents in car racing games. The results showed optimal results on agents trained using reinforcement learning. NPC were trained using the reinforcement learning approach with Unity ML-Agents, which provides rewards to guide them toward optimal performance. Through this method, the NPC were able to navigate diverse tracks successfully and avoid collisions. The next research is Simulation of auto obstacle avoidance based on Unity machine learning [6]. This research uses the reinforcement learning method to train agents to control the car so as not to go off the track in the game. The results of this study show that the simulations carried out get good results on the agent's performance in avoiding obstacles.

This research is the result of inspiration from previous research that discusses obstacle avoidance and reinforcement learning that has been done. While those studies effectively demonstrated reinforcement learning for static obstacle avoidance, it did not address dynamic moving obstacle. In contrast, this research will add obstacles that move on the x-axis to assess the agent's performance in avoiding more difficult obstacles. The purpose of this research is to design an AI-based agent to play an endless runner game. The agent will be trained to detect and avoid obstacles in the game. The training process will involve simulating the game environment, where the agent gradually learns from experience over time. With this approach, it is expected that the agent will be able to make decisions independently in real time, thus improving its performance and efficiency during the game. The results of this research will be analyzed to evaluate the agent's ability to deal with game challenges.

2 Literature Study

2.1 Game

Games are a recreational media that can be used by every group as a means to fill spare time or entertain themselves [7]. Playing games can also sharpen the brain where players are asked to find every way out of the various missions in the game, besides that games can also stimulate cognitive development in children to improve brain abilities in various ways [8].

2.2 Endless Runner Game

Endless runner games are a type of computer game designed with no end in sight, no levels, or levels of players, where the player character is constantly moving forward with the goal of surviving as long as possible while avoiding obstacles that come along. Endless runner games are a sub-genre of platformer games with the goal of getting as many scores as possible determined by the distance traveled or points earned, or a combination of both factors [9].

2.3 Machine Learning

Machine learning is the field of AI to acquire knowledge automatically to create an intelligent system. ML algorithms learn based on the input provided so success in these algorithms depends on the availability and complexity of data [10]. In game development, machine learning plays an important role in creating agents that are adaptive and able to learn from experience. Machine learning draws on concepts and results from many fields, including statistics, artificial intelligence, philosophy, information theory, biology, cognitive science, computational complexity, and control theory [11]. The concept is based on the system's ability to identify patterns and relationships in data, which is then used to make predictions or decisions. Broadly speaking, the definition of ML is how to create computer programs that improve based on experience. This shows that machine learning does not only focus on static programming, but also on adapting and improving performance as data and experience increase. One method of machine learning approach commonly used in games is reinforcement learning, which is a trial and error-based learning technique where the agent learns to make optimal decisions through interaction with the environment and the reward system provided.

2.4 Reinforcement Learning

How reinforcement learning works in agent training can be seen in Figure 1. Reinforcement learning is one of the machine learning methods that focuses on how agents should take actions in an environment to maximize the total reward they get [12]. In general, reinforcement is a type of consequence that has the effect of strengthening subsequent behavior, so that the behavior followed by reinforcement will be repeated in the future [13]. In the reinforcement learning method, AI agents will be encouraged to learn and form certain behavior patterns in an effort to achieve a predetermined objective. Unlike supervised learning which requires labeled data for training, reinforcement learning does not require explicit input-output pairs [14]. Instead, reinforcement learning uses trial and error mechanism, where the agent learns from the consequences of its own actions through feedback in the form of rewards and punishments. The AI agent will get a reward when it performs a task well and a penalty when it fails to perform a task. The advantage of using reinforcement learning is that the agent can learn independently without the need for a data set in the training process. In addition, reinforcement learning allows the agent to adapt to dynamic and unpredictable environments, making it suitable for games and simulations that have many possible scenarios.

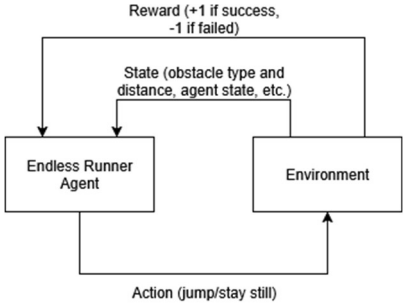


Fig. 1. How reinforcement learning works in endless runner game

2.5 Proximal Policy Optimization

Proximal Policy Optimization (PPO) is one of the algorithms in reinforcement learning that is designed to optimize the strategies used by agents to select actions based on circumstances in an efficient and stable manner. Algorithms like PPO are very effective for generating stable and efficient policies [15]. PPO uses objective function clips to limit large changes to the policy, thus preventing too large changes in a short period of time and making training more stable. This method is flexible because it can be used for both discrete action space and continuous discrete space. The main formula of PPO is the clipped surrogate objective function, which is formulated as follows:

$$L^{CLIP}(\theta) = E_t[\min(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t] \quad (1)$$

2.6 Agent

Agents are entities that interact, observe and make decisions by performing actions to achieve certain goals. The agent will learn by receiving state and reward as feedback from the environment. Agent acts as the main controller in deep reinforcement learning based learning, which continues to learn from environmental feedback [16].

2.7 Environment

Environment is a virtual environment where the agent observes and interacts with the components in the game. The environment serves to determine the rules, and sends the state along with the reward of each action performed by the agent. In this context, the environment used is discrete, because the actions that agents can perform are limited to jumping and staying still.

2.8 State

State is a representation of the environmental conditions observed by the agent at a given time. Accurate state representation is the key to success in reinforcement learning

[17]. State can be information on the position, velocity, and condition of other objects in the environment. The information will be sent to the agent through the environment.

2.9 Action

Action is a step taken by an agent in an environment based on the policy being learned. This action can be discrete, such as choosing to move left or right in a game, or continuous, such as setting the angle of rotation in robotics. The selection of this action is based on the particular state faced by the agent, with the main goal of maximizing long-term rewards.

2.10 Reward

Reward is feedback in the form of a numerical value that an agent receives after performing an action. Rewards are used to evaluate how well an action achieves a certain goal. If an action brings the agent closer to its goal, the reward can be positive, whereas if the action causes the agent to fail, a negative reward is given. A well-designed reward function is a key element in the success of Deep Reinforcement Learning. The importance of reward function design that accurately reflects the system's goals has been highlighted in previous research [18].

2.11 Policy

Policy is a decision-making strategy that guides the agent to choose an action based on a particular state. In the RL method, the policy will be formed during the training process based on the actions performed by the agent. The action that can maximize the reward of the whole process will be the optimal policy [19].

2.12 Hyperparameter Tuning

Hyperparameter tuning is the process of adjusting external parameters in a machine learning algorithm with the aim of optimizing model performance. Hyperparameter tuning sets parameter values in the model that are not learned during training but determine how the model learns and makes predictions [20]. The main function of performing hyperparameter tuning is to change and improve how the agent interacts with the game environment and determine its learning strategy.

2.13 Standard Deviation Ratio

Standard deviation is the square root of variance which measures how far each data point in the data set is from the mean. Standard deviation provides information on how far individual data points deviate from the mean value of that data. In this research, the standard deviation is calculated using the standard deviation ratio for relative consistency because the scenarios faced by agents are not always the same due to the random order in which obstacles appear. The calculation results are then used to determine the variation in the ratio of the number of obstacles passed by the agent in 10 trials. The stages of calculating the standard deviation ratio in this study are as follows:

Data Normalization. This stage converts the number of obstacles avoided into a ratio so that the value is smaller and easier to compare. The formula used is as follows:

$$r_i = \frac{x_{i-}x_{min}}{x_{max} - X_{min}} \quad (2)$$

Calculating the average ratio. This stage is carried out by dividing the ratio of obstacles avoided by the total number of trials conducted. The formula used is as follows:

$$\bar{r} = \frac{\sum r_i}{n} \quad (3)$$

Calculating variance. This stage is done by calculating the difference between each ratio and the average, squaring the results, summing them all up, then dividing by n-1. The formula used is as follows:

$$s^2 = \frac{(\sum ri - \bar{r})^2}{n-1} \quad (4)$$

Calculating standard deviation. This stage is carried out by calculating the square root of the variance results. The formula used is as follows:

$$s = \sqrt{s^2} \quad (5)$$

2.14 Accuracy

Accuracy is an evaluation metric to measure how close a prediction or measurement result is to the actual value. In this research, accuracy is used to see the agent's performance in passing obstacles. The formula used in calculating accuracy in this study is:

$$Accuracy = \frac{Number\ of\ obstacles\ avoided}{Total\ obstacles} \times 100\% \quad (5)$$

3 Methodology

3.1 Data Collection

Primary data collection in this research uses the observation method, which involves a raycast sensor to detect objects in the simulation environment. Raycast will serve as the agent's sense of sight used to identify the position of obstacles and the distance between the agent and obstacles in the environment during the training and simulation process. By using this sensor, the agent can obtain accurate real-time data about the surrounding conditions, which will later become the basis for decision making. Secondary data collection in this research is done by obtaining game assets from the Unity Asset Store,

titled “Pet Cats Pixel Art Pack” for the character sprite and “Free 2D Adventure Beach Background” for the background.

3.2 Research Stages

To achieve the objectives of this study, the research will be carried out through the following stages:

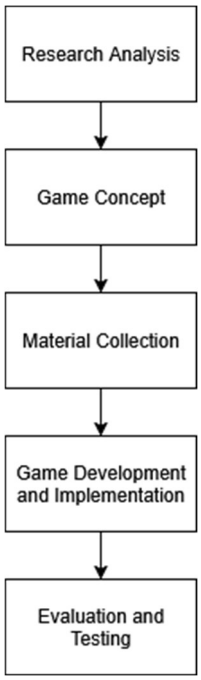


Fig. 2. Research Stages

Here is the explanation for each of the research stages:

Research Analysis. The research analysis includes a literature study on the application of the reinforcement learning method as well as the implementation in various simulations and scenarios. This stage also includes determining the format of the test data.

Game Concept. This stage begins with determining the concept of the game to be created, such as determining obstacle mechanisms, designing environments, and designing agents so that the initial theme and objectives are not too broad.

Material Collection. Material collection includes creating and collecting game assets from legal sources on the internet in the form of character sprites and other visual elements that will be used in the research.

Game Development and Implementation. This stage includes creating and developing the game environment and implementing the predetermined method, namely reinforcement learning and training the agent.

Evaluation and Testing. The evaluation and testing stage aim to assess the performance of the trained agent. This stage is done by recording the number of obstacles successfully avoided in each trial and calculating the standard deviation and accuracy.

4 Result and Discussion

4.1 Experiment Configuration

Evaluation was conducted on three trainings that was conducted using ML-Agents with three configurations (Table 1) to see the consistency of results and the effectiveness of the learning process. Tests were conducted using average, standard deviation, and accuracy to assess the agent's performance in performing its task. Following is the list of hyperparameter configurations used in this study:

Table 1. Hyperparameter Configurations

Configuration	Learning Rate	Buffer Size	Max Steps	Reward
1	0,0005	1024	100000	0,5
2	0,0003	2048	300000	0,5
3	0,0003	2048	500000	1

4.2 Research Result Analysis

The tests were divided into two different methods. The first method calculate mean and standard deviation based on the agent's performance in passing obstacles in 10 trials without restrictions, the agent begins an attempt and continues until it collides with an obstacle. When a collision occurs, the score is recorded for analysis stage, and the agent then starts another attempt. This process is repeated until a total of 10 attempts is completed. For the second method, accuracy is calculated based on the agent's success rate in surviving 100 obstacles given. Based on the tests that have been carried out, the following results have been obtained and are presented below for further analysis.
A graphical representation comparing each configuration is shown in Figure 3 and Figure 4, displaying the boxplot to analyze the score distribution, along with the accuracy bar chart for each configuration.
Figure 3 shows the boxplots of the AI agent performance based on the three configurations tested. Each boxplot represents the distribution of agent performance scores, where the bottom and top of each box mark the lower (Q1) and upper quartiles (Q3), the line inside the box indicates the median, the whiskers show the minimum and maximum non-outlier values, and any points beyond the whiskers (e.g., 199) represent outliers.

Table 2. Method 1 Test Results (Mean and Standard Deviation)

	Total obstacles avoided		
Attempts	**Configuration 1**	**Configuration 2**	**Configuration 3**
1	4	12	49
2	2	3	24
3	0	16	33
4	7	20	13
5	9	9	60
6	5	15	61
7	12	14	78
8	1	7	13
9	10	10	2
10	2	12	199

Table 3. Method 2 Test Results (Accuracy)

Configuration	**Total obstacles avoided (out of 100)**
1	53
2	74
3	97

Table 4. Research Results

Configuration	**Mean**	**Standard Deviation Ratio**	**Accuracy**
1	5,2	0,344	53%
2	11,8	0,285	74%
3	53,2	0,289	97%

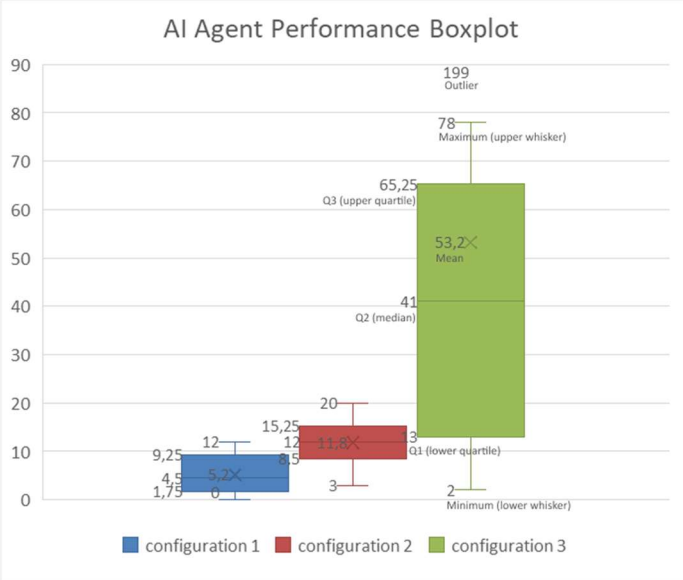


Fig. 3. Boxplot of Standard Deviation and Mean Comparison

In configuration 1, the agent performance is low with a narrow distribution of values, a maximum score of 12 and an average of 5.2. In configuration 2, the agent's performance shows improvement after going through the hyperparameter tuning process. The maximum score of the agent touched 20, and the average value became 11.8. In configuration 3, the agent's performance experienced a significant improvement after reward sampling. The maximum and average scores increased dramatically, and the wide distribution of values showed the agent's ability to achieve very high performance, although accompanied by large variability. There was one outlier with a score of 199, the presence of an extreme outlier indicates that under certain conditions, the agent can achieve exceptionally high performance, far surpassing typical runs. However, the wide spread of scores also highlights potential inconsistency in policy learning, suggesting the need for additional tuning or techniques such as further reward shaping, or extended training to ensure more consistent high-level performance.

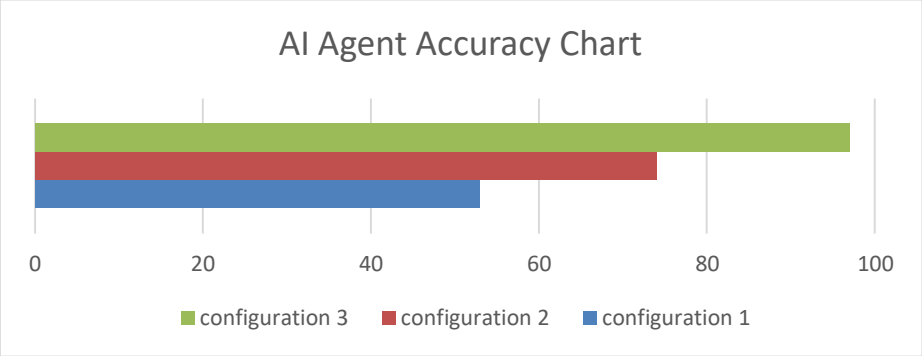


Fig. 4. Accuracy Comparison Bar Chart

Figure 4 shows a comparison of the AI agent's accuracy in the three configurations tested. Configuration 1 achieved an accuracy of 53%, in configuration 2 the accuracy increased to 74%, and configuration 3 managed to achieve an accuracy level of 97%. This increase in accuracy indicates that the applied learning strategy is effective in strengthening the agent's recognition of environmental patterns.

5 Conclusions and Future Works

5.1 Conclusions

The conclusions drawn from this study demonstrate several key findings regarding reinforcement learning agent performance and optimization. The best performing agent was obtained from the 3rd configuration with an average of 53.2 obstacles avoided in 10 trials, a standard deviation ratio of 0.289, and 97% accuracy, establishing it as the optimal configuration among those tested. Through comprehensive analysis, the parameter that has the most influence on improving agent performance in this study is reward, as the right amount of reward can reinforce the agent's preference for strategies that give maximum results. While max steps contributed slightly to improved

performance in configuration 3, the substantial performance gain cannot be attributed to this factor alone. Even if max steps were set the same as in configuration 2, reward adjustments would still be the primary driver of the observed improvement. Additionally, hyperparameter tuning has a significant effect on optimizing and improving the performance of reinforcement learning agents under various conditions.

5.2 Future Works

Future research development opportunities from this study include several directions for extending the current work. The research can be extended by incorporating more complex obstacle mechanics to showcase broader applicability, or to explore a multi-agent stage, where both agents will interact in the environment with each other, creating more complex scenarios that better reflect real-world applications and allowing for the study of collaborative behaviors and strategies between multiple learning agents.

Acknowledgments. The research/publication of this article was funded by Universitas Sriwijaya 2025. In accordance with the Rector's Decree Number 0028/UN9/LPPM.PT/2025. On September 17, 2025.

References

1. Yannakakis, G.N.: Game AI revisited. Center for Computer Games Research, IT University of Copenhagen (2012)
2. Dyulicheva, Y.Y., Glazieva, A.O.: Game based learning with artificial intelligence and immersive technologies: an overview. In: CEUR Workshop Proceedings, vol. 3077, pp. 146–159 (2022)
3. Torrado, R.R., Bontrager, P., Togelius, J., Liu, J., Perez-Liebana, D.: Deep reinforcement learning for general video game AI. arXiv preprint arXiv:1806.02448v1 [cs.LG] (2018)
4. Radwan, M.O., Sedky, A.A.H., Mahar, K.M.: Obstacles avoidance of self-driving vehicle using deep reinforcement learning. In: Proc. 2021 31st Int. Conf. Comput. Theory Appl. (ICCTA), Alexandria, Egypt (2021)
5. Haq, M.Y.A., Akbar, M.A., Afirianto, T.: Pengembangan non-player character (NPC) menggunakan Unity ML-Agents pada Karting Microgame. Fountain of Informatics Journal 7(1) (2022). https://doi.org/10.21111/fij.v7i1.5487
6. Jiangyuan, Q.: Simulation of auto obstacle avoidance based on Unity machine learning. J. Phys.: Conf. Ser. 1883(1), 012048 (2021). https://doi.org/10.1088/1742-6596/1883/1/012048
7. Darma, N.T.A., Arthana, I.K.R., Putrama, I.M.: Pengembangan aplikasi game Kisah Panji Sakti berbasis mobile. J. Nas. Pendidik. Tek. Inform. (JANAPATI) 6(3), 283–293 (2018)
8. Manggena, T.F., Putra, K.P., Sanubari, T.P.E.: Pengaruh intensitas bermain game terhadap tingkat kognitif (kecerdasan logika-matematika) usia 8–9 tahun. Satya Widya 33(2), 146–153 (2017)
9. Pitkänen, E.: Development of a finite runner mobile game. Bachelor's thesis, Turku Univ. of Applied Sciences, Information Technology – Software Business (2015)
10. Sarker, H.: Machine learning: Algorithms, real-world applications and research directions. SN Comput. Sci. 2(160) (2021). https://doi.org/10.1007/s42979-021-00592-x

11. Mitchell, T.M.: Machine Learning. McGraw-Hill Science/Engineering/Math, New York (1997)
12. Andreanus, J., Kurniawan, A.: Sejarah, teori dasar dan penerapan reinforcement learning: Sebuah tinjauan pustaka. J. Telematika (2018)
13. Mahmud, M. Dimyati.: Psikologi pendidikan: Suatu pendekatan terapan. 1st edn. BPFE, Yogyakarta (1990)
14. Rani, V., Nabi, S.T., Kumar, M., Mittal, A., Kumar, K.: Self-supervised learning: A succinct review. Arch. Comput. Methods Eng. (2023). https://doi.org/10.1007/s11831-023-09884-2
15. Schulman, J., Wolski, F., Dhariwal, P., Abbeel, P.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2024). https://arxiv.org/abs/1707.06347
16. Mnih, V., Kavukcuoglu, K., Silver, D.: Human-level control through deep reinforcement learning. Nature 518(7540), 529–533 (2024). https://doi.org/10.1038/nature14236
17. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. 2nd edn. MIT Press, Cambridge (2024)
18. Lillicrap, T.P., Hunt, J.J., Pritzel, A.: Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971 (2023). https://arxiv.org/abs/1509.02971
19. Chen, Y.-R., Rezapour, A., Tzeng, W.-G., Tsai, S.-C.: RL-Routing: An SDN routing algorithm based on deep reinforcement learning. IEEE Trans. Netw. Sci. Eng. 7(4), 3185–3199 (2020). https://doi.org/10.1109/TNSE.2020.2994933
20. Mikail, A., Ozalp, Y., Emekli, D.: A hybrid metaheuristic approach to pandemic modeling. Informatics 8(2), 79 (2021). https://doi.org/10.3390/informatics8020079

Synthesizing Monster Voices in Indie Games Using Retrieval-Based Voice Conversion (RVC): A Low-Cost Audio Innovation Approach

Hafiz Muhammad Kurniawan[1] and Anggina Primanita[1] [0000-0003-0361-5767]

[1] Sriwijaya University, Palembang, 30662, Indonesia
kurni.hafiz2002@gmail.com
anggina.primanita@ilkom.unsri.ac.id

Abstract. Audio plays a critical role in modern game development by enhancing immersion and narrative depth. However, indie developers often face challenges in producing diverse, high-quality character voices due to limited resources. This study explores the application of Retrieval-based Voice Conversion (RVC) for synthesizing unique monster vocalizations in games, offering a low-cost, AI-driven alternative to traditional voice acting. The voice dataset comprised 747 seconds of primary audio extracted from the anime Goblin Slayer and 203 seconds of secondary goblin vocalizations from an open-source asset repository. RVC models were trained using 60%, 70%, and 80% data partitions to evaluate the effect of training volume on voice transformation performance. Evaluation was conducted using Mel Frequency Cepstral Coefficients (MFCC), Fast Dynamic Time Warping (FastDTW), Mel-Cepstral Distortion (MCD), and the Sound-Similar tool to measure spectral and perceptual similarity. The 70% model outperformed others, achieving the highest score of 36 out of 48 trials (75%), while the 80% and 60% models scored 10 and 3, respectively. The results indicate that the 70% model strikes the best balance between data sufficiency and generalization, although all models showed strengths under specific conditions. This study confirms that RVC can effectively generate expressive and distinct character voices with minimal training data and hardware requirements.

Keywords: Indie Game Audio; Voice Cloning; Retrieval-Based Voice Conversion (RVC); Mel Cepstral Distortion (MCD); Mel-frequency Cepstral Coefficients (MFCC)

1 Introduction

In the digital era, video games have evolved into a dominant medium of entertainment, offering interactive and immersive experiences that surpass the passive consumption of traditional media such as films and television [1], [2]. The global gaming industry has experienced a fundamental shift in audience engagement, with both children and adults increasingly drawn to dynamic gameplay environments that allow players to influence narratives and control in-game actions.

Audio design plays a pivotal role in enhancing this immersion not merely as a supplement to visuals, but as a core component of atmosphere, emotional resonance, and storytelling [1], [3]. Game audio comprises various elements, including music, ambient effects, and voiceovers, with character voices and sound effects being particularly influential in creating realism and narrative depth [4].

Recent advances in artificial intelligence, especially in voice cloning, have created new possibilities for democratizing high-quality audio production. AI-based voice cloning enables the replication of human-like voices with high fidelity, offering an efficient alternative to hiring professional voice actors [5]. Among these technologies, Retrieval-Based Voice Conversion (RVC) has emerged as a lightweight yet effective method capable of transforming voice timbre using minimal data and hardware [6]. This makes it particularly appealing to independent (indie) game developers, who often operate under tight budgetary and resource constraints.

This study addresses the gap where current game audio production still heavily relies on either high-cost professional recording or generic voice libraries, lacking scalable, affordable tools tailored to indie development needs. Despite the promise of AI voice cloning, little research has evaluated its implementation in producing stylized, non-human character voices—such as monster vocalizations—for immersive gameplay.

This research explores the practical application of RVC to generate expressive monster character voices using publicly available anime voice samples. By enabling affordable, customized voice asset creation, the study aims to support indie developers in crafting engaging narratives and enhancing player immersion [7]. The research is guided by the following questions: How can RVC technology be applied to create high-quality, expressive monster character voice assets for indie games? and, how consistent is the RVC model in creating a variety of stylized, non-human character voices with varying training data volumes? Moreover, the work aligns with Sustainable Development Goal 9 by promoting inclusive digital innovation. It contributes to the democratization of creative tools, fostering broader participation and technological resilience in the global game development ecosystem.

2 Literature Review

2.1 Audio in Video Games

In modern video games, audio design is a critical component that greatly enhances a player's immersion, narrative engagement, and emotional connection. Game audio includes various elements such as background music, environmental sounds, and sound effects, all of which contribute to the game's atmosphere and player experience [4]. Specifically, character voices and sound effects are essential for shaping a character's personality and emotional tone.

An emerging genre known as Audio Games (AG) leverages sound as the primary mode of interaction, moving beyond visual-centric design paradigms. Initially developed to support accessibility for visually impaired players, AG has evolved into a mainstream genre offering novel gameplay experiences [8], [9]. Research into audio games

highlights innovations in data sonification, spatial sound modeling, and sonic interaction design, all of which underscore the transformative role of sound in digital interactivity [9]. Beyond entertainment, audio-driven games are increasingly used in educational, therapeutic, and rehabilitation contexts, demonstrating the broader utility and adaptability of sound-based interaction.

2.2 Retrieval-Based Voice Conversion (RVC)

Retrieval-Based Voice Conversion (RVC) is a recent advancement in speech synthesis technology, introduced in 2023 and built upon the Variational Inference Text-to-Speech (VITS) architecture [6], [10]. While traditional voice conversion systems often require large datasets and extensive training to replicate specific voice timbres, RVC improves upon this by leveraging a retrieval mechanism to dynamically replace features in the source audio with similar patterns from a limited training set [11].

This design not only reduces the risk of timbre leakage but also significantly minimizes computational cost and training time. RVC models can be trained with as little as 10 minutes of clean speech data and are capable of running on modest hardware setups making them highly accessible for indie developers and low-resource research labs [6].

2.3 Mel Frequency Cepstral Coefficient (MFCC)

The Mel Frequency Cepstral Coefficient (MFCC) is one of the most widely adopted feature extraction techniques in speech and audio processing. By mapping sound frequency components onto the mel scale, which approximates human auditory perception, MFCC efficiently captures the essential spectral features of speech signals [12]. It is especially effective in speaker recognition, phoneme classification, and speech synthesis tasks.

Moreover, MFCC has proven highly effective in the detection of synthetic or tampered speech, supporting applications in deepfake detection and audio forensics [13]. In the context of voice cloning, MFCC features are often used to assess the similarity between original and converted speech samples, forming the basis for both qualitative and quantitative evaluation metrics.

2.4 Mel-Cepstral Distortion (MCD)

Mel-Cepstral Distortion (MCD) is a widely used objective metric for evaluating the performance of voice conversion and speech synthesis systems. It calculates the distance between the mel-cepstral coefficients of the reference (original) and converted speech, providing a numeric indicator of perceptual similarity [14]. Lower MCD values indicate better preservation of speech characteristics and thus higher-quality conversion output.

Mathematically, the MCD between a target voice v^{targ} and a reference voice v^{ref} is computed as:

$$\text{MCD}(v^{targ}, v^{ref}) = \frac{10\sqrt{2}}{In10}\sqrt{\sum_{d=s}^{D} v_d^{targ}(t) - v_d^{ref}(t))^2}$$

While Cepstral Distance (CD) provides an alternative approach based on full-band cepstral coefficients, it often lacks sensitivity to perceptual frequency scales. In contrast, MCD incorporates the mel filter bank, which aligns more closely with human auditory perception, thus offering a more accurate and meaningful assessment of speech quality [13], [15], [16].

2.5 Multimedia Development Life Cycle (MDLC)

The Multimedia Development Life Cycle (MDLC) is a structured software engineering methodology tailored for multimedia applications. Operating in a cyclic and iterative manner, MDLC encompasses systematic stages such as concept development, design, content creation, integration, testing, and deployment [17]. This model ensures that multimedia systems are developed with clear user requirements, consistent asset integration, and quality assurance protocols making it especially relevant for applications involving audio synthesis and game development pipelines.

3 Methodology

3.1 Data Collection

The voice dataset utilized in this study was curated through a targeted extraction of audio segments featuring goblin character vocalizations. The source material consisted of the anime series Goblin Slayer, spanning both Season 1 and Season 2, with a total of 24 episodes.

The audio data underwent a two-step preprocessing workflow to ensure high quality for model training. First, vocal segments were manually isolated from relevant scenes with high precision using Audacity, a widely adopted open-source audio editing tool. This manual process was crucial for precisely selecting the desired goblin voice samples. Second, to address the issue of background noise, an automated process was applied using the MDX-Net method within the UVR5 application to remove any music, dialogue, or environmental noise from the selected audio segments.

The use of 5-second audio segments for training was strategically adopted to manage computational load effectively, particularly due to constraints in available GPU memory. This duration was chosen as an optimal balance and it allows for efficient batch processing and stable model training within our hardware limitations. This approach of splitting audio into smaller segments enables quicker processing and more effective detection of prolonged silence periods, which is crucial for audio preprocessing. Each segmented sample was standardized to a WAV format to ensure consistency across the dataset. The extracted segments were then carefully reviewed to retain only clean, distinct goblin voice samples. This two-step method, combining manual selection with an automated noise removal algorithm, ensures that the training data

is of high quality and distinct, which is essential for optimizing the RVC model. This meticulous approach to data preparation directly contributes to the model's ability to learn and generalize effectively, as the cleanliness and appropriateness of the dataset are key factors influencing successful voice conversion.

3.2 Research Framework

The overall research framework was structured to support the development, training, and evaluation of a voice conversion system tailored for monster character voice generation in games.

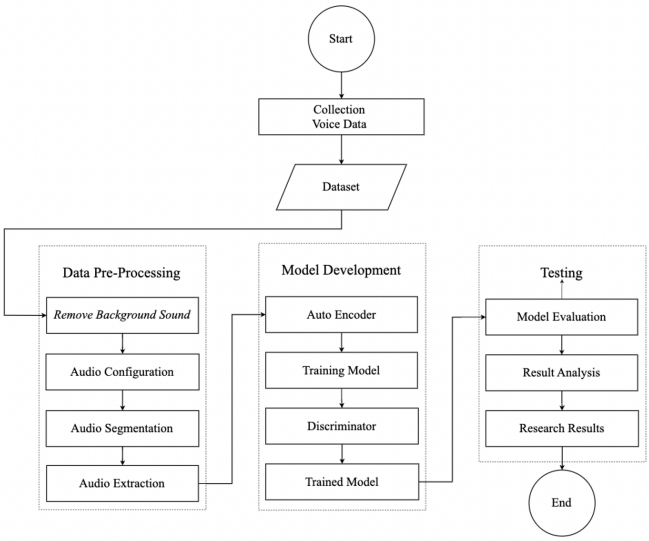


Fig. 1. Software Framework

Model Training in (RVC) models were trained using a curated dataset of goblin voice samples. To investigate the impact of training data volume on model performance, three different data proportion scenarios 60%, 70%, and 80% were tested. This approach allowed for comparative analysis of the models' ability to generalize and synthesize distinct voice outputs under varying resource constraints.

Voice Comparison and Evaluation to assess the quality and effectiveness of the converted voice outputs, a combination of objective and subjective evaluation metrics was employed. MFCC were extracted from both the original and converted audio files to quantify their spectral similarity. These features were further analyzed using Fast Dynamic Time Warping (FastDTW) to compute alignment-based similarity scores across temporal dimensions. MCD was utilized as a perceptual metric to evaluate how closely the converted voices resembled the intended synthetic timbre, with lower values indicating better conversion quality. In addition, the Sound-Similar tool was used to provide a secondary, subjective similarity rating based on psychoacoustic principles. This

combination of metrics offered a comprehensive evaluation framework encompassing both spectral fidelity and perceptual divergence from the source audio.

3.3 Software Design and Implementation

A lightweight multimedia application was developed to facilitate the voice conversion process using the trained RVC models. Designed with flexibility and accessibility in mind, the application supports a wide range of popular audio formats, including MP3, OGG, WAV, and FLAC. It integrates seamlessly with FFmpeg, ensuring compatibility across various input and output pipelines. Additionally, the output formats are optimized for integration with game development platforms such as Unity, allowing developers to embed converted character voices directly into their game environments.

The application workflow is streamlined to accommodate users with varying levels of technical expertise. Users can upload raw or pre-segmented audio files, apply the voice conversion to generate monster-style character voices, and export the resulting audio assets for immediate use in their projects. This tool empowers indie game developers to enrich the auditory experience of their games without the need for professional voice actors or complex audio engineering skills.

4 Results and Discussion

This study investigates the application of RVC for the generation of fictional character voice assets in video games, specifically focusing on monster voice synthesis. To assess the impact of training data volume, three RVC models were trained using voice segment datasets of varying proportions 60%, 70%, and 80% from a total of 190 five-second samples. All models were configured with a target sample rate of 40 kHz and trained over 200 epochs with a batch size of 6, utilizing base_modelG Gf0G40k.pth and base_modelD f0D40k.pth as the foundational checkpoints.

4.1 Evaluation Metrics and Test Setup

To quantitatively evaluate the effectiveness of the voice transformation process, a combination of three key metrics was employed. This approach was chosen to provide a comprehensive evaluation framework that encompasses both spectral fidelity and perceptual transformation quality.

First, MFCC analysis, combined with FastDTW, was used to compute the temporal-spectral distance between the original and converted audio samples, providing insight into structural and acoustic differences. This provided a robust, objective measure of the structural and acoustic differences in the sound signal.

Second, MCD was utilized to generate a perceptual similarity score. While lower values of MCD conventionally indicate higher-quality voice conversion, in the context of this study—where the goal is to create a distinct, non-human voice—a higher MCD value is desirable as it signifies a more successful divergence from the original source audio.

The objective metrics, including MFCC, FastDTW, and MCD, provide a data-driven analysis of the voice conversion. They quantify the structural and acoustic changes from the original voice to the transformed one, offering a precise and repeatable way to measure things like spectral distance and distortion. The value of these metrics lies in their impartiality and consistency, providing scientific rigor to the claims. For this study, where the goal is voice transformation rather than cloning, a higher MCD value is desirable as it indicates a significant and successful divergence from the original source audio.

Lastly, the Sound-Similar tool complements these objective metrics by providing a psychoacoustic assessment that mimics how a human listener would perceive the change. While automated, its value is in measuring the perceptual quality, which is critical for a product like game audio where the goal is to be believable and distinct to a human listener. It confirms that the technical changes validated by the objective metrics are also auditorily effective and meaningful for the intended creative goal.

Therefore, this specific combination was selected to create a holistic evaluation strategy: the objective metrics (MFCC/FastDTW and MCD) quantitatively measure the acoustic signal transformation, while the Sound-Similar tool validates that this transformation is perceptually significant to a human listener, ensuring both technical and creative goals are met.

The testing phase was conducted using newly recorded utterances specifically designed to represent a wide range of vocal characteristics. These utterances were categorized into four distinct speech types to ensure comprehensive model evaluation: Conversational Speech [18] — which reflects natural everyday dialogue, Emotional Expressions [19] — capturing varied affective tones such as anger, joy, and sadness. Phonetically-Rich Sentences [20] — which include diverse phoneme combinations for linguistic variability and Monster Vocalizations — consisting of fictional non-verbal sounds intended to simulate creature-like audio. This diverse dataset enabled a more robust assessment of each model's ability to generalize across different vocal patterns and speech conditions.

This diverse test set ensured a robust evaluation of the models' ability to generalize across speech styles and emotional tones.

4.2 Comparative Analysis

To assess the impact of training data volume on model performance, three different data proportion scenarios—60%, 70%, and 80%—were tested. This specific range was chosen to evaluate the model's ability to balance data sufficiency with generalization. The 60% partition served as a test for limited data and the 80% partition was designed to investigate the risk of overfitting.

Each model was evaluated using three key metrics to assess voice transformation effectiveness. In this context, lower scores from the Sound-Similar tool indicate greater deviation from the original voice, which is advantageous when the goal is to create distinct and fictional character voices. This same logic applies inversely to the other metrics: higher values in both the MFCC distance (measured using FastDTW) and

MCD reflect greater divergence from the original audio, signifying a more successful transformation into non-human or synthetic voice characteristics.

These metrics provide complementary perspectives on the models' performance in voice transformation, as illustrated in Fig 2, 3 and 4 below, where differences in voice characteristics can be observed across the three dataset configurations.

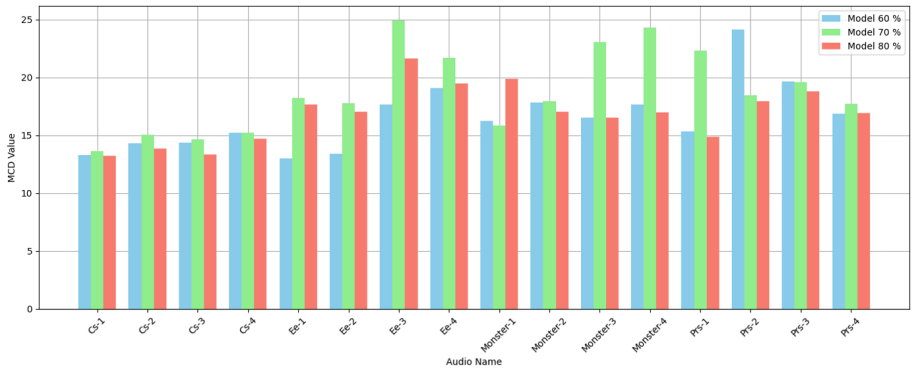


Fig. 2. Comparison Each Audio in MCD

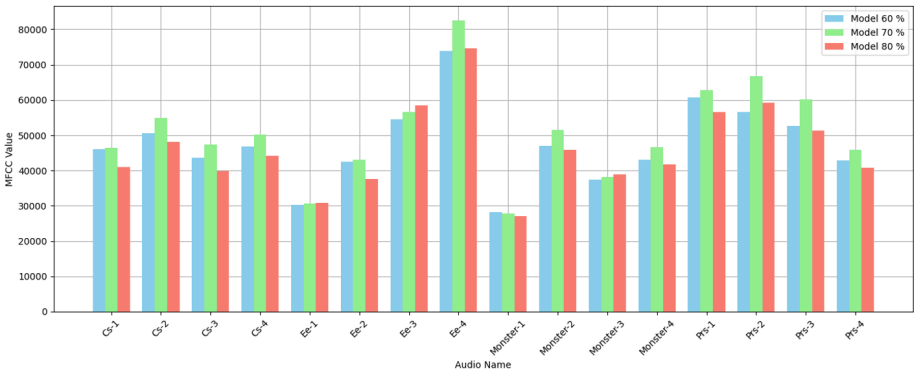


Fig. 3. Comparison Each Audio in FastDTW

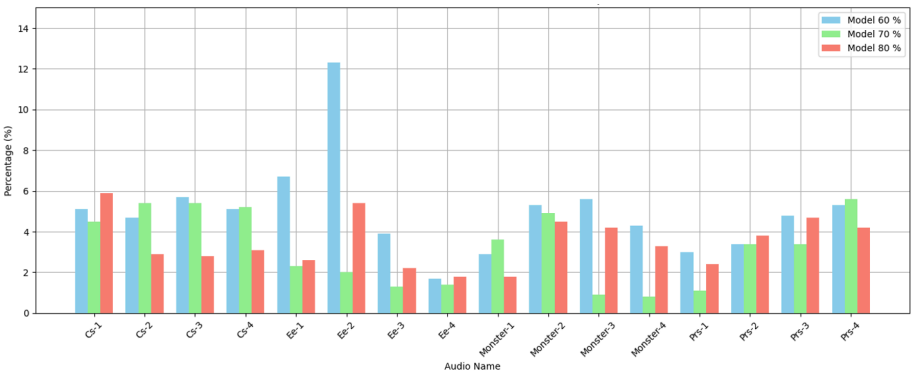


Fig. 4. Comparison Each Audio in Sound-Similar

Table 1. Model Score Value

Speech Type	Model 60%	Model 70%	Model 80%
Conversational Speech	0	10	3
Emotional Expressions	0	10	2
Monster Vocalization	1	7	4
Phonetically-Ruch Speech	2	9	1
Total Score	3	36	10

As presented in Table 1, the 70% model consistently outperformed others across all categories, achieving a total score of 36 out of 48 trials (75%). This suggests that this proportion provides an optimal balance between sufficient training data and generalization capability without overfitting. The 60% model, while the least effective overall, showed better performance in phonetically-rich expressions, suggesting possible data diversity advantages at lower training volumes. The 80% model, while containing more data, exhibited reduced generalizability potentially due to overfitting on the training set.

4.3 Interpretation and Implications

The results of this study highlight several important findings. First, RVC proves to be highly effective in transforming neutral voice samples into diverse and expressive character voices suitable for use in video game environments. This capability enables the creation of rich, immersive audio experiences without the need for professional voice actors. Second, the proportion of training data has a significant impact on model performance. Among the configurations tested, the 70% training dataset offered the best balance between data sufficiency and generalization, resulting in the most consistent and high-quality voice transformations.

Lastly, the use of multiple evaluation metrics Sound-Similar, MFCC with FastDTW, and MCD provided a robust framework for assessing both the perceptual and spectral

fidelity of the converted voices. Together, these findings validate the practical application of RVC for scalable, cost-effective audio asset generation in the gaming industry.

These findings affirm the feasibility of low-resource, AI-driven voice synthesis for game audio production, especially for indie developers who lack access to professional voice actors or high-end computational resources.

5 Conclusion and Future Work

This study successfully demonstrated the implementation of RVC as a practical, scalable, and efficient solution for generating synthetic monster character voices in video game development. By enabling the transformation of original voice samples into expressive, non-human voices using minimal training data and hardware resources, the proposed system addresses critical production challenges commonly faced by indie developers and small studios.

The findings affirm the feasibility of the RVC framework for character voice synthesis. The developed tool supports a range of audio formats (WAV, MP3, OGG, FLAC) and generates output compatible with modern game engines, including Unity. Among the tested configurations, the model trained on 70% of the dataset outperformed others, achieving a top evaluation score of 36 out of 48 tests (75%), indicating a balanced trade-off between training sufficiency and generalization. Nonetheless, the 60% and 80% models showed strengths in specific voice types, suggesting that model selection should be based on project-specific requirements.

This study is limited by the use of a relatively small and domain-specific dataset, focused solely on non-human vocalizations in a single language. Additionally, the voice conversion process was conducted offline, without integration into real-time game environments. Most importantly, the evaluation relied solely on automated metrics and lacked formal human perceptual validation.

Future work may explore several directions to enhance this approach, including:

- Expanding training with multilingual or multi-character datasets, and exploring improvements in model robustness against background noise and silent segments.
- Integrating the RVC engine into real-time game applications, and developing a multi-input batch processing feature to accelerate voice conversion tasks.
- Conducting formal human subjective evaluations, such as Mean Opinion Score (MOS) tests or perceptual listening studies, to quantitatively assess the quality, believability, and immersion factor of the transformed voices from an end-user perspective.
- To enhance practical value for indie developers, the model could be extended to allow users direct control over voice attributes such as pitch, hoarseness, and aggression.
- The voice dataset used in this study was derived from copyrighted material for academic, non-commercial research purposes under fair use provisions. For any commercial applications of this technology, future research and development should utilize licensed or public domain datasets to ensure compliance with intellectual property rights.

Overall, this research contributes to the advancement of inclusive and cost-efficient voice asset generation, aligned with the goals of Sustainable Development Goal 9 by promoting accessible, AI-based digital infrastructure for creative industries.

Acknowledgments. The research/publication of this article was funded by Universitas Sriwijaya 2025. In accordance with the Rector's Decree Number 0028/UN9/LPPM.PT/2025. On September 17, 2025. The author extends heartfelt gratitude to everyone who has supported and contributed to the completion of this research. Special appreciation goes to the supervisor for their invaluable guidance, insight, and knowledge throughout the research process. Sincere thanks are also given to colleagues for their constructive ideas and technical assistance. It is the author's hope that this research will make a meaningful contribution to the advancement of science and technology.

Disclosure of Interests. The voice dataset used for training the RVC models was derived from publicly available audio segments of the anime Goblin Slayer. These audio clips were used solely for academic, non-commercial research purposes under fair use provisions. No redistribution or commercial exploitation of the original or converted audio is intended, and all intellectual property rights remain with the original content owners. For the evaluation phase, additional voice recordings were produced by the authors and research team members specifically for testing purposes. All participants provided informed consent for the use of their voices in the study. These recordings were used exclusively to assess the performance of the voice conversion system and are not publicly distributed.

References

1. J. Sinclair, Principles of Game Audio. Routledge, 2020. [Online]. Available: https://www.routledge.com/Principles-of-Game-Audio-and-Sound-Design-Sound-Design-and-Audio-Implementation-for-Interactive-and-Immersive-Media/Sinclair/p/book/9781138738973
2. N. Bandral and R. Kaur, "A Psychological Inquiry into the Role of Music in Video Games," Language in India, vol. 18, no. 5, p. 297, 2018.
3. S. Horowitz and S. Looney, The Essential Guide to Game Audio. Routledge, 2014. doi: 10.4324/9781315886794.
4. F. Andersen, Danny, C. L. King, and A. A. S. Gunawan, "Audio Influence on Game Atmosphere during Various Game Events," Procedia Comput Sci, vol. 179, no. 2019, pp. 222–231, 2021, doi: 10.1016/j.procs.2021.01.001.
5. S. Jung and H. Kim, "Neural voice cloning with a few low-quality samples," no. NeurIPS, pp. 1–11, 2020, [Online]. Available: http://arxiv.org/abs/2006.06940
6. Z. Ren, "Selection of Optimal Solution for Example and Model of Retrieval Based Voice Conversion," 2024, pp. 468–475. doi: 10.2991/978-94-6463-370-2_48.
7. Onuh Matthew Ijiga, Idoko Peter Idoko, Lawrence Anebi Enyejo, Omachile Akoh, Solomon Ileanaju Ugbane, and Akan Ime Ibokette, "Harmonizing the voices of AI: Exploring generative music models, voice cloning, and voice transfer for creative expression," World Journal of Advanced Engineering Technology and Sciences, vol. 11, no. 1, pp. 372–394, Feb. 2024, doi: 10.30574/wjaets.2024.11.1.0072.
8. T. Farkaš, "Understanding Auditory Space in Digital Games for Visually Impaired People," Acta Ludologica, vol. 7, no. 1, pp. 136–150, 2024, doi: 10.34135/actaludologica.2024-7-1.136-150.

9. E. Rovithis, N. Moustakas, A. Floros, and K. Vogklis, “Audio legends: Investigating sonic interaction in an augmented reality audio game,” Multimodal Technologies and Interaction, vol. 3, no. 4, pp. 1–18, 2019, doi: 10.3390/mti3040073.
10. J. J. Bird and A. Lotfi, “Real-time Detection of AI-Generated Speech for DeepFake Voice Conversion,” no. Ml, Aug. 2023, [Online]. Available: http://arxiv.org/abs/2308.12734
11. A. Kamble, A. Tathe, S. Kumbharkar, A. Bhandare, and A. C. Mitra, “Custom Data Augmentation for low resource ASR using Bark and Retrieval-Based Voice Conversion,” Nov. 2023, [Online]. Available: http://arxiv.org/abs/2311.14836
12. A. Putrasyah, J. Magister, I. Komputer, F. I. Komputer, and U. Sriwijaya, “Classification of Recorded Voice Similarity Using Mel-Frequency,” vol. 6, pp. 959–969, 2023.
13. Z. K. Abdul and A. K. Al-Talabani, “Mel Frequency Cepstral Coefficient and its Applications: A Review,” IEEE Access, vol. 10, no. November, pp. 122136–122158, 2022, doi: 10.1109/ACCESS.2022.3223444.
14. J. Kominek, T. Schultz, and A. W. Black, “Synthesizer Voice Quality of New Languages Calibrated With Mean Mel Cepstral Distortion,” SLTU 2008 - 1st International Workshop on Spoken Languages Technologies for Under-Resourced Languages, pp. 63–68, 2008.
15. R. F. Kubichek, “Mel-Cepstral distance measure for objective speech quality assessment,” IEEE Pac Rim Conf Commun Comput Signal Process, pp. 125–128, 1993, doi: 10.1109/pacrim.1993.407206.
16. A. Vasilijević and D. Petrinović, “Perceptual significance of cepstral distortion measures in digital speech processing,” Automatika, vol. 52, no. 2, pp. 132–146, 2011, doi: 10.1080/00051144.2011.11828412.
17. I. Binanto, Multimedia Digital - Dasar Teori dan Pengembangannya, no. 25. 2010.
18. P. Flipsen, “Measuring the intelligibility of conversational speech in children,” Clin Linguist Phon, vol. 20, no. 4, pp. 303–312, Jun. 2006, doi: 10.1080/02699200400024863.
19. D. Keltner, D. Sauter, J. Tracy, and A. Cowen, “Emotional Expression: Advances in Basic Emotion Theory,” Jun. 15, 2019, Springer New York LLC. doi: 10.1007/s10919-019-00293-3.
20. A. A. Raza, S. Hussain, H. Sarfraz, I. Ullah, and Z. Sarfraz, “Design and development of phonetically rich Urdu speech corpus,” in 2009 Oriental COCOSDA International Conference on Speech Database and Assessments, IEEE, Aug. 2009, pp. 38–43. doi: 10.1109/ICSDA.2009.5278380.

Interpretable Machine Learning for Assessing Winter Outdoor Thermal Comfort: Exploring Built Environment Impacts in a Historic Waterfront Street

Haitao Lian[1], Zhenghui Han[1], Zeyu Ma[1], Yulin Yang[1]

[1] School of Architecture and Art, Hebei University of Engineering, Handan 056038, China
h574021519@gmail.com

Abstract. Urban renewal and climate change have heightened concerns over winter outdoor thermal comfort in historic districts, especially in "hot summer–cold winter" regions where low temperatures suppress public activities. Using Pingjiang Street in Suzhou (a typical hot summer–cold winter region) as a case, this study specifically investigates how built environment factors influence winter Physiological Equivalent Temperature (PET), aiming to fill the research gap of winter outdoor thermal comfort in this climate zone.

A high-resolution ENVI-met model was developed to simulate PET under typical winter conditions. Built environment indicators—BP (Interface Permeability), TAC (Number of Tables and Chairs), GVR (Green view ratio), SVF (Sky View Factor), and WSR (Water Surface Ratio)—were extracted via image semantic segmentation and GIS. A Gradient Boosting Decision Tree (GBDT) combined with SHAP analysis was applied to identify key variables and nonlinear effects.

Results show: (1) PET has an inverted U-shaped relationship with BP, optimal at 40–60%; (2) SVF is the dominant factor, with <40% enhancing and >80% reducing PET; (3) high GVR + high SVF weakens thermal comfort; (4) BP and SVF interact—high BP (≥60%) with low SVF (≤40%) markedly improves PET. Findings highlight the need to coordinate permeability, enclosure, and greenery for balanced visual and thermal performance in winter street design.

Keywords: Built environment; Outdoor thermal comfort; ENVI-met; GBDT; SHAP; Historic district.

1 Introduction

Enhancing the winter climatic adaptability of outdoor spaces in hot summer–cold winter regions is critical for sustainable development: low temperatures, wind chill, and limited solar radiation reduce pedestrian activities, harming residents' social interactions and the cultural value of historic public spaces [1,2]. As a key indicator linking humans, space, and climate, outdoor thermal comfort has attracted growing attention in urban design and policy-making [3].

Thermal comfort is influenced by macroclimate, microclimate, human behavior, and spatial form. Among these, the built environment is the most directly modifiable factor, regulating wind, radiation, and shading. Spatial openness, greenery, water features, and façade permeability affect comfort both physically and psychologically [4,5]. Enclosure can reduce wind exposure, greenery can block drafts,

and water surfaces can buffer temperature fluctuations. These effects are often nonlinear, involving thresholds and trade-offs rather than simple linear trends [6].

However, existing studies have two main limitations. First, most research emphasizes summer heat stress, while winter conditions in hot summer–cold winter regions have received limited attention [7,8]. Second, conventional linear methods, such as regression, struggle to reveal nonlinear interactions among multiple spatial factors, which are critical for design practice [9].

Machine learning offers a way forward. Gradient Boosting Decision Trees (GBDT) can handle high-dimensional nonlinear regression, and explainable AI techniques such as SHAP (Shapley Additive Explanations) enhance interpretability by quantifying feature contributions and interactions [10,11]. This combination allows not only accurate prediction but also transparent insights into how spatial elements shape microclimate and thermal comfort.

This study investigates Pingjiang Street, a historic waterfront neighborhood in Suzhou, China, located in a hot summer–cold winter climate zone. Notably, this study focuses exclusively on winter conditions, as the hot summer–cold winter region has seen abundant research on summer heat stress mitigation but insufficient exploration of winter thermal comfort, which directly restricts the climate adaptability of historic waterfront streets in cold seasons. Using an integrated framework of ENVI-met simulation and interpretable machine learning, five built environment indicators—SVF, GVR, BP, TAC, and WSR—are extracted. A GBDT model predicts Physiological Equivalent Temperature (PET), and SHAP analysis identifies dominant effects, thresholds, and interaction mechanisms. The goal is to clarify how built environment factors influence winter outdoor thermal comfort in historic waterfront streets, and to provide quantitative, explainable guidance for climate-adaptive urban design [12–14].

2 Literature Review

2.1 Outdoor Thermal Comfort

Outdoor thermal comfort reflects the interaction between human perception, microclimate, and spatial form, serving as a key indicator for climate-adaptive urban design [15]. It is influenced by temperature, wind, solar radiation, and human behavior. In winter, particularly in hot summer–cold winter regions, low temperatures and insufficient solar exposure reduce outdoor activity and social participation [16]. Empirical studies indicate that thermal comfort is not only a physical response but also a psychological experience, emphasizing the need to design spaces that mitigate adverse winter conditions [17].

2.2 Influence of Built Environment Factors on the Urban Thermal Environment

The built environment, comprising SVF, GVR, BP, TAC, and WSR, directly shapes local microclimate by regulating wind, radiation, and shading [18,19]. In winter, spatial enclosure and greenery can reduce wind chill, while water surfaces and building geometry influence heat retention. Prior research shows that these factors interact nonlinearly, and combined effects often produce threshold behaviors that cannot be captured by linear models [20,21]. For example, moderate SVF can enhance solar gain, whereas excessively high SVF increases convective heat loss. Similarly, moderate GVR provides wind protection, while excessive vegetation can block sunlight [22,23]. BP exhibits an inverted U-shaped effect: very low permeability restricts sunlight, while very high permeability facilitates cold air intrusion [24].

2.3 Applications of Explainable Machine Learning in Thermal Comfort Studies

Traditional linear methods, such as regression and ANOVA, identify significant relationships but are limited in capturing nonlinear interactions among multiple spatial factors [25]. Machine learning models, particularly GBDT, excel at modeling complex, high-dimensional relationships [26]. Explainable AI techniques, such as SHAP, decompose predictions into feature contributions, enabling identification of dominant factors, interaction effects, and thresholds—critical for evidence-based design interventions [27,28].

Despite advances in thermal comfort research, two key gaps remain in winter-focused studies of historic waterfront streets:

(1) Which built environment variables play a dominant role in influencing outdoor thermal comfort during winter, and can their mechanisms be clearly explained?

(2) Under winter conditions, what are the nonlinear effects and threshold characteristics of built environment factors on outdoor thermal comfort?

3 Methodology

3.1 Study Area

This study focuses on Pingjiang Street Historic District in Suzhou, located in the hot summer–cold winter climatic zone of eastern China. Winter conditions are characterized by low temperatures and pronounced wind chill, limiting pedestrian activity. Located in the hot summer–cold winter zone, Pingjiang Street faces distinct winter challenges: low temperatures (daily average 2–5℃ in January), strong wind chill, and limited solar radiation, which significantly reduce pedestrian activity. In contrast, summer thermal comfort in this region has been well studied in existing literature, so this study focuses on winter to address the research gap. Pingjiang Street features a “parallel waterway and street” layout, continuous street interfaces, diverse spatial scales, and mixed functions, making it a representative high-density historic waterfront street. Variations in BP, SVF, GVR, WSR, and TAC provide an ideal setting for analyzing winter thermal comfort mechanisms.

Fig.1. Study area: (a) Suzhou, Jiangsu; (b) Pingjiang Street Neighborhood; (c) Photos of Pingjiang Street

3.2 Variable Selection and Data Acquisition

Thermal Comfort Simulation and Model Validation. Physiological Equivalent Temperature (PET) was adopted to represent winter street thermal comfort, integrating air temperature, wind speed, relative humidity, and mean radiant temperature under standardized human conditions [32]. ENVI-met 5.7.1 simulated microclimate conditions, capturing interactions among surfaces, atmosphere, vegetation, water bodies, and buildings at 5 m × 5 m × 5 m resolution. Field validation was conducted on January 4, 2025 (08:00–22:00) using temperature–humidity sensors, anemometers, and globe thermometers at 1.5 m height. The RMSE between simulated and measured PET was 1.9 °C, confirming model reliability [33].

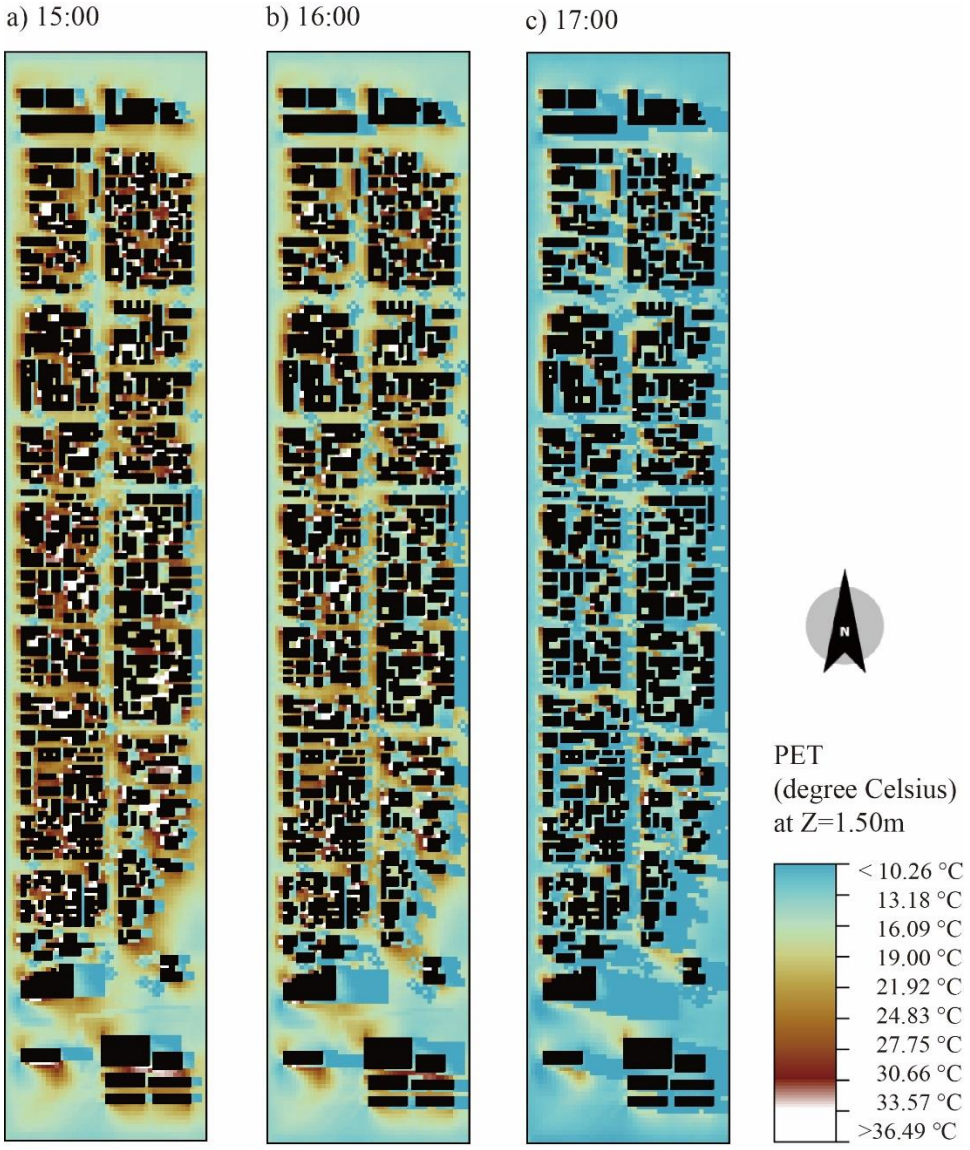


Fig.2.PETs in the Pingjiang Street during winter

Built Environment Factors. Five categories of spatial variables were extracted, with data acquisition methods specified as follows:

Building Interface Permeability (BP). Ratio of transparent façade area (e.g., glass) to total façade area at street level, reflecting visual openness. Data were obtained via semantic segmentation of street-view images to identify transparent materials, combined with field measurements of façade dimensions.

Table and Chair Count (TAC). Public seating (tables and chairs) normalized by street segment length, indicating resting facility provision. Quantified through semantic segmentation of street-view images to count facilities, then standardized by segment length to account for spatial scale differences.

Green View Ratio (GVR). Proportion of vegetation (trees, shrubs, etc.) visible in street-view images. Calculated using semantic segmentation to extract vegetation pixels and compute their ratio to total image pixels, validated by field checks.

Sky View Factor (SVF). Proportion of visible sky from street level, representing spatial openness. Derived from upward-facing street-view images via semantic segmentation to isolate sky pixels, with ratios calculated and cross-checked using GIS spatial analysis.

Water Surface Ratio (WSR). Ratio of water features (rivers, canals) to total street space area, reflecting thermal buffering capacity. Extracted using GIS overlay analysis of hydrological data and satellite imagery, with water boundaries verified through field measurements.

All visual variables (SVF, GVR, BP, TAC) were extracted from Street View images (resolution 2048×1024 px) using a DeepLabV3+ semantic segmentation model trained on the ADE20K dataset, ensuring over 90% accuracy in object classification. Results were cross-validated with manual annotation on 10% of samples. Detailed calculation formulas are summarized in Table 1.

Table 1. Spatial element indicators and their calculation methods

Indicators	Full name	Calculation method
BP	Interface Permeability	$BP = \frac{G_a}{B_j}$ BP, permeability of the bottom interface of the street; Ga, the ratio of the glass area along the street; Bj, facade area
TAC	Number of Tables and Chairs	The number of public tables and chairs identified through street view images in the commercial street.
GVR	Green view ratio	The percentage of greenery detected in images of pedestrian commercial streetscapes.
SVF	Sky View Factor	Upward visible sky area ratio derived from semantic segmentation of street view images, reflecting street space openness.
WSR	Water Surface Ratio	$WSR = \frac{A_w}{A_s}$ Aw, waterbody area; As, total street unit area; Water boundaries extracted using GIS vector overlays and field validation.

3.3 Machine Learning Analysis

Model Selection and Modeling Approach. Gradient Boosting Decision Tree (GBDT) modeled nonlinear relationships between built environment factors (SVF/GVR/BP/TAC/WSR) and PET, while SHAP analysis quantified feature contributions and interactions [31,32].

Data Input and Training. Street segments were divided into homogeneous grid units, and all features were standardized. The dataset was split into training and testing sets at a 5:1 ratio. GBDT parameters (learning rate, tree depth, number of leaves) were optimized via cross-validation, ensuring robust PET prediction.

Model Interpretation and Partial Dependence. SHAP provided global and local insights into variable importance and interactions. Partial dependence plots (PDPs) identified nonlinear inflection points and sensitive intervals, while response surface plots illustrated joint effects of variable pairs, supporting climate-adaptive design strategies [33].

Summary of the Technical Framework. The framework integrates ENVI-met microclimate simulations with perceptual built environment indicators, using GBDT + SHAP + PDP to predict and interpret winter thermal comfort. This approach identifies nonlinear mechanisms, threshold effects, and key pathways of built environment influence, offering actionable guidance for historic street design in cold-season contexts (Fig. 3).

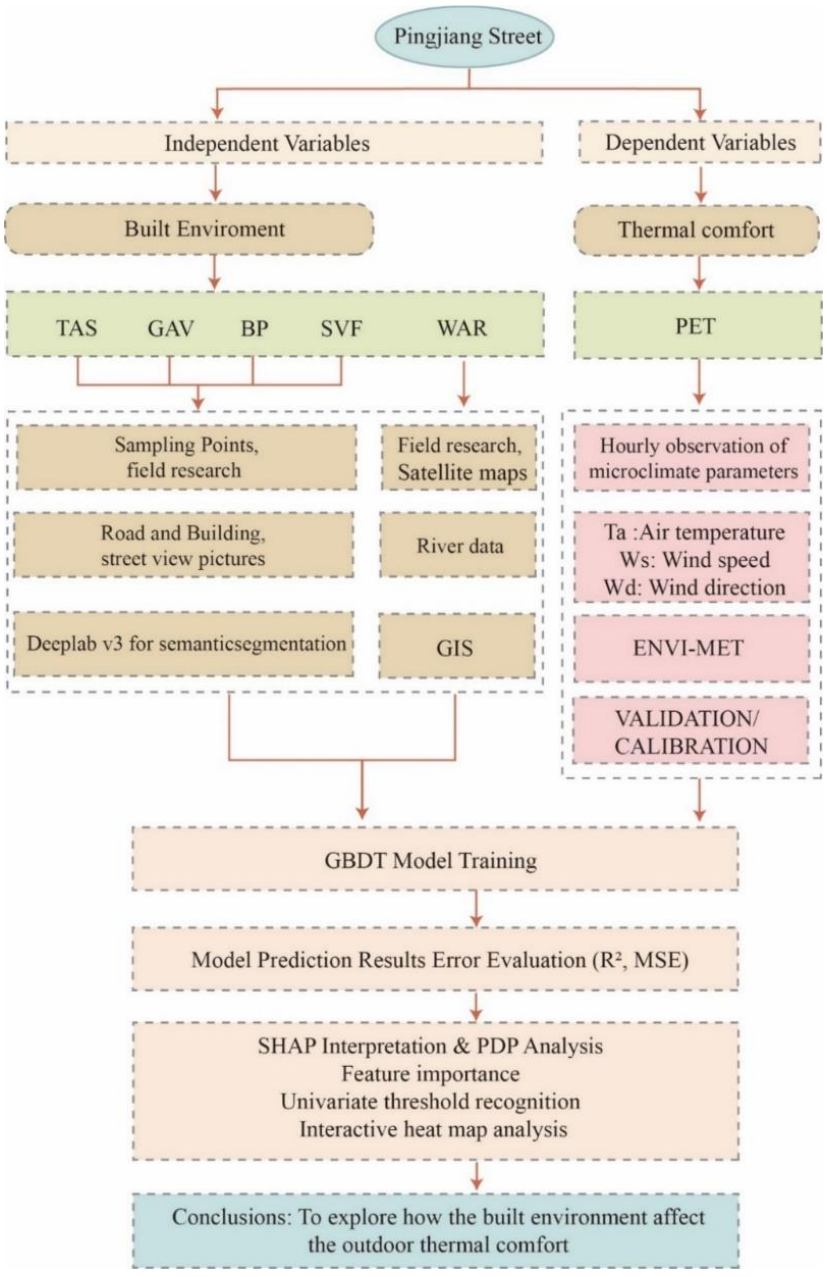


Fig.3. The flowchart of this study.

4 Results

4.1 Variable Importance Ranking

The GBDT model indicates that, among the built environment factors, SVF has the strongest influence on winter PET, followed by GVR and BP, while WSR and TAC contribute relatively less (Fig.4). This highlights that street enclosure and vegetation configuration exert a more pronounced effect on thermal comfort than small-scale street furniture or water features during winter.

SHAP summary analysis confirms this ranking: SVF shows the highest mean absolute contribution (0.43), making it the primary driver of PET variation. GVR and BP follow with contributions of 0.29 and 0.21, respectively, whereas WSR and TAC remain below 0.15, indicating minor influence on thermal conditions.

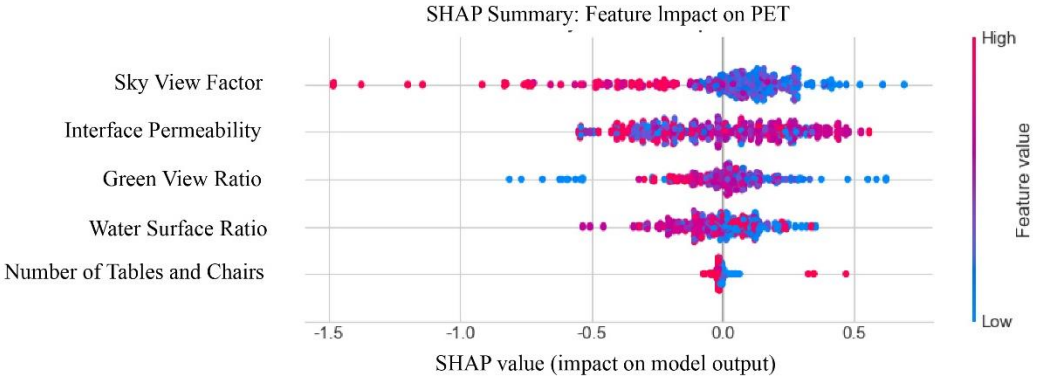


Fig.4.SHAP values for the building environment factors of the Pingjiang street

4.2 Nonlinear Response Characteristics

SHAP dependence plots reveal clear nonlinear response patterns. PET generally decreases with increasing SVF, but the decline steepens beyond approximately 60%. When SVF is below 40%, a more enclosed street form helps block cold winds and retain heat, producing relatively favorable thermal conditions.

GVR exhibits a threshold effect: moderate vegetation coverage (20–40%) improves PET by providing partial wind shielding without excessive shading, while values above 60% reduce PET due to limited solar radiation at pedestrian level. BP demonstrates an inverted U-shaped relationship with PET, with optimal thermal comfort achieved at 40–60% (Fig.6). Extremely low BP limits sunlight penetration, while overly high BP facilitates cold air intrusion, both reducing comfort.

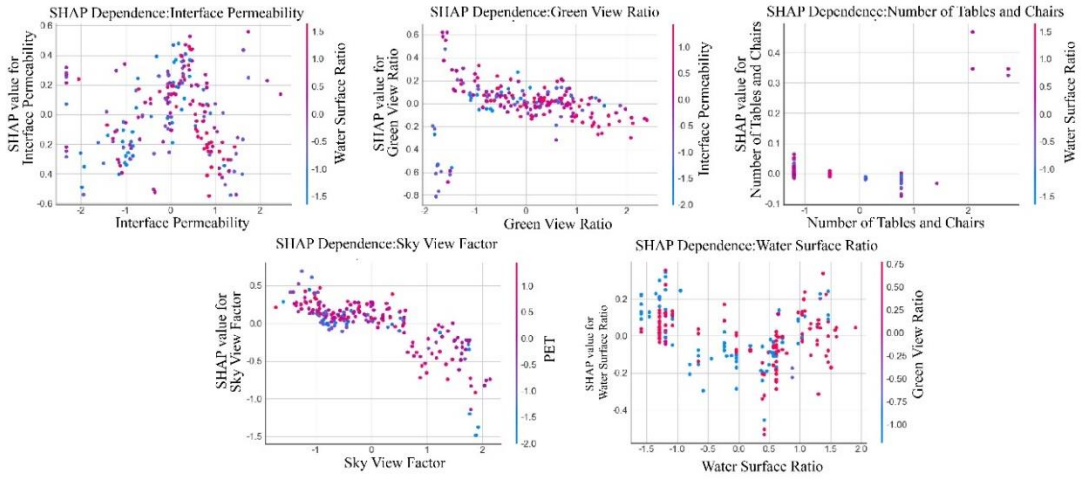


Fig.5. SHAP Dependence Plots for the building environment factors of Pingjiang Street

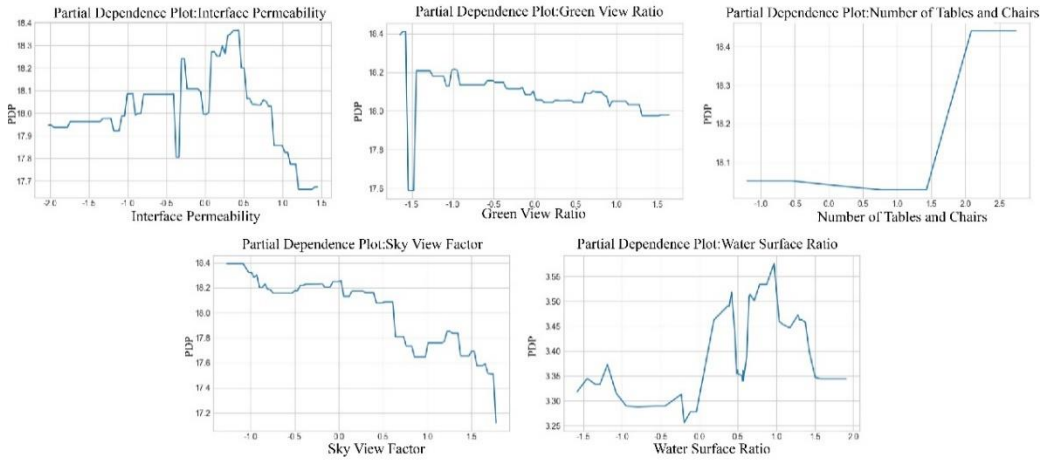


Fig.6. Partial Dependence Plots (PDP) of Building Environment Factors on PET for Pingjiang Street

4.3 Interaction Effects

The SHAP interaction analysis further identifies several notable combined effects among the key variables. A combination of high BP and low SVF is particularly beneficial for winter PET, as adequate enclosure limits wind intrusion while permeable facades maintain visual connectivity and support street vitality. Conversely, when both GVR and SVF are high, PET declines sharply, indicating a compounded cooling effect caused by vegetation shading coupled with excessive openness (Fig.7). For BP and GVR, moderate permeability appears to buffer the negative effects of increased vegetation, but when BP is either too low or too high, higher GVR tends to exacerbate heat loss.

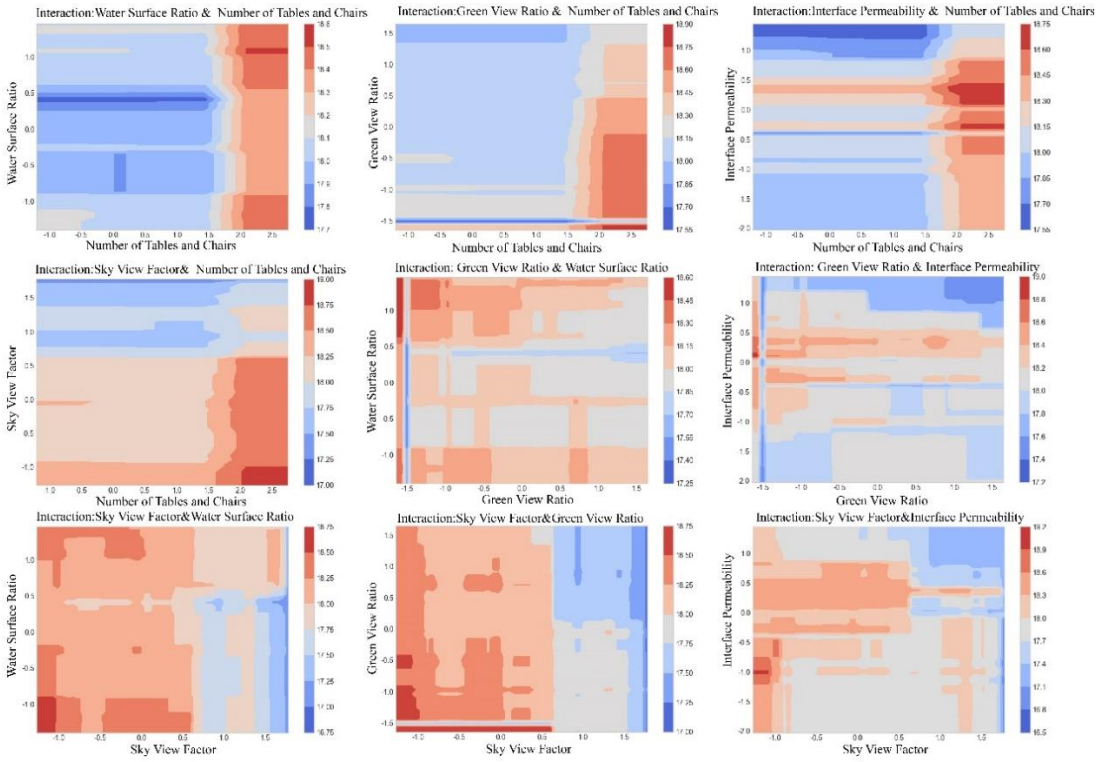


Fig.7. Interaction Plots of Building Environment Factors for Pingjiang Street

4.4 Partial Dependence Analysis

The partial dependence analysis corroborates the SHAP-based findings, further clarifying the presence of threshold effects. Specifically, SVF values exceeding 80% and GVR values above 60% both contribute to substantial reductions in PET during winter. By contrast, the most favorable PET outcomes are observed when SVF ranges between 30% and 50%, GVR between 20% and 40%, and BP between 40% and 60%. These ranges constitute “comfort windows” that can serve as quantitative references for optimizing winter microclimates in historic street environments.

5 Discussion

5.1 Nonlinear Mechanisms of Spatial Factors

For the hot summer–cold winter region, winter outdoor thermal comfort has long been overlooked compared to summer. The nonlinear mechanisms identified in this study (e.g., inverted U-shaped BP-PET relationship) are specific to winter conditions, providing targeted guidance for cold-season design of historic districts in this climate zone. The findings indicate that built environment elements influence winter thermal comfort through distinct nonlinear mechanisms, with clear optimal response ranges. For instance, a moderate Sky View Factor (SVF) improves Physiological Equivalent Temperature (PET), whereas an excessively high SVF can accelerate heat loss and reduce perceived warmth. Similarly, Green View Ratio (GVR) and Building Interface Permeability (BP) exhibit inverted U-shaped or plateau-type

marginal effects. This underscores the need for urban space design to avoid the extreme maximization of single indicators, instead seeking a dynamic balance and coordinated adjustment among scale, form, and functional configuration.

5.2 Spatial Strategies for Climate-Responsive Design

Drawing on SHAP and partial dependence analyses, several targeted strategies for optimizing winter street thermal environments can be proposed. First, street interface enclosure-permeability balance (targeting BP=40%-60%, SVF<40%): For historic facades, retain traditional enclosure (e.g., wooden lattice windows) while replacing non-essential solid panels with low-emissivity glass (transmittance >70%), ensuring cold wind shielding (from low SVF) and visual connectivity (from moderate BP) without damaging heritage features (Fig.7: SVF&BP interaction shows optimal PET at low SVF+high BP).Second, winter-friendly vegetation configuration (targeting GVR=20%-40%): Prioritize deciduous trees (e.g., Ginkgo biloba) with defoliation in winter—their bare branches reduce shading (avoiding GVR>60%'s cooling effect, Fig.5GVR/SHAP) while maintaining windbreak function; avoid evergreen shrubs in south-facing street sections to preserve solar access (Fig.6GVR/PDP shows PET decline at GVR>60%).Third, micro-scale facility layout: Place public seating (TAC) in south-facing zones with SVF=30%-50% (PET peak range, Fig.6SVF/PDP) and adjacent to low GVR (⩽ 40%), balancing sun exposure and wind protection—this aligns with short-term winter activity needs (e.g., 10-15 minute rests) without overcrowding historic street corridors.

5.3 Methodological Value and Research Implications

The GBDT–SHAP framework employed in this study offers notable advantages in exploring the relationships between urban spatial elements and microclimatic responses. It effectively models nonlinear and interactive mechanisms among multiple variables, provides interpretable results that overcome the "black box" limitations of traditional machine learning, supports continuous spatial sampling and variable integration, and is well-suited for high-resolution block-scale analysis. Furthermore, it can be extended to different climate zones, seasons, and multifunctional street contexts for thermal comfort assessments. This work also confirms the feasibility and value of integrating machine learning methods into thermal comfort research, offering a reference pathway for data-driven, climate-adaptive spatial interventions.

For practical application, this study provides three targeted directions:

Policy integration: Propose adding 'winter thermal comfort indicators' (e.g., SVF⩽50%, GVR⩽40%) to the Technical Guidelines for Conservation and Renewal of Historic Districts in hot summer–cold winter regions, requiring microclimate simulation (e.g., ENVI-met) for winter scenarios in renewal project approval.

Urban design optimization: For new-built blocks adjacent to historic districts, adopt 'stepwise SVF control'—gradually increasing SVF from 30% (historic street core) to 50% (peripheral blocks) to avoid abrupt thermal comfort differences (Fig.4 confirms SVF's dominant role in PET).

Heritage conservation balance: For historic waterfront streets (e.g., Pingjiang Road), use removable windbreak facilities (e.g., transparent acrylic screens) in winter—installed along canal-side corridors (WSR zones) to reduce cold wind intrusion (without altering historic water-street layout) and removed in other seasons to preserve landscape authenticity."

5.4 Limitations and Future Directions

This study has several limitations. First, the temporal coverage is restricted, as field validation and ENVI-met simulations were conducted only for a single representative winter day. Although high-frequency

Wi-Fi probe monitoring provided abundant activity records, this short validation period may constrain the robustness of the GBDT training and limit generalizability to other weather conditions such as cloudy, windy, or extremely cold days. Second, the framework primarily relied on physical simulations and objective indices, without incorporating dynamic human behaviors or subjective thermal perceptions. Future studies should therefore extend data collection to multi-day and multi-seasonal periods and integrate field surveys or wearable sensing to enhance comprehensiveness and adaptability.

6 Conclusion

This study reveals four key insights into the relationship between built environment elements and winter thermal comfort in historic districts. First, moderate interface permeability (approximately 40%–60%) produces the highest PET values, as overly low permeability limits insulation benefits, while excessively high permeability allows cold air intrusion and accelerates heat loss (Fig.6). Second, reducing SVF is an effective way to minimize winter heat loss. Streets with SVF below 40% can leverage building enclosures, tree canopies, and pergolas to block cold winds and reduce convective heat loss, whereas SVF above 80% leads to poor heat retention even under abundant sunlight. Third, the combination of high GVR and high SVF can significantly impair thermal comfort, as excessive vegetation shading coupled with openness increases heat loss, creating visually pleasant but thermally cold spaces. Finally, the interaction of high BP and low SVF can substantially enhance winter comfort by maintaining visual continuity and street vitality while reducing wind intrusion and retaining heat. These findings highlight the necessity of integrated spatial optimization that aligns permeability and enclosure, providing practical guidance for climate-adaptive design in traditional urban contexts.

References

1. Bruse, M., Fleer, H.: Simulating surface–plant–air interactions inside urban environments with a three dimensional numerical model. Environmental Modelling & Software 13(3–4), 373–384 (1998)
2. Huttner, S.: Further development and application of the 3D microclimate simulation ENVI-met. PhD Thesis, University of Mainz, Germany (2012)
3. Höppe, P.: The physiological equivalent temperature – A universal index for the biometeorological assessment of the thermal environment. International Journal of Biometeorology 43(2), 71–75 (1999)
4. Matzarakis, A., Mayer, H., Iziomon, M.G.: Applications of a universal thermal index: physiological equivalent temperature. International Journal of Biometeorology 43(2), 76–84 (1999)
5. He, B.-J., Yang, L., Ye, M., Mou, B., Zhou, Y.: Simulation-based evaluation of outdoor thermal comfort in different street layouts under hot summer conditions. Building and Environment 126, 159–172 (2017)
6. Samadpour Shahrak, M., Karimimoshaver, M.: Evaluation of the effect of tree planting pattern on thermal comfort around residential blocks. Motaleate Shahri 12(47), 105–114 (2023)
7. Karimimoshaver, M., Shahrak, M.S.: The effect of height and orientation of buildings on thermal comfort. Sustainable Cities and Society 79, 103720 (2022)
8. Salata, F., Golasi, I., Petitti, D., de Lieto Vollaro, R.: Relating microclimate, human thermal comfort and health during heat waves: an analysis of microclimate change mitigation strategies through ENVI-met simulations. Sustainable Cities and Society 30, 79–91 (2017)
9. Taleghani, M., Berardi, U.: The effect of pavement characteristics on pedestrians' thermal comfort in Toronto. Urban Climate 24, 449–459 (2018)

10. Wang, K., Ozbilen, B.: Synergistic and threshold effects of telework and residential location choice on travel time allocation. Sustainable Cities and Society 63, 102468 (2020)
11. Ki, D., Lee, S.: Analyzing the effects of Green View Index of neighborhood streets on walking time using Google Street View and deep learning. Landscape and Urban Planning 205, 103920 (2021)
12. Li, X., Zhou, W., Ouyang, Z.: Forty years of urban expansion in Beijing: What is the relative importance of physical, socioeconomic, and neighborhood factors? Applied Geography 38, 1–10 (2013)
13. Ma, J., Dong, G., Chen, Y., Zhang, W.: Does satisfactory neighbourhood environment lead to a satisfying life? An investigation of the association between neighbourhood environment and life satisfaction in Beijing. Cities 74, 229–239 (2018)
14. Zheng, Y., Ye, R., Hong, X., Tao, Y., Li, Z.: What factors revitalize the street vitality of old cities? A case study in Nanjing, China. ISPRS International Journal of Geo-Information 13(8), 282 (2024)
15. Yung, E.H.K., Wang, S., Chau, C.K.: Thermal perceptions of the elderly, use patterns and satisfaction with open space. Landscape and Urban Planning 185, 44–60 (2019)
16. Li, Y., Yabuki, N., Fukuda, T.: Exploring the association between street built environment and street vitality using deep learning methods. Sustainable Cities and Society 79, 103656 (2022)
17. Kim, Y., Brown, R.D.: Climate-sensitive street design: Evaluating summer pedestrian activity and behavioral thermal adaptation on the High Line, NYC. Building and Environment 113203 (2025)
18. Cheng, C.Y., Lin, T.P.: Decision tree analysis of thermal comfort in the courtyard of a senior residence in hot and humid climate. Sustainable Cities and Society 101, 105165 (2024)
19. Chen, S., Wang, X., Lun, I., Chen, Y., Wu, J., Ge, J.: Effect of inhabitant behavioral responses on adaptive thermal comfort under hot summer and cold winter climate in China. Building and Environment 168, 106492 (2020)
20. Jin, H., Liu, S., Kang, J.: Thermal comfort range and influence factor of urban pedestrian streets in severe cold regions. Energy and Buildings 198, 197–206 (2019)
21. Ma, X., Chau, C.K., Lu, S., Leung, T.M., Li, H.: Modelling the effects of neighbourhood and street geometry on pedestrian thermal comfort in Hong Kong. Architectural Science Review, 1–16 (2024)
22. Lin, J., Chen, S., Yang, J., Li, Z.: Research on summer outdoor thermal comfort based on COMFA model in an urban park of Fuzhou, China. Theoretical and Applied Climatology 155(3) (2024)
23. Sun, C., Lian, W., Liu, L., Dong, Q., Han, Y.: The impact of street geometry on outdoor thermal comfort within three different urban forms in severe cold region of China. Building and Environment 222, 109342 (2022)
24. Abdallah, A.S.H., Mahmoud, R.M.A.: Urban morphology as an adaptation strategy to improve outdoor thermal comfort in urban residential community of New Assiut City, Egypt. Sustainable Cities and Society 78, 103648 (2022)
25. Altunkasa, C., Uslu, C.: Use of outdoor microclimate simulation maps for a planting design to improve thermal comfort. Sustainable Cities and Society 57, 102137 (2020)
26. Khaire, J.D., Madrigal, L.O., Lanzarote, B.S.: Outdoor thermal comfort in built environment: A review of studies in India. Energy and Buildings 303, 113758 (2024)
27. Huang, X., Ma, X., Zhang, Q.: Effect of building interface form on thermal comfort in gymnasiums in hot and humid climates. Frontiers of Architectural Research 8(1), 32–43 (2019)
28. Deng, X., Nie, W., Li, X., Wu, J., Yin, Z., Han, J., Lam, C.K.C.: Influence of built environment on outdoor thermal comfort: A comparative study of new and old urban blocks in Guangzhou. Building and Environment 234, 110133 (2023)

29. Ladi, T., Jabalameli, S., Sharifi, A.: Applications of machine learning and deep learning methods for climate change mitigation and adaptation. Environment and Planning B: Urban Analytics and City Science 49(4), 1314–1330 (2022)
30. Yang, X., Li, H., Ma, X., Zhang, B.: Research on the coupling relationship between street built environment and thermal comfort based on deep learning of street view images: A case study of Chaowai Block in Beijing. Buildings 15(9), 1449 (2025)
31. Zhang, J., Li, X., Lian, H., Li, H., Zhang, J.: Day–night synergy between built environment and thermal comfort and its impact on pedestrian street vitality: Beijing–Chengdu comparison. Buildings 15(12), 2118 (2025)
32. Nakano, J., Tanabe, S.I.: Thermal adaptation and comfort zones in urban semi-outdoor environments. Frontiers in Built Environment 6, 34 (2020)
33. Berezsky, O., Kovalchuk, O., Berezka, K., Ivanytskyy, R.: Assessing smart cities' effectiveness: Machine learning approaches. Frontiers in Sustainable Cities 7, 1400917 (2025)

EduGame: Making Math Magical with 3D Learning Adventures

Shahirah Mohamed Hatim[1[0000-0001-7399-1325]], Haryani Haron[2[0000-0002-0550-8351]], Muhammad Daniel Aiman Mohd Rozli[2]

[1] Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Perak Branch Tapah Campus, 35400 Tapah Road, Perak, Malaysia
[2] Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Malaysia
shahirah88@uitm.edu.my

Abstract. This study presents the design, development and evaluation of *EduGame*, a three-dimensional (3D) educational game aimed at enhancing mathematics learning among Malaysian primary school students. Developed using the ADDIE instructional design model, *EduGame* integrates curriculum-aligned content with interactive gameplay to address common challenges in mathematics education, including negative student perceptions, limited stakeholder support, and disengagement with traditional teaching methods. The game was built on the Sandbox Game Maker platform and features intuitive controls, tutorials, achievement systems, and interactive question-and-answer mechanics. Usability testing conducted with students, teachers, and parents yielded an average System Usability Scale (SUS) score of 77.78, indicating good usability and positive user experience. Observations revealed increased student motivation and engagement, particularly among learners with low initial interest in mathematics. Limitations such as repetitive gameplay and platform constraints were identified, leading to recommendations for future development, including the use of advanced game engines, expanded content coverage, and improved navigation support. *EduGame* demonstrates the potential of game-based learning to transform mathematics education in early childhood settings.

Keywords: Primary education, visual leaning, 3D educational game.

1 Introduction

1.1 Background of the Study

The integration of three-dimensional (3D) educational games into early childhood education is based on the understanding that computer games can serve as effective teaching and learning tools. According to [1], computer games play a significant role in influencing the behavior of the students and fostering a positive learning environment. This underscores the need for educational games that are not only informative but also engaging and visually stimulating for young learners. It is essential to consider the learning preferences and cognitive styles of kindergarten children during game

development. The use of familiar, colorful models with animations is particularly important for capturing and maintaining their attention. Furthermore, incorporating multimedia elements such as music and interactive activities can significantly enhance the learning experience.

In Malaysia, the incorporation of games into education is already being practiced across various subjects, including history. Traditional methods of teaching history have often been criticized for low levels of student engagement and limited appeal. The use of 3D game technology has emerged as an effective approach to help school children better understand and appreciate historical topics, particularly Malay history. According to [2], the primary objective of such initiatives is to create an engaging and interactive learning environment that revitalizes the teaching and learning of history. This approach addresses several challenges, such as the lengthy and text-heavy nature of history textbooks, which can discourage students, especially those who are not inclined towards reading. As a result, many students struggle to retain historical information and gradually lose interest in the subject. Moreover, students often perceive history learning as requiring extensive reading and passive listening to lectures, both of which they find challenging and unengaging.

Beyond the local context, education in the gaming industry is dynamic, interdisciplinary, and continuously evolving, reflecting the complexity and innovative nature of one of the most progressive sectors in the modern era. A wide range of educational opportunities is now available to aspiring game designers, developers, researchers, and enthusiasts. These opportunities include formal degree programs, online courses, workshops, and collaborations with industry partners. By fostering a culture of creativity, collaboration, and lifelong learning, education in game development empowers individuals to make meaningful contributions to the field, explore new creative frontiers, and influence future direction of interactive entertainment on a global scale [3][4].

The persistent reliance on traditional teaching methods within primary education constitutes a significant concern requiring urgent intervention. While such methods have historically been regarded as foundational, their capacity to address the diverse needs of contemporary learners is increasingly questioned. This pedagogical rigidity may impede children's cognitive development and hinder their ability to adapt and thrive in a rapidly evolving global environment [5]. Consequently, it is imperative to critically examine and address the limitations of conventional instructional strategies in primary education, with the aim of fostering a learning environment conducive to nurturing lifelong learners equipped with the competencies necessary to navigate the complexities of the twenty-first century. In the context of mathematics education in Malaysia, three key challenges have been identified. Specifically, traditional teaching methods appear increasingly ineffective, as students' widespread access to advanced technologies such as smartphones and tablets diminishes the effectiveness of conventional instructional approaches. Many students develop negative early perceptions of mathematics, which can hinder their long-term engagement and performance in the subject. In addition, the effectiveness of mathematics education is often constrained by insufficient support from key stakeholders, including educators, parents, and policy makers. Thus, many students face learning challenges arising from the educational environment, which can significantly affect their academic performance and motivation [6][7].

Although mathematics differs from history in subject matter, the narrative style drawn from historical storytelling was intentionally employed to increase learner engagement and contextualize mathematical challenges. Story-driven tasks provide a meaningful backdrop that encourages problem-solving and persistence, aligning with constructivist principles of learning. This study aims to develop an educational game named *EduGame* that integrates instructional content with interactive gameplay. The primary objective is to positively influence the perceptions of students in mathematics by presenting mathematical concepts in an engaging and accessible manner. The game will incorporate educational features designed to support learning, while employing intuitive controls to ensure ease of use and a comfortable gaming experience for students.

2 Related Works

Gaming in education refers to the integration of video games as a tool to support student learning. This approach introduces enjoyment and interactivity into the educational process. Instead of relying solely on traditional methods such as textbooks and lectures, students can engage with educational content through games, enabling them to solve problems, explore new concepts, and receive immediate feedback. These games also help foster essential skills such as creativity, collaboration, and technological literacy which are the competencies valuable for their future. The core gameplay loop of the game guides players through a sequence of missions in which they solve progressively more complex mathematics puzzles to unlock tools and explore new environments. Each mechanic which are timed problem-solving, adaptive hints and immediate feedback, is intentionally linked to specific learning outcomes such as number sense, algebraic reasoning and logical deduction, with scoring that rewards accuracy and strategy rather than speed to strengthen conceptual understanding. Overall, incorporating gaming into the classroom is transforming both teaching and learning, making the experience more dynamic and engaging for students and educators alike.

There are several gaming applications developed by previous researchers which assist the students to learn mathematics better. Several educational games are available across different platforms, each offering unique learning experiences. Unlike open-ended sandbox platforms such as Roblox or Minecraft, EduGame offers a structured learning pathway where objectives, constraints, and feedback are explicitly aligned to curricular standards. While sandbox games support creativity and social interaction, EduGame integrates guided challenges and teacher dashboards that track mathematical skill acquisition, offering stronger support for formal assessment and classroom use. In a similar effort to provide a structured, curriculum-aligned experience, DragonBox Numbers is a math-focused game for young children, available on mobile devices and tablets, and it uses a top-down perspective to engage learners in foundational number concepts [8]. Zombinis combines puzzle and adventure gameplay to enhance spatial-temporal reasoning, using a third-person view on personal computers, consoles and mobile devices [9]. ST Math offers a math learning experience through bingo-style activities, also using a top-down perspective on mobile and tablet platforms [10]. Prodigy

is an online math platform with role-playing game (RPG) elements, using a third-person perspective on both web and mobile [11][12][13]. Lastly, Splash Math delivers interactive math lessons through a top-down perspective, available on mobile devices and tablets [14]. These games collectively highlight the diverse ways educational content can be delivered to students.

While these educational applications offer interactive and engaging learning experiences, they also present certain limitations. DragonBox Numbers, ST Math, and Splash Math primarily focus on basic mathematical concepts, which may not sufficiently challenge older or more advanced learners. Their top-down perspectives and simple gameplay mechanics could limit engagement for students seeking more immersive experiences. Zombinis, though effective for developing reasoning skills, may lack direct curriculum alignment and modern graphics, potentially reducing its appeal to today's learners. Prodigy, while offering an RPG-based approach, has been criticized for emphasizing in-game rewards over actual learning and may encourage gameplay for progression rather than conceptual understanding. A notable limitation of Splash Math is its reliance on repetitive gameplay mechanics. The platform focus on drill-based exercises may initially support skill reinforcement; however, over time, the lack of variation in task design can lead to reduced student engagement, particularly among learners who require more cognitively stimulating or advanced challenges. In addition to the repetitive nature of activities, Splash Math tends to prioritize procedural fluency over the development of deeper conceptual understanding. The design of the game emphasizes practice in basic mathematical operations without fostering higher-order thinking or problem-solving strategies. Consequently, while students may exhibit improvements in surface-level skills, their ability to apply mathematical concepts in unfamiliar contexts may remain underdeveloped. Additionally, most of these applications depend on device availability and stable internet connections, which can be a barrier in less technologically equipped environments [8-14]. Table 1 provides the summary of all mathematical based game applications.

Table 1. The comparison of available mathematical-based game applications

Game Title	Description	Type of Gameplay	Perspective
DragonBox Numbers	Educational math game for young children	Educational	Top-Down
Zombinis	Puzzle with adventure game spatial temporal reasoning	Puzzle Adventure	Third Person
ST Math	Educational game with a bingo	Educational	Top-Down
Prodigy	Online math platform with a role-playing theme	RPG	Third Person
Splash Math	Educational math app with interactive lessons	Educational	Top-Down

3 System Development and Implementation

The development and implementation of *EduGame* were systematically conducted using the Analysis, Design, Development, Implementation, and Evaluation (ADDIE) instructional design model, ensuring a structured and iterative approach throughout the research lifecycle [15]. Figure 1 shows the ADDIE model cycles.

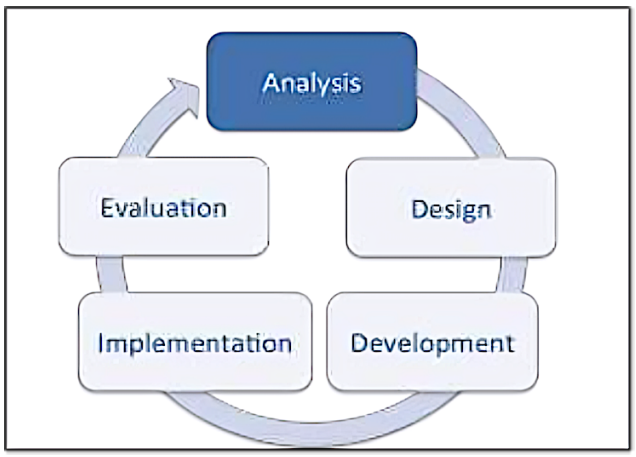


Fig. 1. ADDIE model cycles [15][17]

The analysis phase of the *EduGame* focused on identifying educational needs, user requirements, and system specifications necessary for developing an effective 3D educational game. A comprehensive review of existing literature on educational games, particularly those targeting primary school mathematics, highlighted key limitations in current teaching approaches, such as student disengagement and negative perceptions of mathematics. User personas representing teachers, students, and parents were developed to guide the identification of functional and non-functional system requirements. Functional requirements included essential gameplay features such as tutorials, achievement systems, and interactive question-and-answer mechanics, while non-functional requirements emphasized usability, performance stability, and accessibility across standard computing devices. Stakeholder expectations were carefully considered to ensure that the game design addressed the specific needs of Malaysian primary school students. Additionally, potential limitations related to technological infrastructure, such as varying levels of digital literacy and inconsistent internet access, were acknowledged during this phase. The outcomes of the analysis phase provided a clear foundation for subsequent design and development activities.

In the design phase, the focus shifted towards translating analytical insights into concrete system structures and interaction models. Low-fidelity prototypes, including wireframes and flowcharts, were developed to outline the user interface and in-game navigation pathways. Storyboards were created to visualize gameplay progression and learning objectives, while use case diagrams defined user-system interactions, encompassing key actions such as starting the game, accessing tutorials, and completing

mathematics challenges. Special attention was given to ensuring that the interface was intuitive and accessible to young learners. Visual elements including icons and objective markers were incorporated to guide players through tasks with minimal cognitive load. The game mechanics were designed to balance educational content with interactive engagement, incorporating elements such as point systems and feedback loops to motivate learners. The selection of Sandbox Game Maker as the development tool informed technical design choices, ensuring alignment with the platform's capabilities. Overall, the design phase produced a detailed blueprint of the structure, gameplay flow and user interactions of the *EduGame*, ensuring readiness for the development phase. Figure 2 shows the storyboard for the proposed *EduGame*.

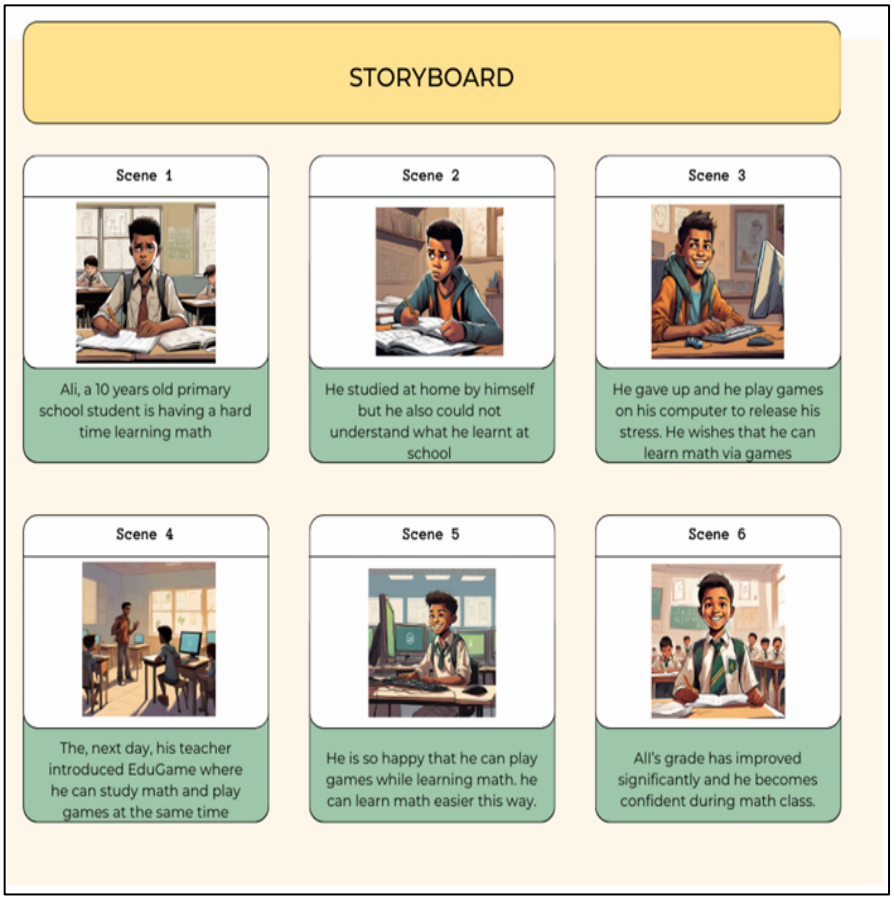


Fig. 2. The storyboard of *EduGame* (OpenAI, 2025)

Figure 2 explains on *EduGame* environment concept art. In the development phase, significant emphasis was placed on constructing a user-friendly and pedagogically sound 3D educational game using the Sandbox Game Maker platform. This phase involved the creation of detailed user personas, storyboards and flowcharts, ensuring it

addressed the specific learning needs of primary school students. Interactive educational content, such as mathematics challenges, and clear navigation paths were integrated to enhance engagement and learning efficacy. Asset placement, environment design, and the development of in-game logic were carefully managed to create a cohesive and immersive experience [2][16][17]. Figure 3 illustrates the prototype for *EduGame*.

Interface	Description
	This is the first page of the Interface where the user will see. On the top left, there are username and player's health points. The user needs to interact by choosing the dialogue option in order to proceed.
	After choosing the dialogue option, the game will be started. The objective will be available for the user to see.
	The instructions will be given to the user when the game starts running.
	This is the part where the user will be given a tutorial or a guide in order to complete the game. Complete the dialogue option in order to proceed.

	The user needs to answer the mathematics question in order to proceed. A few options/answers will be provided by the game. The user needs to choose the correct answer to proceed.

Fig. 3. Prototype for *EduGame*

Upon completion of development, *EduGame* was implemented in controlled classroom environments and individual home settings to assess its operational effectiveness and user experience. The implementation phase aimed to evaluate the accessibility, usability, and educational value of the game. Clear instructions and in-game tutorials were provided to facilitate independent use by the target audience of primary school students. Two trained researchers conducted structured observations during classroom sessions using a predefined checklist that captured engagement, collaboration, and time-on-task. Notes and time-stamped screenshots were recorded unobtrusively, and feedback from teachers and students was collected immediately afterward through brief semi-structured interviews. System usability testing using the System Usability Scale (SUS), was conducted with a sample group of students, teachers, and parents to gather quantitative and qualitative feedback regarding user satisfaction, interface clarity, and overall gameplay experience. Observations from this phase highlighted both strengths and areas requiring improvement, such as enhancing game variety and optimizing user interaction. Data collected during implementation informed iterative refinements to the design of the game and its functionality. The implementation phase demonstrated that *EduGame* could be effectively used both within formal classroom settings and at home, offering a flexible and engaging tool to support mathematics learning.

4 Result and Discussion

During the implementation phase, the game was deployed in both classroom and home environments. Preliminary testing confirmed that students could interact with the game without technical issues, and the inclusion of tutorials and objective markers facilitated initial engagement. However, observations revealed that while most users navigated the game effectively, some students experienced initial confusion, particularly with task progression in certain stages. A formal assessment was conducted using the System Usability Scale (SUS) with participants involving students, teachers and parents, who each completed a standardized SUS questionnaire after interacting with the game to evaluate overall usability. Each session lasted 45 minutes over a two-week period. Pre- and post-tests measured mathematical understanding, and in-game analytics recorded completion rates and error patterns. Notes and time-stamped screenshots were taken unobtrusively. Feedback from teachers and students was collected immediately

afterward through brief semi-structured interviews. Figure 4 illustrates the SUS scores recorded from each participant.

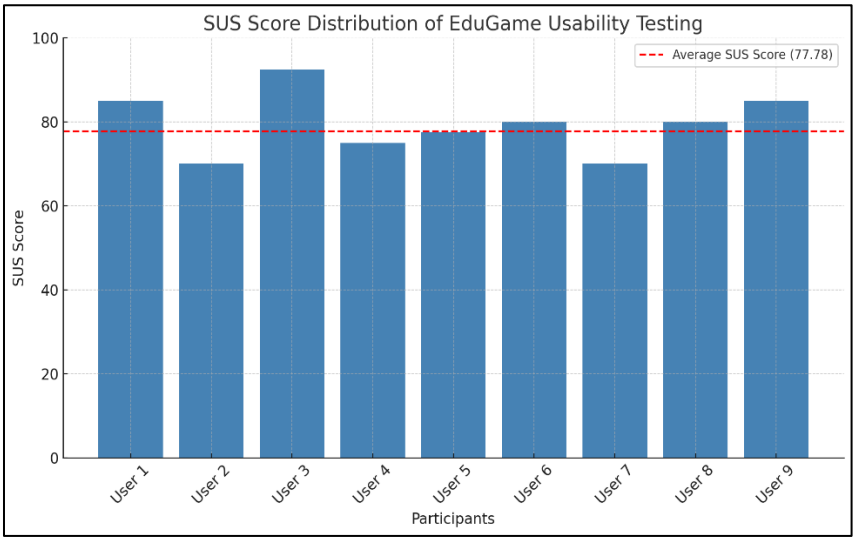


Fig. 4. SUS score distribution of *EduGame* usability testing

Figure 4 above shows the average SUS score of 77.78, places *EduGame* within the "Good" usability category. Participants generally found the system easy to use, though certain responses indicated areas where guidance and navigation could be improved. The highest score (92.5) suggests a highly positive user experience for some, while the lower scores (70) reflect variability in user interaction fluency, possibly influenced by prior experience with digital games.

Observations during the implementation phase revealed that students were highly engaged with the interactive mechanics of *EduGame*, with most showing motivation to complete tasks to earn points and rewards. Teachers reported that students who typically displayed low enthusiasm for mathematics exhibited increased interest during game sessions. Parental feedback supported these findings, noting that students voluntarily interacted with the game outside of classroom sessions.

However, repetitive question structures and task loops were identified as potential contributors to reduced engagement over extended play sessions. This feedback aligns with existing literature emphasizing the need for varied challenges within educational games to sustain long-term interest. Qualitative observations suggest that *EduGame* fostered a more positive perception of mathematics among its target users.

5 Conclusion and Future Work

Overall, the results indicate that *EduGame* achieved its primary goal of providing an accessible, engaging educational tool for mathematics learning. The SUS evaluation

validates the general usability of the system, though iterative improvements are required to address navigation challenges and reduce gameplay monotony. Game Maker proved advantageous for rapid prototyping and multi-platform deployment, with its drag-and-drop interface significantly reducing development time while built-in scripting supported the creation of custom math-puzzle mechanics. However, scalability for large multiplayer environments and advanced 3-D rendering remains limited, suggesting that future expansions could benefit from more robust engines such as Unity or Godot. Furthermore, expanding content coverage to encompass a wider range of mathematics topics and adaptive difficulty levels will enhance the educational value and applicability across various primary education stages.

In conclusion, *EduGame* achieved its core objective of creating an accessible and engaging educational tool tailored to the learning needs of primary school students. Nevertheless, further enhancements are recommended, such as diversifying educational content, improving navigation support, and leveraging more advanced game development engines to enrich the gameplay experience. This research contributes to the growing body of evidence supporting the integration of educational games into formal learning environments, offering a promising approach to addressing challenges in primary mathematics education. In order to enhance the effectiveness and sustainability of *EduGame*, several key recommendations are proposed. First, the content of game can be expanded to include a broader range of mathematics topics and adaptive difficulty levels to accommodate diverse learning abilities. Second, varied gameplay elements, such as mini-games and narrative-based challenges, can be integrated to reduce repetition and maintain long-term user engagement. Third, improvements to navigation support, including clearer tutorials, visual cues, and simplified interfaces, are necessary to assist novice users. The future development should consider migrating to more advanced game engines, such as Unity or Unreal Engine, to enable better graphics, flexible content design, and multi-platform deployment.

Acknowledgments. The authors wish to thank Universiti Teknologi MARA, Malaysia for funding this research.

Disclosure of Interests. The authors hereby declare that there are no known financial, commercial, or personal relationships that could be construed as potential competing interests in the development, implementation, or reporting of this study. All aspects of the research, including data collection, analysis, and interpretation, were conducted independently and without influence from any external parties or organizations. This study was undertaken solely for academic purposes, with the intention of contributing to the field of educational technology and supporting future research.

References

1. Eridani, D.: The Development of 3D Educational Game to Maximize Children's Memory. 1st International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE) (2014).

2. Mahamarowi, N., H., Mustapha, S.: Game-Based Approach to Learning Ancient Malay Heritage's History. 14th Control and System Graduate Research Colloquium (ICSGRC), IEEE, (2023).
3. Lin, C., Yang, H., Y.: Research on the Value of Physical Education in School to the Physical & Mental Health Development of Teenagers. 2020 International Conference on Modern Education and Information Management (ICMEIM), pp. 722-725, (2020).
4. Ayob, A., Hussain, A., Majid, R. A.: A review of research on creative teachers in higher education. International 8–14, (2020).
5. Pollard, V., Hains-Wesson, R., Young, K.: Creative teaching in STEM. Teaching in Higher Education, 23(2), 178-193, (2018).
6. Rocha, M., Dondio, P.: Effects of a videogame in math performance and anxiety in primary school. Int. J. Serious Games, 8, 45-70, (2021).
7. Lawson, M., Hodge, L., L.: Blurring the boundaries between home and school: Supporting parent and student learning with family math. In M. B. Wood, E. E. Turner, M. Civil, & J. A. Eli (Eds.), Proceedings of the 38th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education (p. 1131). Tucson, AZ: The University of Arizona (2016).
8. Gibbs, P., S.: Game Based Learning: The Effects of DragonBox 12+ on Algebraic Performance of Middle School Students. PhD Thesis.
9. Hui, C., S.: Learning mathematics through computer games. National Institute of Education, Nanyang Technological University (2014).
10. Wendt, S., Rice, J., Nakamoto, J.: Evaluation of the MIND Research Institute's Spatial-Temporal Math (ST Math) Program in California (2014).
11. Haliza, R., V., Lestari, I., Usman, H., Sarifah, I.: The Effect of Prodigy Math Learning Media Based on Digital Based Learning on Student Learning Outcomes in Grade IV Mathematics Subjects. Proceedings of the International Conference on Education Practice (ICEP), Advances in Social Science, Education and Humanities Research 906 (2024).
12. Bledsaw, J., A.: Investigating Prodigy Math Program to Improve Students' Investigating Prodigy Math Program to Improve Students' Success in Mathematics. Master Theses (2024).
13. Sjöberg, J., Brooks, E.: Collaborative interactions in problem-solving activities: School children's orientations while developing digital game designs using smart mobile technology. International Journal of Child-Computer Interaction, Volume 33, 100456, Elsevier (2022).
14. Zhang, M., Trussell, R., P., Gallegos, B., Asam, R., R.: Using Math Apps for Improving Student Learning: An Exploratory Study in an Inclusive Fourth Grade Classroom. TechTrends, Volume 59, Number 2 (2015).
15. Kurt, S.: ADDIE model: Instructional design. (2018).
16. Candiasa, I., M.: Application of Instructional Design Models by Prospective Teacher Students. Jurnal Pendidikan dan Pengajaran, Volume 55, Nomor 3, pp. 640-652 (2022).
17. Thabran, Y., Ali, R., D., M.: Reconstructing Syllabus and Teaching Materials for Creative Writing Class. (2019).

Redefining Campus Orientation: A Web-Based Virtual Tour Experience for Students

Haryani Haron[1[0000-0002-0550-8351]], Shahirah Mohamed Hatim[2[0000-0001-7399-1325]] and Allysha Zull Hizam[1]

[1] Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Malaysia
[2] Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Perak Branch Tapah Campus, 35400 Tapah Road, Perak, Malaysia
harya265@uitm.edu.my

Abstract. Navigating large university campuses can be challenging for new students, often leading to disorientation and reduced engagement. This study presents *LensaKPPIM*, a web-based virtual tour application developed to enhance campus orientation and student experience at Universiti Teknologi MARA. Utilizing the Web Development Life Cycle (WDLC) methodology, the system integrates panoramic imagery, augmented reality (AR), user authentication and analytics to provide an immersive and interactive navigation tool. The application enables students to explore key campus landmarks virtually without requiring additional software installation. Usability testing using the System Usability Scale (SUS) yielded an average score of 91.88, indicating excellent user satisfaction. The findings demonstrate the effectiveness of immersive technologies in fostering spatial awareness and a sense of belonging within academic environments. Future enhancements include real-time navigation, multilingual support, and expanded campus coverage, positioning *LensaKPPIM* as a scalable solution for digital student engagement.

Keywords: Campus navigation, Immersive experience, Augmented reality.

1 Introduction

1.1 Research Background

Over the past decades, the progressive integration of technology into academic environments has significantly transformed traditional learning methods and enhanced student engagement. Within the dynamic landscape of higher education, students increasingly face challenges in navigating extensive university campuses. They are often densely populated with buildings, facilities, and resources. [1] highlighted that technological advancements have profoundly impacted human life, influencing interpersonal interactions and everyday activities. One of the primary challenges faced by new students at the university is the difficulty in navigating the campus and locating essential academic and administrative services. Traditional orientation methods, such as printed

maps or verbal directions, often lack interactivity and fail to provide real-time guidance. This gap in spatial awareness and access to information can lead to confusion, inefficiency, and a sense of disconnection among students.

In response to the growing needs of university students for immersive methods to familiarize themselves with their surroundings. Virtual tours have emerged as a viable solution to enhance their university experience. It is defined as information, communication, and technology-driven methods that allow users to experience real-world locations virtually [2]. It provides a three-dimensional (3D) representation of physical spaces, offering users an immersive experience that simulates real-world exploration. These dynamic and interactive tours are typically accessible via web platforms, mobile applications, or virtual reality (VR) devices, utilizing panoramic images and video recordings to assist students in navigating and understanding their campus environment [3]. The incorporation of multimedia elements, such as visual graphics and sound effects, further enhances the user experience, creating a realistic and engaging navigation tool [2]. The adoption of virtual tours in Malaysia is exemplified by initiatives such as the Ministry of Tourism, Arts, and Culture's Klonfoot.com platform, which offers tourists a virtual experience of exploring Kuala Lumpur [4].

Immersive virtual tours provide highly interactive experiences, integrating panoramic views, auditory cues, and interactive hotspots that supply users with additional contextual information [3]. By allowing users to interact with and navigate virtual environments, such technologies facilitate intuitive exploration of university campuses. Advanced technologies, including virtual reality (VR) and augmented reality (AR), further enhance immersion by creating a stronger sense of presence within the environment. Employing key metaverse components, particularly AR, within virtual tours enables the creation of immersive digital environments [5]. According to [6], the metaverse represents an intermediary space between individuals and the physical world, where AR technologies can overlay digital content onto real-world settings, making educational material more engaging and accessible [7]. AR applications support the development of dynamic; 3D content that sustains student interest and improves learning outcomes [8]. Recent market forecasts predict significant growth in the Asia-Pacific virtual tourism sector, driven by increased investment in AR technologies, with sustained growth expected through 2027 [9]. As web-based tools, virtual tours stimulate sensory engagement and support both individual and collaborative learning contexts. Their broad platform compatibility ensures seamless user experiences across various devices via standard web browsers, such as Google Chrome, without the need for additional software installations [10]. Thus, the web-based delivery of virtual tours effectively combines technological innovation with user accessibility, providing an adaptable platform that offers immersive, interactive experiences within academic settings.

Emerging technologies, such as virtual tours, have significantly influenced digital navigation by offering immersive and interactive solutions. For students, adapting to a new campus environment can be particularly challenging, as they are required to adjust themselves with unfamiliar surroundings in a relatively short period. This highlights the need for the development of a dynamic and user-centered application that can facilitate effective navigation in such contexts. Consequently, it is crucial to formulate an

innovative strategy that leverages technologies such as virtual tours to provide students with immersive, informative, and engaging campus navigation experiences. Three key issues have been identified within the current university navigation systems are limited provision of information that accurately captures the real campus ambiance [11]; insufficient detail within existing navigational resources [12-14]; and feelings of demotivation among students caused by environmental challenges encountered on campus [15].

The purpose of this research is to propose a web application, named *LensaKPPIM*, aimed at supporting students in navigating from Melati Residential College to the Al-Khawarizmi Building in UiTM Shah Alam Campus through an immersive, technology-enhanced experience. By integrating interactive virtual elements, the system is intended to facilitate students' spatial understanding of campus infrastructure, enabling them to explore key buildings, pathways, and landmarks with greater confidence. Ultimately, this initiative seeks to promote a sense of familiarity, belonging, and connection within the academic environment, thereby enhancing overall campus experience.

2 Related Works

Virtual tours (VTs) have emerged as a significant innovation in digital navigation and education, allowing users to explore real-world environments virtually through digital imagery and interactive elements [16]. Modern VTs leverage technologies such as Augmented Reality (AR), Virtual Reality (VR), and Internet of Things (IoT) integrations to provide immersive, three-dimensional (3D) experiences. These technologies create dynamic virtual environments that simulate physical navigation, facilitating spatial orientation and engagement without requiring physical presence [17].

In educational contexts, virtual tours offer critical benefits by improving students' spatial awareness, enhancing orientation processes, and fostering a sense of presence within unfamiliar environments [18-19]. Virtual campus tours assist new students in adapting to their academic surroundings, reducing anxiety linked to environmental unfamiliarity [20]. The inclusion of visual graphics, multimedia elements, panoramic images, and landmark labelling contributes to a richer, more interactive learning and navigation experience [2-3].

Existing implementations in Malaysian institutions of higher learning (MIHL), have demonstrated widespread adoption of virtual campus tours. However, these systems often lack advanced features such as personalized user authentication, dynamic analytics such as journey timelines and interactive commenting functionalities, limiting their capacity for personalized, data-driven navigation [21-22]. Studies highlighted that while panoramic views and landmark labelling are common, deeper interactivity and secure, personalized navigation remain underexplored. The application of Human-Computer Interaction (HCI) principles in designing intuitive, responsive, and visually appealing interfaces supports user engagement and system usability [23]. Additionally, integrating AR in marker-less formats simplifies content delivery by overlaying virtual content onto real-world environments without reliance on physical markers [24].

Although, VT technologies have been implemented within Malaysian higher education institutions, the current existing systems remain limited in offering integrated

features that support immersive navigation, personalized user interaction, and interactive engagement. Existing solutions predominantly focus on panoramic visuals and basic landmark labelling, while neglecting advanced functionalities such as secure user authentication, analytical tools, and real-time user feedback mechanisms. The development of the *LensaKPPIM* web application aims to address these limitations by incorporating augmented reality, responsive web interfaces, and dynamic multimedia content, thereby enhancing campus navigation and improving the overall student experience.

3 Development and Implementation

The development and implementation of *LensaKPPIM* were systematically conducted using the Web Development Life Cycle (WDLC) methodology, a comprehensive life cycle model for web-based systems that covers all stages of web application development [25]. In the initial phase, the overall goals and direction of the study were defined. This includes identifying the objectives, understanding the target users, who are the students of the College of Computing, Informatics, and Mathematics (KPPIM) and determining the scope and feasibility of the proposed application. Project timeline and task distribution plan were established using the Substitute, Combine, Adapt, Modify, Purpose, Eliminate, and Rearrange (SCAMPER) technique to stimulate innovation. This phase ensured that development efforts aligned with user needs and institutional goals. Table 1. shows the *LensaKPPIM* SCAMPER table.

Table 1. SCAMPER table of *LensaKPPIM*

	Transformation	Solution Ideas
S	**SUBSTITUTE**	Replacing the traditional navigation method for student from using text like physical map and signages to a digital approach by implementing the virtual tour feature. This offer students a more interactive and engaging way to explore campus locations.
C	**COMBINE**	Integrating panoramic view with navigational tools like zooming in/out and full-screen view to allow a comprehensive exploration of campus landmarks.
A	**ADAPT**	Adapting the emerging technologies based on the Progressive Web Application (PWA) to ensure that the web app remains relevant and effective in the evolving needs of users over time.

M	**MODIFY/MAGNIFY/ MINIFY**	Enhancing the interface design to be more visually appealing, aesthetic and user friendly for a familiar design and structure.
P	**PURPOSE/PUT TO OTHER USES**	Implementing features like user profiling, journey estimation, and commenting, provide additional functionalities that cater to user needs and preferences.
E	**ELIMINATE**	Simplifying the functionality of the web app streamlines the user experience, making it easier for users to navigate the virtual tour and access relevant information without unnecessary complexities.
R	**REVERSE/ REARRANGE**	Rearranging features based on stakeholder feedback ensures that the web app aligns with user expectations and preferences, optimising usability, and effectiveness.

The analysis phase focused on gathering and analyzing the application requirements. Both functional and non-functional requirements were outlined to determine what the application must do and under which conditions. Hardware and software specifications were also defined, ensuring that the infrastructure could support the intended features of the application. Key deliverables in this phase included requirement checklists, resource identification and the foundation for application design.

The design phase plays a crucial role in shaping the visual identity and structural layout of the web application, while also ensuring a systematic approach to development. Establishing clear design guidelines is essential to support the attainment of the secondary objective, particularly in aligning user experience with functional goals. A variety of tools and techniques are utilized to optimize the design process and ensure its effective execution. One such tool is the development of user personas, a fictional representation of target users which serve to capture user needs, behaviors, preferences, and goals [27-28]. These personas are instrumental in guiding design decisions, ensuring that the web application aligns with the expectations and requirements of its intended users. The usability and relevance of the application within the context is enhanced by constructing detailed user profiles that encompass demographic characteristics, behavioral patterns and user motivations. The design process gains valuable insights into the diverse needs of the user base. Figure 1 illustrates the storyboard of *LensaKPPIM* respectively.

Fig. 1. Storyboard of *LensaKPPIM*

According to [29], storyboards serve as visual tools that depict a sequence of user interactions through annotated sketches or illustrations, offering contextual understanding of a proposed concept. In web development, storyboarding is essential for clarifying project objectives and anticipating the user journey. It enables designers to visualize user interactions with the interface, identify usability issues, and propose early solutions to mitigate potential risks [30]. In this study, the storyboard in Figure 1 presents a detailed overview of a user named Ashraf, illustrating his navigation experience on campus before and after using the *LensaKPPIM* web application. This visual narrative supports the design and development process by highlighting key user behaviors and needs.

In order to ensure the platform is intuitive for all users, including those unfamiliar with the university environment, *LensaKPPIM* is designed to explore several key areas that are essential for student orientation and daily navigation. These include academic zones such as faculty buildings, lecture halls, and laboratories; administrative offices like the registrar, finance, and student affairs; and student-centric spaces such as activity centers, event venues, and recreational facilities. While these areas may be naturally understood by the authors or long-time university members, they can be confusing for new students or external users. Therefore, the platform provides clear labels, contextual descriptions, and interactive guidance to help users locate and understand the purpose of each area within the campus ecosystem.

Building upon the conceptual framework and design principles outlined earlier, the next phase focuses on translating these ideas into a functional digital solution. The development phase represents a critical stage in the creation of the *LensaKPPIM* web application, focusing on the implementation of an interactive, user-friendly system using dynamic web development technologies. This phase involves the construction of both the user interface (UI) and user experience (UX) components, which form the primary points of interaction for end-users. To ensure the functionality and reliability of the system, user testing is essential upon completion of development, as it facilitates the identification and rectification of potential errors or bugs. The front-end development was carried out using Visual Studio Code (VSC) as the primary integrated

development environment (IDE), offering a flexible and efficient platform for code management. Adobe Lightroom was employed for stitching and enhancing the panoramic images used in the virtual tour component. Figure 2 to Figure 5 illustrate the prototype for *LensaKPPIM*.

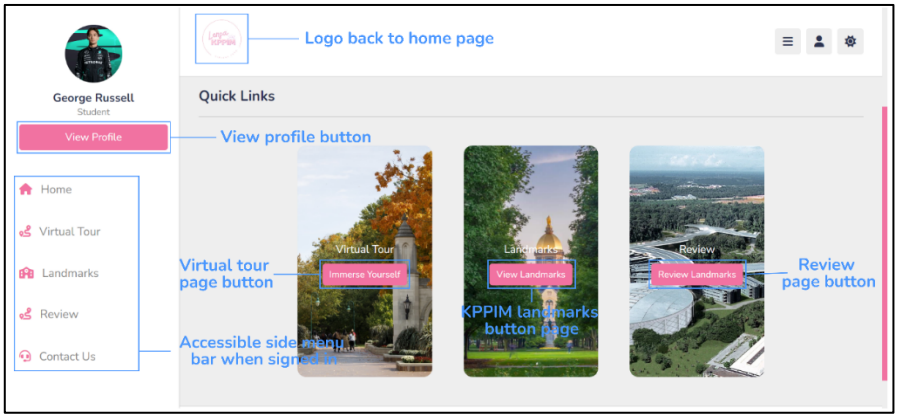


Fig. 2 Home page

The Home page as in Figure 2 acts as the central hub for the *LensaKPPIM* web application, welcoming users after successful authentication. It provides an overview of the primary functions, including access to the virtual tour, landmarks, user profile, and system announcements. Dynamic content is featured to highlight updates, featured landmarks, or new functionalities. The layout reflects user-centred design principles, ensuring that both novice and experienced users can locate features with ease.

Fig. 3. Virtual Tour

The Virtual Tour page as in Figure 3, is a core feature of the *LensaKPPIM* web application, offering an immersive and interactive experience that allows students to explore key campus landmarks virtually. Powered by CloudPano, the page displays panoramic images that simulate real-world environments, enabling users to navigate through locations such as the Al-Khawarizmi building and Melati Residential College. Users can zoom in and out, view labelled landmarks, and switch to full-screen mode for a more engaging experience. The interface supports markerless AR elements, enhancing the realism and interactivity of the tour. Navigation tools and clickable hotspots provide additional information about each location, aiding students in familiarising themselves with their surroundings. This page plays a significant role in reducing anxiety associated with campus navigation, particularly for new students, by offering a sense of orientation and familiarity before physically visiting the campus. Overall, the Virtual Tour page embodies the aim to integrate technology into student life, improving campus engagement and accessibility.

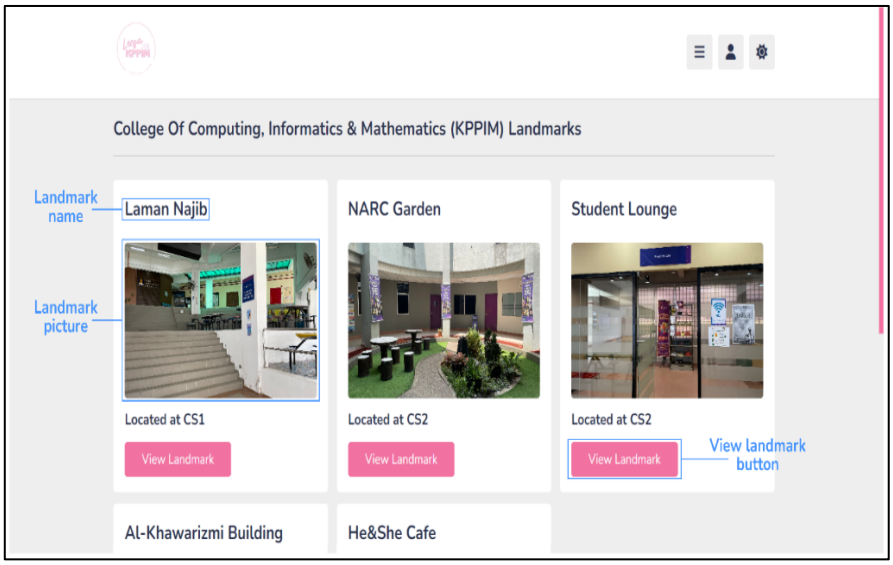


Fig. 4. Landmarks (1)

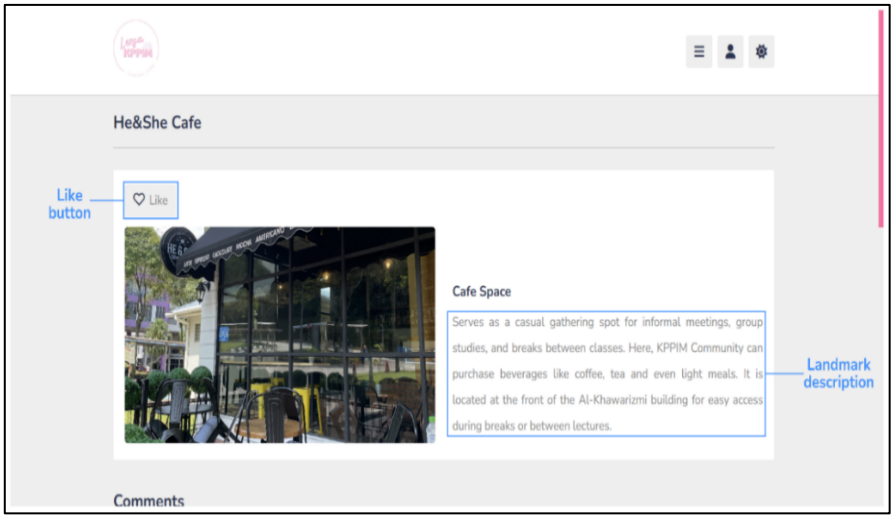


Fig. 5. Landmarks (2)

The Landmark pages as refelected in Figure 4 and 5 decribe on dedicated feature within the *LensaKPPIM* web application that provides detailed information and interactive media for specific campus landmarks. It serves as an informative and visually engaging section where users can explore significant locations such as the Al-Khawarizmi building, Melati Residential College and other key facilities within the KPPIM area. Each landmark entry includes a combination of panoramic images, descriptive texts, and location-specific details to enhance user understanding and navigation. The page integrates interactive features such as labelled points of interest and zoom capabilities, allowing users to examine structural layouts and key areas more closely. Additionally, the page supports a commenting system, enabling users to add, view, edit, and delete their feedback or personal reflections on each location. This fosters peer-to-peer engagement and community building among students. The Landmark page not only supports spatial awareness for new students but also contributes to the overall immersive experience of the virtual tour. This page facilitates better preparation for on-campus navigation while promoting a sense of familiarity and confidence within the academic environment.

4 Result and Discussion

The user testing for the *LensaKPPIM* web application was conducted using the System Usability Scale (SUS), the evaluation of the functionality and overall user experience of the application. The SUS participants consisted of undergraduate students from various faculties within the university, primarily in their first and second years of study. They were intentionally selected based on their limited familiarity with the campus layout and administrative structures, which aligns with the core objective of the research which is to address the challenge of orientation and navigation within the university environment. By involving users who are most likely to experience

disorientation such as new students, the evaluation could more accurately reflect the usability and effectiveness of the LensaKPPIM platform in guiding users through key academic and administrative areas. Their diverse academic backgrounds also ensured that feedback captured a broad range of user experiences and expectations. Figure 6 demonstrates the SUS score for *LensaKPPIM*.

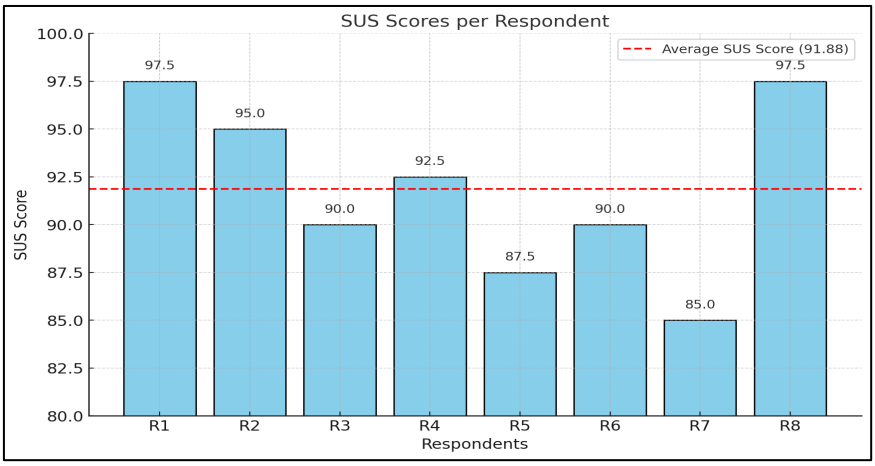


Fig. 6. SUS result for *LensaKPPIM*

As reflected in Figure 6, the individual SUS scores ranged from 85.0 to 97.5, indicating a consistently high level of user satisfaction. Specifically, Respondents 1 and 8 achieved the highest SUS scores of 97.5, while Respondent 7 recorded the lowest score of 85.0. The average SUS score across all eight participants was calculated to be 91.875, which is considered to be in the "excellent" range according to standard SUS interpretation guidelines. This suggests that users found the system to be highly usable, intuitive and well-designed. The high scores reflect positive user experiences in terms of ease of use, system efficiency, and confidence in navigating the application. Furthermore, the relatively narrow range of scores indicates consistent usability across different user backgrounds and environments. These results affirm the effectiveness of the system design and validate its readiness for broader implementation among university students.

5 Conclusion

This research presented the design, development and evaluation of *LensaKPPIM*, a virtual tour web application aimed at enhancing campus navigation and engagement among university students at UiTM Shah Alam. The application incorporates web-based technologies and panoramic virtual content to simulate real-world navigation between key landmarks, particularly from Melati Residential College to the Al-Khawarizmi building. This application indicates an excellent level of user satisfaction,

confirming that the application is intuitive, reliable, and effective for its intended purpose. These findings support the feasibility and value of integrating immersive technologies within academic environments to address challenges related to orientation and spatial awareness.

Acknowledgments. The authors wish to thank Universiti Teknologi MARA, Malaysia for funding this research.

Disclosure of Interests. The authors hereby declare that there are no known financial, commercial, or personal relationships that could be construed as potential competing interests in the development, implementation, or reporting of this study. All aspects of the research, including data collection, analysis, and interpretation, were conducted independently and without influence from any external parties or organizations. This study was undertaken solely for academic purposes, with the intention of contributing to the field of educational technology and supporting future research.

References

1. Polishchuk, E., Bujdosó, Z., El Archi, Y., Benbba, B., Zhu, K., & Dávid, L. D.: The Theoretical Background of Virtual Reality and Its Implications for The Tourism Industry. *Sustainability*, *15*(13), 10534 (2023).
2. Ankomah, P., & Larson, T.: Virtual Tourism and Its Potential for Tourism Development in Sub-Saharan Africa. *Encyclopedia of Information Science and Technology, Fourth Edition*, 4113–4122 (2018).
3. Tengku Wook, T. S., Zairon, I. Y., Sahari@Ashaari, N., Idris, M., Mat Zin, N. A., Mohamad Judi, H., & Jailani, N.: Campus Virtual Tour Design to Enhance Visitor Experience and Interaction in A Natural Environment. The International Journal of Multimedia & Its Applications, 10(1/2/3), 77–92 (2018).
4. Rosli, N., Johar, E. R., Omar Zaki, H., & Mohd Farid Fernandez, D. F.: The Rise of Virtual Tour in Tourism: A Bibliographic Review and Future Research Agenda. International Journal of Academic Research in Business and Social Sciences, 13(2) (2023).
5. Dwivedi, Y. K., Hughes, L., Baabdullah, A. M., Ribeiro-Navarrete, S., Giannakis, M., Al-Debei, M. M., Dennehy, D., Metri, B., Buhalis, D., Cheung, C. M. K., Conboy, K., Doyle, R., Dubey, R., Dutot, V., Felix, R., Goyal, D. P., Gustafsson, A., Hinsch, C., Jebabli, I., & Wamba, S. F.: Metaverse Beyond The Hype: Multidisciplinary Perspectives on Emerging Challenges, Opportunities, and Agenda for Research, Practice and Policy. International Journal of Information Management, 66, 102542 (2022).
6. Damar, M.: Metaverse Shape of Your Life for Future: A Bibliometric Snapshot. Journal of Metaverse, 1(1), 1-8. (2021).
7. Young, G., Stehle, S., Walsh, B., Tiri, E.: Exploring Virtual Reality in The Higher Education Classroom: Using VR to Build Knowledge and Understanding. JUCS - Journal of Universal Computer Science, 26(8), 904–928 (2020).
8. Gargrish, S., Mantri, A., Kaur, D., P.: Augmented Reality-based Learning Environment to Enhance Teaching-learning Experience in Geometry Education. Procedia Computer Science, 172, 1039–1046 (2020).
9. Market Data Forecast (2023). Virtual Tourism Market Research Report - Industry Analysis & Forecast (2023-2028).

10. Sarhan, Q., I., Gawdan, I., S.: Web Applications and Web Services: A Comparative Study. Science Journal of University of Zakho, 6(1), 35 9 (2018).
11. Meliana, M., Su Mon, C.: A Preliminary Study on Requirement of SMART Tour Guide Application Using Augmented Reality. 10th International Conference on Software and Computer Applications (2021).
12. Harwood, J.: 7 Reasons Why You Need A Virtual Campus Tour. Concept3D (2023).
13. Khan, T., Johnston, K., Ophoff, J.: The Impact of An Augmented Reality Application on Learning Motivation of Students. Advances in Human-Computer Interaction, 1–14 (2019).
14. Figueroa, R., B., Mendoza, G., A., Fajardo, J., C., Tan, S., E., Yassin, E., Thian, T., H.: Virtualizing A University Campus Tour: A Pilot Study on Its Usability and User Experience, and Perception. International Journal in Information Technology in Governance, Education and Business, 2(1), 1–8 (2020).
15. Ding, F., Curtis, F.: 'I Feel Lost and Somehow Messy': A Narrative Inquiry into The Identity Struggle of A First-year University Student. Higher Education Research & Development, 40(6), 1146–1160 (2020).
16. Zhao, W., Huang, Y.: How Does Virtual Tourism Affect The Real tourism: A Perceptual Perspective of The "New Generation" in China. Travel and Tourism Research Association: Advancing Tourism Research Globally, 30 (2022).
17. Arena, F., Collotta, M., Pau, G., Termine, F.: An Overview of Augmented Reality. Computers, 11(2), 28 (2022).
18. Chang, M., Lee, G., Hyun Lee, J., Lee, M., Lee, J.-H.: The Influence of Virtual Tour on Urban Visitor Using A Network Approach. Advanced Engineering Informatics, 56, 102025 (2023).
19. Verma, S., Warrier, L., Bolia, B., Mehta, S.: Past, Present, and Future of Virtual Tourism-A Literature Review. International Journal of Information Management Data Insights, 2(2), 100085 (2022).
20. Suwarno, Murnaka, N., M.: Virtual Campus Tour (Student Perception of University Virtual Environment). Journal of Critical Reviews, 7(19) (2020).
21. Rohizan, R., B., Vistro, D., M., Puasa, M., R.: Enhanced Visitor Experience through Campus Virtual Tour. Journal of Physics: Conference Series, 1228(1), 012067 (2019).
22. Wu, X., Lai, I., K.: The Use of 360-degree Virtual Tours to Promote Mountain Walking Tourism: Stimulus–Organism–Response Model. Information Technology & Tourism, 24(1), 85–107 (2021).
23. R, P., Sanjaya, K., Rathika, S., Hussein Alawadi, A., Makhzuna, K., Venkatesh, S., Rajalakshmi, B.: Human-Computer Interaction: Enhancing User Experience in Interactive Systems. E3S Web of Conferences, 399, 04037 (2023).
24. Hajirasouli, A., Banihashemi, S.: Augmented Reality in Architecture and Construction Education: State of The Field and Opportunities. International Journal of Educational Technology in Higher Education, 19(1) (2022).
25. Amadi, D., N., Utomo, P., Budiman, A.: Design and Build of Road Damage Information System in Madiun Regency using Web Development Life Cycle Methods. Journal of Information Systems and Informatics, 4(4), 1112–1125 (2022).
26. Mohamad Nowawi, N., L., Ahmad, N., A.: Malay Language Learning for Kindergarten Students through Interactive Web-based Application. International Journal of Academic Research in Progressive Education and Development, 12(2) (2023).
27. Dam, R. F., Siang, T., Y.: Personas – A Simple Introduction. The Interaction Design Foundation (2024).

28. Salminen, J., Wenyun Guan, K., Jung, S.-G., Jansen, B.: Use Cases for Design Personas: A Systematic Review and New Frontiers. CHI Conference on Human Factors in Computing Systems (2022).
29. Wan Husain, W., S., Che Hassan, S., H., Nik Kamaruzaman, N. N., Wan Aziz, W., A. H., Wan Abdul Rahman, W., W.: From Scratch to Storyboard: Incorporating Techniques for Novice Users. Journal of Mathematics and Computing Science, 6(2), 9–19.4.2.5 Use Case Diagram (2020).
30. Rahmi, A., Mahyuddin, N.: Design & Application of Storyboard in Teaching Characters for Children Aged 6–8 Years. Proceedings of the International Conference of Early Childhood Education (ICECE 2019).

Classification Hate Speech in Boosting Algorithms for Trending Twitter cases Domestic Violence in Indonesia

Agus Nursikuwagus[1][0000-0001-8435-7522] and Syahrul Mauluddin[2][0000-0002-6125-6523]

[1,2] Universitas Komputer Indonesia, Bandung 40132, Indonesia
agusnursikuwagus@email.unikom.ac.id

Abstract. Boosting accuracy is one technique to increase a good result especially in text processing. This strategy used to boost the classify the sentence, is the sentence classify to hate speech or not. The appropriate of result depends to how a good the machine learning parameter tuning. The problem of sentence classification is not only parameter tuning in machine learning, but in feature extraction manner. The suitable process for text processing is to be a successful classification and reach a high accuracy. One of the datasets uses for the experiment is a "domestic violence" as many as 562 datasets consisting of 520 neutral sentences, 33 positive sentences, and 9 negative sentences. The dataset "domestic violence" (KDRT) is one of the objective datasets in Bahasa which uses for classification. We propose a boosting algorithm like XGBoost to improve the classification which compares with SVM and KNN. KNN are the baseline model that used for primary machine learning to get the gap an accuracy. To conduct the research, we used many stages of methods such as text preprocessing, modelling, evaluating, and visualization. On the experiment, we reached the accuracy for each machine learning with the parameter used. The accuracy for KNN is 0.703, SVM is 1.0, and XGBoost is 1.0 for 112 testing datasets. KNN for K=5 is still difficult to reach high accuracy than SVM dan XGBoost. Some challenges found when the experiment like not each slang word defines correctly and imbalance process not initiate in this research.

Keywords: Classification, Boosting Algorithm, Hate Speech, Performance, Social Media, Product Innovation, health and wellbeing.

1 Introduction

1.1 Background

Social media serves as a medium for articulating a diverse array of thoughts and feelings [1]. Social contact encompasses three fundamental elements: recognition, communication, and cooperation. Twitter is one of the most widely utilized social media platforms in Indonesia at present [2,3]. Twitter is a social networking platform that enables users to disseminate messages referred to as "tweets." Jack Dorsey founded Twitter on March 21, 2006, with its headquarters located in San Francisco, California. Twitter has multiple functions, including the trending topic function, which emphasizes prevalent

talks among users. Twitter grants users the liberty to articulate their views and emotions, resulting in both beneficial and detrimental effects. This liberty of expression may also lead to hate speech. Davidson et al. (2017) define hate speech as a mode of expression that incites, disseminates, justifies, or advocates hatred, discrimination, and violence against an individual or group for diverse causes. Text categorization seeks to ascertain the categorical class of a specific data instance. Common techniques for text classification include Naïve Bayes [4], Support Vector Machines (SVM) [5], Logistic Regression[6], Neural Networks [7], and K-Nearest Neighbors (KNN) [8].

One of the trending topics on Twitter in October 2022 was “KDRT” (domestic violence), which generated significant reactions from users and led to the emergence of various forms of hate speech. In order to assess hate speech, it is necessary to calculate accuracy metrics to determine which algorithm is most suitable for performing the classification task.

1.2 Problem

The main problem in this study is how to automatically and accurately classify hate speech from Indonesian texts sourced from Twitter, especially in the context of the trending topic "domestic violence". Challenges in text classification include the diversity of language styles, the use of slang, as well as the cultural context that influences the interpretation of speech. Based on the problems faced, this research has the following objectives:

- Preprocessing carried out on Twitter's domestic violence dataset such as remove stopwords , slang words, punctuation, stemming, lematization, and balancing class was able to provide classification results with high accuracy [9].
- Vectorization using TF-IDF as a word transformation method drives sentiment analysis results with high accuracy [10].
- The balancing method with SMOTE for Twitter's domestic violence dataset is able to provide fairer predictive results in classifying results [11].
- The use of XGBoosting as an ensemble algorithm is able to provide model results with an accuracy that exceeds KNN as a baseline [12].

The writing of this article is arranged in the prescribed order. The order starts from the first part, which is an introduction which contains ideas, problems, contributions, and systematics of writing. The second part is about the method that contains the method used to solve the prediction problem of this sentiment analysis. The third part is about the results and discussion which contains the presentation of the results of the research and the discussion which contains a comparison between the proposed model and the baseline. The fourth part is the conclusion which contains the alignment of the results with the contribution obtained from the results of the experiment.

2 Classification Method

This section describes the methods used and the steps of the research. The stages of research from data analysis to website development are as follows:

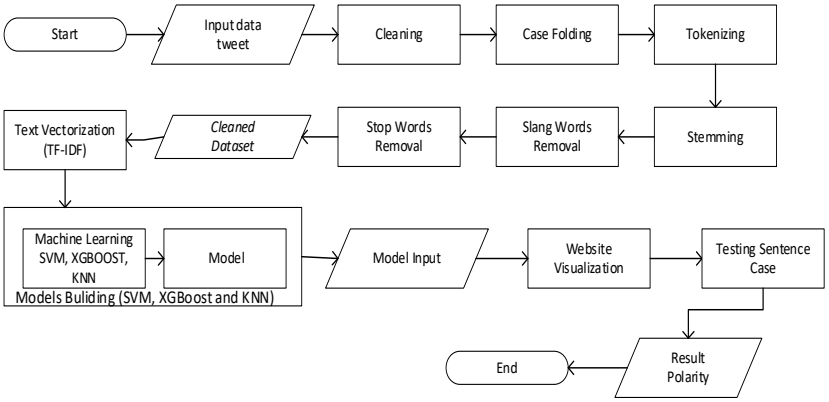


Fig. 1. A Pipeline chart classification model

2.1 Text Preprocessing

Case folding is one of the early stages in text preprocessing that aims to standardize capital and lowercase letters into a single shape, usually all letters are converted into lowercase letters. The goal is to reduce word variability due to differences in capitalization so as to make it easier to analyze the text.

Slang removal or slang normalization is the process of identifying and replacing non-standard words (slang, abbreviations, slangs, typo) into standard or formal words so that the text is easier to analyze by machines. Slang removal is especially important in data from social media (Twitter, Instagram, WhatsApp, etc.) because many users write in informal language that is not recognized by standard NLP dictionaries.

TF-IDF (Term Frequency–Inverse Document Frequency) is a word weighting method used in text mining and Natural Language Processing (NLP) to evaluate how important a word is in a document relative to the entire body of documents (corpus). Increase the weight of words that are infrequent but meaningful [10,13–15].

$$\mathrm{TF}(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \tag{1}$$

Where: t: word, d: document, $f_{t,d}$: the number of occurrences of the word t in document d. Inverse Document Frequency (IDF): Measures how rarely a word appears in all documents:[14]

$$\mathrm{IDF}(t) = \log\left(\frac{N}{1 + |\{d \in D : t \in d\}|}\right) \tag{2}$$

Where: N: total number of documents, Number of documents containing the word t|{d ∈ D : t ∈ d}|, Addition **of +1** to denominator to prevent division by zero. A word that appears frequently in one document but rarely appears in another will get a high TF-IDF score — signifying that the word is important.

$$\mathrm{TF} - \mathrm{IDF}(\mathrm{t}, \mathrm{d}) = \mathrm{TF}(\mathrm{t}, \mathrm{d}) \mathrm{x}\ \mathrm{IDF}(\mathrm{t}) \tag{3}$$

2.2 Machine Learning Classification

In the context of text classification, SVM is very effective because it is able to handle high-dimensional data, such as text representations in the form of TF-IDF or bag-of-words. SVM works by finding the best hyperplane that separates data from the two classes with the largest margin [16]. Objective Function for binary classifications with SVM labels $y_i \in \{-1, 1\}$ to maximize the margin between two classes $\min_{w,b} = \frac{1}{2}\|w\|^2$, the constraint hand: $y_i(w \,.\, x_i + b) \geq 1, \text{for all i}$. Where w: vector weight, b: bias (intercept), x_i : features of the first data, y_i: class labels of the i data. Soft Margin SVM (for data not linearly separable): $\min_{w,b,\xi} = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i$, with constraint $y_i(w \,.\, x_i + b) \geq 1 - \xi_i$ dan $\xi_i > 0$, ξ_i slack variable error, C is the regularization parameter (trade-off between margin and error). Kernel Function (if the data cannot be separated linearly : $K(x_i, x_j) = \varphi(x_i).\varphi(x_j)$ with the type of linear kernel $K(x_i, x_j) = (x_i).(x_j)$, Polynomial $K(x_i, x_j) = (x_i \,.\, x_j + 1)^d$, RBF (gaussian) $K(x_i, x_j) = \exp\left(-\gamma\|x_i - x_j\|^2\right)$.

K-Nearest Neighbors (KNN) is an instance-based learning-based non-parametric classification algorithm. This means that the KNN does not build an explicit model during training, but classifies documents based on their proximity to the training data. In text classification, KNN is used to classify a text document (e.g., tweet, email, or review) into one of the classes (e.g., positive/negative, spam/non-spam) based on similarity to previous documents. KNN Suitable for high-dimensional datasets (such as TF-IDF), Easy to implement, Does not require complicated training processes, can be used for binary as well as multi-class classification. The way KNN works is Measure the Distance (Similarity) between Documents, with Cosine Similarity (most common in text), Euclidean Distance, Manhattan Distance [8].

KNN Algorithm:

1. Represent the document as a feature vector (e.g. with TF-IDF)
2. Calculate the distance/similarity between the test document and all the training documents
3. Choose the nearest neighbor K
4. Use the majority of labels from the K neighbor to determine the class of the test document, $\hat{y} = \mathrm{majority}_{\mathrm{vote}}(y_1, y_2 \ldots y_k)$

XGBoost (Extreme Gradient Boosting) is a very popular and powerful tree ensemble-based machine learning algorithm for classification and regression [17]. XGBoost

is a development of the Gradient Boosting Decision Tree (GBDT) algorithm that is optimized for high efficiency, speed, and accuracy. Fast and efficient (using parallel computation system), Supports missing value handling, Avoids overfitting through regularization (L1 and L2), Can be used for big data and competitions such as Kaggle, Suitable for text classification, fraud detection, customer churn.

XGBoost Algorithm:

1. XGBoost creates an initial model (target bias or average).
2. The next model learns from the rest of the (residual) errors of the previous model.
3. This process is repeated by adding a new decision tree to correct previous mistakes.
4. All the trees are combined (ensemble) to make the final prediction.

The basic formula of XGBoost, for example. Suppose y_i : actual target, $\widehat{y_i}^t$: prediction at iteration t^{th}, $f_i(x_i)$: tree model at model pohon iteration t^{th}, $\mathcal{L}$: loss function, $\Omega(f)$: regulatization model complexity. The prediction model can be written as follows: $\widehat{y_i} = \sum_{t=1}^{T} f_t(x_i), f_t \in F$. Where F is the function space of the decision tree.

Objective Function: $\mathcal{L} = \sum_{i=1}^{n} l(y_i, \widehat{y_i}) + \sum_{t=1}^{T} \Omega(f_t)$. Taylor Expansion Approach (up to 2nd order): To efficiently calculate, the loss function is developed using Taylor expansion:): $\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_i(x_i) + \frac{1}{2} h_i f_t (x_i)^2 \right] + \Omega(f_t)$. Where $g_i = \frac{\partial l\left(y_i, \widehat{y_i}^{(t-1)}\right)}{\partial \widehat{y_i}^{(t-1)}}$, and $f_i = \frac{\partial^2 l\left(y_i, \widehat{y_i}^{(t-1)}\right)}{\partial \widehat{y_i}^{(t-1)^2}}$.

The parameter for each machine learning, we set fro SVC using SVC (*, C=1.0, kernel='rbf', degree=3, gamma='scale', coef0=0.0, shrinking=True, probability= False, tol=0.001, cache_size=200, class_weigh t= None, verbose = False, max_iter=-1, decision_function_shape='ovr', break_ties=False, random_state=Non). For KNN, we used K=5 selected point, and XGBoost, we leveraged default mode.

3 Result and Discussion

Following the title section, we explore some experiment referring to the model such as using of SVM, KNN, and Boosting algorithm. We ordered the subsection like dataset, result, and discussion.

3.1 Dataset

A total of 562 hate speech tweet data about domestic violence will be tested by sharing datasets, namely 80% training data and 20% test data. Text preprocessing is done first on tweet data, so that the document becomes a collection of basic words that have been cleaned of words, letters, and symbols as well as numbers that are not needed for weighting. Here is an example of a research data structure at Table 1 [18].

The dataset used still needs to be tidied up so that when classifying it can be processed and read clearly. Fig. 2, there were 562 datasets consisting of 520 neutral sentences, 33 positive sentences, and 9 negative sentences.

3.2 Text Preprocessing

The text must be cleaned up in order to be able to select according to the relevant features. The text preprocessing carried out in this experiment involves case holding, removing stopwords, slang words, and tokenizing. The perform of slang word, we used the Indonesia slang word corpus https://raw.githubusercontent.com/louis-owen6/NLP_bahasa_resources/master/combined_slang_words.txt to identify dan remove where the slang or not. The following are the results of the text preprocessing carried out [9].

Table 1. An example of dataset in Bahasa [18].

No	Tweet	Sentiment
1	billar disebut kena mental pasca dilaporkan lesti kejora kini kepergok sukai video tak senonoh beritaterkini rizkybillar lestikejora taksenonoh kepergok kd klik selengkapnya gt gt	Negative
2	kekerasan dalam rumah tangga kdrt berkontribusi memberikan pengaruh terhadap mental anak dalam pertumbuhannya kdrt	Negative
3	bukan sekali saja ini video bukti kdrizky billar saat lesti kejora sedang hamil kdlestikejora rizkybillar	Neutral
4	ingat bang markus dulu pernah digugat cerai oleh istrinya sekarang giliran lesti kejora bang bima sakti adem ayem tiba tiba menghentakan prestasi good luck sks timnas markus kdlestikejora rizkybillar	Positive
…	…	…

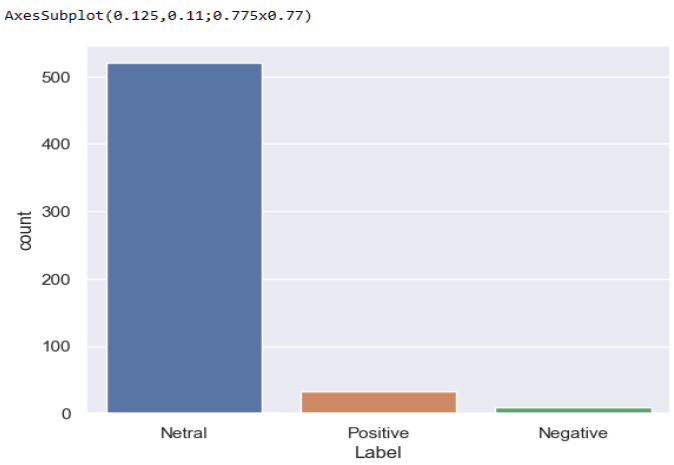


Fig. 2 . A distribution for every instance in targeting.

Tokenizing is the process of dividing text that can be in the form of sentences, paragraphs or documents, into specific tokens/sections. In this text preprocessing, tweet data will be processed to then produce a document compiled from a collection of basic words. After Tokenizing, process continue to stemming. Stemming is a necessary step to reduce and solve words, becoming a combination of root words that are in accordance with the word that will be processed later. The next is remove slang words. At this stage of slang words will remove slang words or words that are not standard and unnecessary. The continue process is stop word. Stop word removal is a command used to remove characters or words that are not needed. The result of text preprocessing can see at Fig. 3.

	Label	Tweet	length	Case_folded	Tokenized	Stemmed	No_Slang	No_Stop
0	Netral	rekaman cctv rizki billar melempar bola bill...	145	rekaman cctv rizki billar melempar bola billia...	[rekaman, cctv, rizki, billar, melempar, bola,...	[rekam, cctv, rizki, billar, lempar, bola, bil...	[rekam, cctv, rizki, billar, lempar, bola, bil...	[rekam, cctv, rizki, billar, lempar, bola, bil...
1	Netral	bukan sekali saja ini video bukti kdrizky bi...	107	bukan sekali saja ini video bukti kdrizky bill...	[bukan, sekali, saja, ini, video, bukti, kdriz...	[bukan, sekali, saja, ini, video, bukti, kdriz...	[bukan, sekali, saja, ini, video, bukti, kdriz...	[video, bukti, kdrizky, billar, lesti, kejora,...
2	Netral	kasus prank kdyang dilakukan baim wong dan p...	204	kasus prank kdyang dilakukan baim wong dan pau...	[kasus, prank, kdyang, dilakukan, baim, wong, ...	[kasus, prank, kdyang, laku, baim, wong, dan, ...	[kasus, prank, kdyang, laku, baim, wong, dan, ...	[prank, kdyang, laku, baim, paula, verhoeven, ...
3	Netral	menurut kasandra putranto psikolog dari ui m...	237	menurut kasandra putranto psikolog dari ui men...	[menurut, kasandra, putranto, psikolog, dari, ...	[turut, kasandra, putranto, psikolog, dari, ui...	[turut, kasandra, putranto, psikolog, dari, ui...	[kasandra, putranto, psikolog, ui, kdpada, aki...
4	Netral	soal prank kdrt kameramen baim wong dicecar ...	105	soal prank kdrt kameramen baim wong dicecar pe...	[soal, prank, kdrt, kameramen, baim, wong, dic...	[soal, prank, kdrt, kameramen, baim, wong, cec...	[soal, prank, kdrt, kameramen, baim, wong, cec...	[prank, kdrt, kameramen, baim, cecar, baimwong...

Fig. 3. A result of text preprocessing.

3.3 Vectorizing

The following is the result of the TF-IDF (Term Frequency-Inverse Document Frequency) which was carried out and the output was obtained as in Fig. 4 with the output shape or dimension is 562 x 1907.

```
In [44]: from sklearn.feature_extraction.text import TfidfVectorizer
         X = dataset['Ready']
         label = dataset['Label']
         tfidf_vectorizer = TfidfVectorizer()
         tfidf_vector = tfidf_vectorizer.fit_transform(X)
         tfidf_vector.shape

Out[44]: (562, 1907)
```

Fig. 4. A Shape dataset after TF-IDF utilized

From the wordcloud Hate Speech Fig. 5(a), it can be seen that the word that is widely used is Mental. From the wordcloud Non-Hate Speech Fig. 5(b), it can be seen that the word that is widely used is rizky billar.

3.4 Modelling

The K-Nearest Neighbor (KNN) algorithm showed perfect performance on classifications with values of TP = 31, TN = 110, FP = 81, and FN = 0 at Fig. 6. All data is successfully classified correctly without errors. The accuracy, precision, recall, and F1-score values are 100% each, which indicates that the model is highly accurate in

predicting positive and negative classes. Although simple, KNN is quite effective if the data is distributed in a balanced manner, but sensitive to data that is large in volume or has high dimensions.

(a) (b)

Fig. 5. A result hate speech wordcloud (a) and Non-hate speech (b)

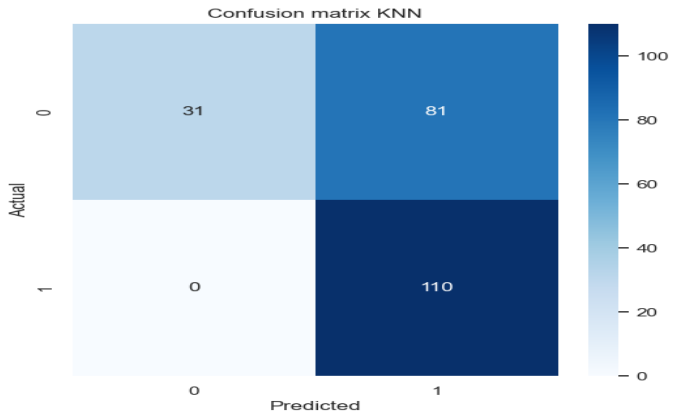


Fig. 6. An example matrix confusion of KNN

Extreme Gradient Boosting (XGBoost) is a decision tree-based algorithm that is highly efficient in the classification of hate speech and sentiment [19]. The advantage of XGBoost lies in its ability to handle large data with high efficiency and reduce overfitting through regularization. These algorithms often produce high accuracy on complex text classification tasks. The accuracy, precision, recall, and F1-score values are 100% respectively. This reflects that XGBoost is very effective at recognizing both positive and negative data perfectly. Perfect precision means all positive predictions are correct, and perfect recall indicates no positive data is missed.

Based on the results of the classification using the Support Vector Classification (SVC) algorithm with values of True Positive (TP) = 112, False Positive (FP) = 0, True Negative (TN) = 110, and False Negative (FN) = 0 [20,21]. It can be concluded that the model works very well without making misclassifications. The SVC model showed perfect performance in the classification of the data, characterized by the absence of misclassification (FP = 0 and FN = 0).

Here is the calculation of the evaluation SVC metrics: Accuracy = (TP+TN)/(TP+TN+FP+FN) = (112+110)/(112+110+0+0)=222/222=1.00 or 100%. Precision =

TP/(TP+FP) =112/(112+0)= 1.00 or 100%. Recall = TP / (TP + FN)= 112/(112 + 0) = 1.00 or 100%. F1-Score = 2 × (Precision × Recall)/(Precision + Recall) = 2 × (1 × 1)/(1 + 1)= 2 / 2= 1.00 or 100%.

3.5 Discussion

The performance evaluation of the Support Vector Classifier (SVC), K-Nearest Neighbors (KNN), and Extreme Gradient Boosting (XGBoost) models was conducted based on four key classification metrics: Accuracy, Precision, Recall, and F1-Score and can see at Table 2.

Table 2. A comparison the evaluation of performance

Model	TP	FP	TN	FN	Acc.	Prec.	Rec.	F1-score
SVC	112	0	110	0	1.00	1.00	1.00	1.00
KNN	31	0	110	81	0.64	1.00	0.28	0.43
XGBoost	112	0	110	0	1.00	1.00	1.00	1.00

Table shows the both SVC and XGBoost have similar resulting from different parameter. We convenience for the experiment has successful separating the instance event each model has a different parameters.

The SVC model exhibited optimal classification capability, achieving an Accuracy of 1.00, Precision of 1.00, Recall of 1.00, and an F1-Score of 1.00. This indicates perfect discrimination between the positive and negative classes, with no instances of false positive (FP = 0) or false negative (FN = 0) predictions. The performance is attributable to the SVC's ability to construct an optimal hyperplane in high-dimensional feature space, maximizing the margin between classes and ensuring robust separation even in complex data distributions [22].

Similarly, the XGBoost classifier achieved identical performance to the SVC, with all four metrics reaching 1.00. The absence of misclassifications (TP = 112, FP = 0, TN = 110, FN = 0) suggests that the gradient boosting framework employed by XGBoost efficiently captured the underlying patterns within the dataset. Given these results, both SVC and XGBoost can be regarded as highly reliable models for the classification task under consideration [17]. All performance for KNN, SVM, and XGBoost can be seen at Fig. 7.

In contrast, the KNN classifier demonstrated substantially lower performance, with an Accuracy of 0.64, Precision of 1.00, Recall of 0.28, and an F1-Score of 0.43. Although the model achieved perfect Precision—indicating no false positive classifications—it suffered from a markedly low Recall due to a high number of false negatives (FN = 81) [8]. This suggests that the KNN algorithm, in this configuration, was overly conservative in labeling instances as positive, leading to a substantial proportion of actual positive cases being misclassified as negative . The reduced performance can be attributed to several factors, including sensitivity to the choice of k, the curse of dimensionality, and the influence of unscaled features on distance calculations.

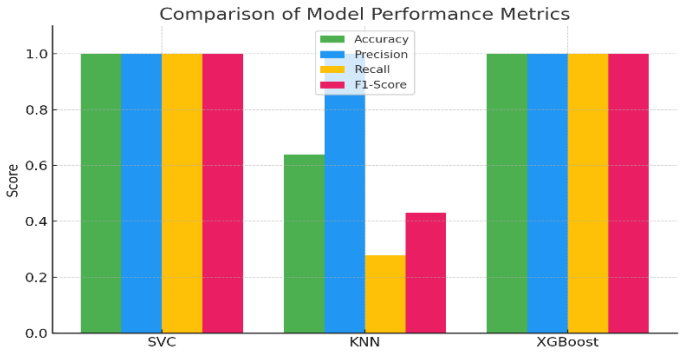


Fig. 7. A comparison chart for every machine learning performance

In contrast, the KNN classifier demonstrated substantially lower performance, with an Accuracy of 0.64, Precision of 1.00, Recall of 0.28, and an F1-Score of 0.43. Although the model achieved perfect Precision—indicating no false positive classifications—it suffered from a markedly low Recall due to a high number of false negatives (FN = 81) [8]. This suggests that the KNN algorithm, in this configuration, was overly conservative in labeling instances as positive, leading to a substantial proportion of actual positive cases being misclassified as negative . The reduced performance can be attributed to several factors, including sensitivity to the choice of k, the curse of dimensionality, and the influence of unscaled features on distance calculations.

The test gives more information about the different kinds of mistakes. A Type I error (false positive) happens when a negative case is wrongly labeled as positive. Both SVC and XGBoost got rid of Type I errors completely in this dataset, as evidenced by FP = 0. KNN also cut down on Type I errors, which gave it a flawless Precision score. A Type II error, or false negative, happens when a positive case is incorrectly labeled as negative. This is where the models start to look different. The KNN model made a lot of Type II mistakes (FN = 81), however SVC and XGBoost did not (FN = 0). The model always failed to detect positive cases, which is why its Recall is so low [23,24].

Overall, the findings indicate that SVC and XGBoost substantially outperform KNN for this dataset. Both high-performing models achieved perfect classification, while KNN struggled to identify the positive class effectively. These results are consistent with prior research, which has shown that distance-based classifiers like KNN are less robust in high-dimensional spaces or when feature distributions are complex, whereas margin-based (SVC) and ensemble-based (XGBoost) approaches are more resilient.

4 Conclusion

Significant differences exist in the classification efficacy of the Support Vector Classifier (SVC), K-Nearest Neighbour (KNN), and XGBoost. BothSVC and XGBoost achieved perfect scores across all evaluation metrics, including accuracy, precision,

recall, and F1-score. The findings demonstrate the ability to circumvent both Type I and Type II errors while maintaining high sensitivity in identifying all positive cases. Their endurance indicates optimal suitability for applications necessitating comprehensive and dependable detection. The KNN model performed notably poorer, particularly regarding sensitivity. Although the precision score was high due to the absence of false positives, the model produced a significant number of false negatives, leading to a markedly low recall value. This indicates that KNN is prone to Type II errors, reducing its effectiveness in scenarios where overlooking affirmative cases presents significant risks. The findings emphasize the importance of considering both accuracy and sensitivity in the evaluation of classification models. Reducing Type II errors is essential in domains including healthcare, fraud detection, and hate speech classification. Consequently, models such as SVC and XGBoost, which offer high precision and exceptional sensitivity, are preferable. In contrast, KNN necessitates extensive testing and may be appropriate only in scenarios with reduced sensitivity.

Future research should focus on evaluating these models on larger and more diverse datasets to see how well they perform under different data distributions and class imbalances. Further research could look into additional techniques for enhancing KNN sensitivity, such as distance-weighted voting, hybrid ensemble methods, or feature selection strategies that minimize the influence of sparsely distributed positive examples. Furthermore, incorporating cost-sensitive learning frameworks may help balance Type I and Type II errors more effectively, ensuring that the models are tailored to application domains where the consequences of misclassification are critical, such as healthcare analytics, cybersecurity, or social media content moderation.

Acknowledgments: This study was funded by Universitas Komputer Indonesia by grant number No.1523/ST/REKTOR/VI/2025.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Tukinah U, Tri Adriyanto A, Mumpuni Hartarini Y, Tinggi Ilmu Ekonomi Semarang S, Semarang U. Social Media Application as a New Paradigm for Business Communication Strategies: The Role of Knowledge, Attitude, and Practices (KAP). JDM (Jurnal Dinamika Manajemen) 2025;16:124–41. https://doi.org/doi.org/10.15294/jdm.v16i1.15534.
2. Amien Ibrahim M, Arifin S, Gusti Agung Anom Yudistira I, Nariswari R, Azis Abdillah A, Pranuta Murnaka N, et al. An Explainable AI Model for Hate Speech Detection on Indonesian Twitter. vol. 16. 2022.
3. Abdusyukur F. Penerapan Algoritma Support Vector Machine (Svm) Untuk Klasifikasi Pencemaran Nama Baik Di Media Sosial Twitter. Komputa : Jurnal Ilmiah Komputer Dan Informatika 2023;12.
4. Peretz O, Koren M, Koren O. Naive Bayes classifier – An ensemble procedure for recall and precision enrichment. Eng Appl Artif Intell 2024;136. https://doi.org/10.1016/j.engappai.2024.108972.

5. Huda Ovirianti N, Zarlis M, Mawengkang H. Support Vector Machine Using A Classification Algorithm. Jurnal Dan Penelitian Teknik Informatika 2022;6. https://doi.org/10.33395/sinkron.v7i3.
6. Rudd MPH JM, Priestley JL, Mph JM, Rudd JM, Lewis Priestley J. A Comparison of Decision Tree with Logistic Regression Model for Prediction of Worst Non-Financial Payment Status in Commercial Credit. vol. 5. 2017.
7. Tavanaei A, Ghodrati M, Reza S. Deep learning in spiking neural networks. Neural Networks 2019;111:47–63. https://doi.org/10.1016/j.neunet.2018.12.002.
8. Rajeswari RP, Juliet K. Text Classification for Student Data Set using Naive Bayes Classifier and KNN Classifier. International Journal of Computer Trends and Technology 2017;43:8–12. https://doi.org/10.14445/22312803/IJCTT-V43P103.
9. Granizo SL, Caraguay ALV, Lopez LIB, Hernandez-Alvarez M. Detection of Possible Illicit Messages Using Natural Language Processing and Computer Vision on Twitter and Linked Websites. IEEE Access 2020;8:44534–46. https://doi.org/10.1109/ACCESS.2020.2976530.
10. Liang M, Niu T. Research on Text Classification Techniques Based on Improved TF-IDF Algorithm and LSTM Inputs. Procedia Comput Sci 2022;208:460–70. https://doi.org/https://doi.org/10.1016/j.procs.2022.10.064.
11. Shamshuzzoha Md, Audry TTB, Alam MdJ, Bhuiyan ZA, Motaharul Islam M, Hassan MM. A novel framework for seasonal affective disorder detection: Comprehensive machine learning analysis using multimodal social media data and SMOTE. Acta Psychol (Amst) 2025;256:105005. https://doi.org/https://doi.org/10.1016/j.actpsy.2025.105005.
12. Indah YM, Aristawidya R, Fitrianto A, Erfiani E, Jumansyah LMRD. Comparison of Random Forest, XGBoost, and LightGBM Methods for the Human Development Index Classification. Jambura Journal of Mathematics 2025;7:14–8. https://doi.org/10.37905/jjom.v7i1.28290.
13. Delibaş E. Efficient TF-IDF method for alignment-free DNA sequence similarity analysis. J Mol Graph Model 2025;137:109011. https://doi.org/https://doi.org/10.1016/j.jmgm.2025.109011.
14. Wan Q, Xu X, Han J. A dimensionality reduction method for large-scale group decision-making using TF-IDF feature similarity and information loss entropy. Appl Soft Comput 2024;150:111039. https://doi.org/https://doi.org/10.1016/j.asoc.2023.111039.
15. Zhou J, Ye Z, Zhang S, Geng Z, Han N, Yang T. Investigating response behavior through TF-IDF and Word2vec text analysis: A case study of PISA 2012 problem-solving process data. Heliyon 2024;10:e35945. https://doi.org/https://doi.org/10.1016/j.heliyon.2024.e35945.
16. Jain DK, Dubey SB, Choubey RK, Sinhal A, Arjaria SK, Jain A, et al. An approach for hyperspectral image classification by optimizing SVM using self organizing map. J Comput Sci 2018;25:252–9. https://doi.org/https://doi.org/10.1016/j.jocs.2017.07.016.
17. Andriansyah D, Eka Wulansari Fridayanthie. Optimization of Support Vector Machine and XGBoost Methods Using Feature Selection to Improve Classification Performance. Journal Of Informatics And Telecommunication Engineering 2023;6:484–93. https://doi.org/10.31289/jite.v6i2.8373.
18. Hatijah M, Shasrini T. Opini Pengguna Twitter Terhadap Upaya Memaafkan Yang Dilakukan Oleh Lesty Kejora Atas Kasus KDRT. Jurnal Ilmiah Wahana Pendidikan, 2024;10(7):729-37. https://doi.org/10.5281/zenodo.11170889.
19. Anderson T, Sarkar S, Kelley R. Analyzing public sentiment on sustainability: A comprehensive review and application of sentiment analysis techniques. Natural Language Processing Journal 2024;8:100097. https://doi.org/https://doi.org/10.1016/j.nlp.2024.100097.
20. Liang J. Confusion matrix: Machine learning. POGIL Activity Clearinghouse 2022;3.

21. Heydarian M, Doyle TE, Samavi R. MLCM: Multi-label confusion matrix. IEEE Access 2022;10:19083–95. https://doi.org/10.3390/s24175752.
22. Ahmad M, Aftab S, Salman Bashir M, Hameed N, Ali I, Nawaz Z. SVM Optimization for Sentiment Analysis. vol. 9. 2018. https://doi.org/10.14569/IJACSA.2018.090455
23. Toleva B, Ivanov I, Hooper V. Feature selection for support vector machines in imbalanced data. Bulletin of Electrical Engineering & Informatics, 2025. https://doi.org/10.11591/eei.v14i4.9556.
24. Halder R. K, Uddin M. N, Ashraf Uddin M, Aryal S, Khraisat A. Enhancing K-Nearest Neighbor Algorithm: A Comprehensive Review and Performance Analysis of Modifications. Journal of Big Data, 2024;11(1):113. https://doi.org/10.1186/s40537-024-00973-y.

Research on the Development Path of Industrial Integration in Health and Wellness Tourism

Zhai Juntian[0009-0001-4189-064X]

Hebei University of Engineering, No. 19 Taiji Road, Handan 056107, Hebei, China

Abstract. Health and wellness tourism is a new form and trend in the development of tourism in the new era, with enormous potential and development space for cross-industry integration with diverse sectors. It not only provides people with a new type of tourism experience but also offers an innovative way to maintain physical and mental health, meeting people's needs for improving their physical and mental health levels. From the perspective of industrial integration, this paper conducts a typical case analysis of health and wellness tourism in China, summarizes and refines the development models of health and wellness tourism, and elaborates on the development paths of industrial integration. The practical and theoretical research on health and wellness tourism is of great significance for promoting local economic development, improving residents' quality of life, and advancing the construction of "Healthy China."

Keywords: Health and wellness tourism, industrial integration, path.

1 Introduction

In recent years, as people's attention to healthy lifestyles continues to increase, the "grand health" industry is accelerating into a new stage of growth. As a product of the integration of tourism and the "grand health" industry, health and wellness tourism relies on a favorable market environment and represents a blue ocean market with huge development potential. By 2023, the scale of China's health and wellness tourism market had approached 90 billion yuan. The *"Healthy China 2030" Planning Outline* has set a clear goal: to reach 16 trillion yuan by 2030. With "Healthy China" officially becoming a core concept of China's development, the health industry has become a key driving force for the development of the service industry under the new normal. The National Tourism Administration has also gradually promoted the standardized development of this field by building national health and wellness tourism demonstration bases. It is predicted that by 2029, the scale of China's health and wellness tourism market will approach 160 billion yuan. In the future, factors such as the upgrading of tourism consumption levels, the innovation and upgrading of intelligent technologies, and the rising demand for health and wellness after the pandemic will become the core driving forces for the development of health and wellness tourism. Internationally, the potential forces promoting the growth of the

health and wellness tourism market remain very strong. Predictions from the Global Wellness Institute (GWI) show that from 2020 to 2025, health and wellness tourism, as one of the three fastest-growing wellness segments in the global wellness industry, will achieve an annual growth rate of over 20%[1].

From the perspective of demand, with the continuous development of China's economy and society and the gradual improvement of people's living standards, the people's demand for health maintenance and physical fitness enhancement is constantly rising. At the same time, China has gradually entered an aging society, and the problem of aging is becoming increasingly serious. Currently, more than 70% of people in China are in a sub - healthy state, 22% are elderly people over 60 years old, and there are more than 400 million middle - class groups pursuing a high - quality life. All these provide a huge market foundation for health and wellness tourism. As a comprehensive way of health maintenance and promotion, health and wellness tourism focuses on the organic integration of tourism with health preservation, rehabilitation and healing. It is conducive to improving people's overall health level, quality of life and quality of life, and is becoming a new engine driving the growth of tourism economy.

From the perspective of supply, health and wellness tourism shoulders the mission of upgrading the tourism industry and promoting the "health-oriented" transformation of destinations. Health and wellness tourism is characterized by long stay duration, high revisit rate, high consumption level, and strong industrial correlation[2]. It serves as a key means to optimize the supply of tourism products, adjust the traditional industrial structure, and enhance the competitiveness of regional health and wellness industries. Tourism enterprises and destinations need to seize the opportunity, shift from solely focusing on economic benefits to balancing multiple benefits such as social benefits, ecological benefits, and people’s health and harmony, build a health and wellness tourism industry chain, and improve the competitiveness and vitality of the regional economy[3].

Health and wellness tourism, generally referred to as medical and health tourism, is a new type of tourism format with the core orientation of maintaining health status and improving quality of life. In essence, it integrates natural ecological resources, leisure experience scenarios and medical and health care services to provide participants with multi-dimensional services covering functional regulation, physical and mental healing, and life quality enhancement. Different from the superficial consumption attribute of traditional sightseeing tourism, health and wellness tourism emphasizes a progressive value creation logic of prevention, healing and improvement. It presents diversified forms such as hot spring health preservation, forest therapy, and TCM (Traditional Chinese Medicine) health therapy, and forms a composite industrial structure with ecological space as the carrier, health services as the core, and cultural experience as the link[4]. The development process of health and wellness tourism relies on in-depth integration and collaborative innovation across industries. Its core lies in breaking the boundary restrictions between traditional industries. There is mutual penetration and cross-integration between tourism and health service industries such as wellness services, wellness sports, and health management, and it also involves the cross-border integration of tourism with the

three major industrial fields of agriculture, industry, and services. Integration does not mean the complete convergence of different industries, but rather forms a certain degree of complementary effect on the original health and wellness tourism formats in the process of integration, building a key support platform for the innovative development and quality upgrading of health and wellness tourism[5]. As an emerging health tourism industry, health and wellness tourism is of great significance for promoting the overall sustainable development of the tourism industry. While meeting the diversified needs of a large number of consumers, it also helps protect the tourism environment, enrich tourism elements, and improve tourism services, thus comprehensively meeting the actual needs of different consumer groups. Vigorously promoting the development of the health and wellness tourism industry can not only promote the transformation and upgrading of traditional tourism models, but also blur the boundaries between different industries and help the horizontal expansion and extension of the tourism industry chain.

2 Analysis of Typical Cases of Health and Wellness Tourism

Three typical cases of health and wellness tourism have been selected, namely Dayu Yashan Tourism Resort in Jiangxi Province, Boao Lecheng International Medical Tourism Pilot Zone in Hainan Province, and Mount Emei. The primary reason for this selection is that they possess significant advantages in aspects such as resource characteristics, industrial models, and market influence. Additionally, data like tourist volume can intuitively reflect the effectiveness of their development.

Dayu Yashan Tourism Resort in Jiangxi Province is located on the world-recognized "golden ecological belt" at 25°N latitude, with a forest coverage rate as high as 92.6%. It holds multiple national-level titles, including National Tourism Resort, National AAAA-Level Tourist Attraction, China's Most Beautiful Village, and National Forest Health and Wellness Base. Yashan is committed to building a world-class forest health and wellness platform and has independently developed a comprehensive green healing system across the entire resort that focuses on "preventing diseases before onset, treating existing diseases, and aiding recovery after illness". Against the backdrop of its natural healing environment, Yashan has introduced advanced international concepts of comprehensive health management, and successively built high-end health and wellness service platforms such as the Yashan Health Care Center, Yashan Hydrotherapy Center, and Bamboo Forest Medicinal Spring Valley. In addition, nearly 50 wellness and leisure programs—including traditional physical therapy, tea ceremony, flower arrangement, incense ceremony, and health-preserving dietary therapy—are organically integrated with ecological vacation services, pioneering a new lifestyle of natural wellness for the body, mind, and spirit during vacations. Yashan has made every effort to promote the high-quality development of ecological tourism. It has gradually improved 12 major projects covering all tourism elements, including catering, accommodation, transportation, sightseeing, entertainment, shopping, business, education, and wellness. It has blazed a path of sustainable ecological tourism development that

integrates rural leisure, forest wellness, mountain vacation, and nature education. Today, it has effectively become a one-stop, all-age leisure and wellness vacation destination. Dayu County thoroughly practices the development concept that "lucid waters and lush mountains are invaluable assets". Based on its unique advantages in location, transportation, and ecology, the county has activated and utilized local historical and cultural resources, optimized the supply of public cultural services for tourists, and strived to build itself into an "ecological backyard" for health and wellness tourism connecting with the Guangdong-Hong Kong-Macao Greater Bay Area, as well as a well-known domestic and international destination for ecological tourism and wellness. In 2024, Dayu County received over 8.014 million tourist visits, with a comprehensive tourism income of 7.965 billion yuan, representing year-on-year growth of 29.14% and 19.05% respectively. From January to August 2025, the county accumulated over 6.0938 million tourist visits (a year-on-year increase of 23.56%), and its comprehensive tourism income reached 5.958 billion yuan (a year-on-year increase of 17.42%)[6].

Boao Lecheng International Medical Tourism Pilot Zone in Hainan is known as the "Second Movement of the Boao Forum for Asia". Approved by the State Council on February 28, 2013, it was granted nine special preferential policies. The zone pilots the development of international medical tourism-related industries, including licensed medical services, medical aesthetics and anti-aging, care and rehabilitation, and health management. It aims to gather high-end international and domestic medical tourism services as well as resources of cutting-edge international medical and pharmaceutical technology achievements, building an international industrial cluster for medical technology services. Catering to the needs of the elderly in medical and health care and other fields, the zone provides characteristic services centered on licensed medical care, health management, and care and rehabilitation. Based on this, it improves the construction of health service facilities and systems, supports the development of diversified health and wellness institutions, promotes the optimization of health insurance industries, and accelerates the process of cross-regional medical expense settlement—all to provide convenient conditions for visiting tourists who aim for disease prevention, treatment, and rehabilitation-based health preservation. Relying on the integrated development model of "medical services, pharmaceuticals, research, industry, and city", the zone advances toward the goal of "building a world-class international medical tourism destination and medical technology innovation platform", and strives to become an important gateway leading the opening-up of the large health sector in the new era. Data shows that in 2024, the number of medical tourism visits reached 413,700, a year-on-year increase of 36.8%. Among these, the per capita consumption of international medical tourism groups exceeded 12,000 yuan, driving the entire industrial chain consumption including physical examinations, rehabilitation, and medical aesthetics. In the first quarter of 2025, medical institutions in the pilot zone received a total of 111,500 medical tourism visits, a year-on-year increase of 29.8%; the number of visits involving licensed medical devices and drugs reached 16,000, a year-on-year increase of 44.14%. To meet the diversified health consumption needs of tourists, during the 5th China International Consumer Products Expo, the zone innovatively launched 5

characteristic medical tourism routes, such as the Cardiac and Cerebral Health Tour and the Silver-Age Longevity Tour, accelerating the integrated development of medical wellness and tourism industries[7]. Up to now, it has established cooperative relationships with more than 180 medical device and pharmaceutical enterprises from 20 countries and regions, introduced a total of 512 types of international innovative medical devices and drugs for "the first use in China", and provided services to over 180,000 Chinese and international patients[8].

Mount Emei is located in the southwestern part of Sichuan Province. As one of the Four Great Buddhist Mountains in China, it is traditionally believed to be the sacred site of Samantabhadra Bodhisattva in Buddhism and boasts a profound cultural heritage. As a key scenic spot in China and a national model of civilized scenic tourism destinations, Mount Emei was inscribed on the UNESCO World Heritage List (for both cultural and natural values) in 1996. In 2007, it was rated as a National AAAAA-Level Scenic Spot by the China National Tourism Administration, and in 2019, it was named one of the "Top 10 National Health and Wellness Destinations" through online voting. It is a typical health and wellness tourism destination that integrates both ecological and cultural attributes. Focusing on the development of health and wellness tourism themed around Buddhist culture, Mount Emei Scenic Area takes its superior natural resources as the foundation, deeply explores Buddhist cultural resources, and has put forward the tourism image positioning of "Graceful Mount Emei, a Sacred Land for Buddhist Wellness". It aims to build a national-level elderly care destination featuring valley landforms. The Mount Emei International Elderly Care Sacred City aligns with Mount Emei's development direction of building an international livable and elderly-friendly city, highlights the unique charm of Mount Emei's Buddhist culture, and takes "valuing life and respecting the elderly, honoring filial piety and worshipping the sage" as its core concept to promote the integrated development of the elderly care industry and the health industry. Mount Emei has built a corridor for in-depth integration of traditional Chinese medicine (TCM), cultural tourism, and health and wellness in the Greater Mount Emei area, successfully becoming a provincial-level medical and health wellness service cluster. At the same time, it has created the "Emei Studies" study tour brand and established a study tourism pattern covering Emei Snow Bud tea culture study tours, intangible cultural heritage study tours of Emei Wushu (martial arts), and the Gaoqiaoli Shanyuexi Nature School. In 2024, the number of tourists received by Mount Emei Scenic Area reached a historic high of 6.21 million, with ticket revenue amounting to 475 million yuan. Among these, inbound tourists numbered 190,000, an increase of 202% year-on-year. The cultural and tourism industry drove the contribution rate of the service sector to economic growth up to 58%, and the total volume of cultural and tourism economy ranked first among prefecture-level cities in Sichuan Province. From January to August 2025, the city received a total of 85.8 million domestic tourists, with tourism consumption reaching 110 billion yuan, increasing by over 20% and 18% year-on-year respectively; the number of inbound tourists increased by 35.25% year-on-year[9].

3 Analysis of the Basic Modes of Current Health and Wellness Tourism

Health and wellness tourism products are becoming increasingly diversified, forming a multi-level market supply system. This paper classifies health and wellness tourism products into four main types: ecological resource-based, medical technology-based, sports and fitness-based, and cultural experience-based.

Ecological health-preserving wellness tourism mainly relies on its superior natural resources such as forests, hot springs, and oceans. Based on ecological resources, it expands and designs in-depth experiential leisure vacation tourism products, builds wellness tourism bases, and carries out wellness tourism activities themed around health preservation and fitness. First, forest wellness. Relying on mountains and woodlands, it focuses on themes such as forest bathing, forest meditation, and forest yoga. It takes the research and development of wellness and healthcare pharmaceuticals as an industrial extension direction, constructs an industry chain development model integrating medical care and health preservation, builds national-level forest wellness tourism resorts with wellness functions, and forms a comprehensive forest wellness tourism brand. Second, hot spring therapy. Leveraging hot spring resources, it gives full play to their wellness functions, deeply explores the inherent concepts of hot spring wellness, integrates modern wellness technology means, develops diversified hot spring wellness projects, and promotes the comprehensive development and utilization of hot spring resources. Third, coastal recuperation. Depending on water resources and combined with modern wellness technologies, it builds facilities such as coastal river section wellness centers, dietary therapy areas, and leisure fishing gardens, exerting the special effects of water-based wellness resources in nourishing the heart, body, and spirit. Fourth, ecological rural wellness. It integrates rural areas, villages, and the surrounding natural environment to construct an organically integrated ecological wellness tourism space, and carries out tourism activities such as rural sightseeing, farming experience, and rural cultural leisure.

Medical technology-based health and wellness tourism mainly relies on the research, development and application of advanced medical technologies, links with mid-to-high-end medical and wellness resources, and precisely targets domestic and foreign customer groups to enhance market appeal, thereby building professional medical and wellness bases. First, developing traditional Chinese medicine (TCM) health therapy. On the premise of leveraging China’s existing medical resources, it builds medical-nursing integrated residential bases, integrates TCM resources, ethnic medical therapies with longevity and blessing cultures, creates elderly care tourism towns and hot spring health resorts, and develops wellness tourism products featuring TCM health preservation and ethnic medical diagnosis and treatment systems, while improving the medical prevention and treatment service network. Second, expanding the wellness medical market. By analyzing the physical and mental needs of domestic wellness tourism groups, it taps into expert-level medical resources, establishes remote collaboration channels with core medical institutions, introduces core medical

institutions to set up branch bases, and develops high-end wellness medical tourism products in a targeted manner.

Cultural nourishment-based health and wellness tourism mainly focuses on fully exploring the cultural resources and historical heritage of rural areas, and creating rural wellness tourism products centered on cultural experience, cultural inheritance, and cultural health preservation. This model aims to enable tourists to achieve both physical and mental wellness by integrating into and participating in the leisurely local life through cultural edification and experience, combined with the region's basic tourism facilities that meet tourists' needs for food, accommodation, transportation, and entertainment.

Sports and fitness-based health and wellness tourism mainly relies on the regional terrain and topography such as valleys and woodlands that are suitable for carrying out outdoor sports. It can provide tourists with good sports resources such as sports venues and sports routes. At the same time, with supporting leisure and health-preserving facilities and entertainment projects as auxiliary, it achieves the tourism purpose of promoting tourists' physical health. The characteristic of this type of health and wellness tourism industry is that outdoor sports are the main form. Generally, the main consumers are tourists who are in good physical and mental health and have high pursuit of quality of life and health status.

4 Industrial Integration Development Paths

When the development of a single industry hits a bottleneck, the industrial integration development path of "Health and Wellness + Tourism +" and even "+N" should be adopted. For coordinated development between industries, it is necessary to ensure the circulation of various factors, give full play to the complementary advantages and integration and collaboration potential of different industries, activate the vitality of industrial development, and promote the continuous progress of the health and wellness tourism industry relying on new models and new products.

4.1 The Primary Industry

The integration of health and wellness tourism with the primary industry has formed the "Health and Wellness + Tourism + Agriculture" development model. In-depth exploration and rational development of agricultural resources can endow health and wellness tourism with unique charm and promote the industrial integration to move toward a higher level. Building Agricultural Health and Wellness Complexes. With agricultural resources such as farmland, orchards and tea gardens as carriers, agricultural health and wellness complexes are built to integrate functions including agricultural sightseeing, leisure vacation and health-preserving elderly care. By providing programs like agricultural product picking, agricultural science popularization and farming experience, parents and children are encouraged to participate in farming work together. In the process of getting close to nature, parent-child bonds are strengthened; meanwhile, children are taught agricultural knowledge to cultivate their interest in the agricultural field. Developing Characteristic Agricultural Health and Wellness Products. Based on local agricultural characteristics

and health and wellness needs, a series of agricultural health and wellness products with local features are developed. Agricultural products with health-preserving and healthcare functions are cultivated, and tailored health-preserving agricultural product packages are prepared according to the health needs of different groups: nutrient-rich and easy-to-digest agricultural product combinations are customized for the elderly to support their health maintenance; portable agricultural product snacks that can quickly replenish energy are prepared for working professionals[10].

Creating a Health and Wellness Environment. The deep connection between tourists and nature is strengthened by stimulating their "five senses". The visual experience primarily creates visual aesthetics in the ecological environment through elements such as color, form, and space, thereby achieving the health and wellness effect of physical and mental relaxation; The auditory experience involves the design of natural sounds and artificial sounds, enhancing tourists' immersive experience; The olfactory experience creates a pleasant and comfortable environment by properly arranging aromatic plants or using natural aromatherapy; The gustatory experience mainly emphasizes the natural properties and nutritional value of food ingredients, deepening the connection with the health and wellness environment through tasting; The tactile experience involves direct contact between the human body and the natural environment, allowing one to fully obtain the healing effect of integrating the body and mind with nature[11]. Integration and Development of Folk Customs. Folk customs and agricultural resources are explored and integrated into health and wellness tourism programs. Traditional farming techniques such as manual rice transplanting and ancient winemaking methods are demonstrated, allowing tourists to personally experience ancient agricultural production methods and appreciate the wisdom of their ancestors. Agriculture-themed study tour courses are developed, and agricultural experts and folk culture scholars are invited to explain local agricultural ecological knowledge and the history of folk culture to tourists. This enhances tourists' in-depth understanding and recognition of agricultural ecological resources.

4.2 The Secondary Industry

The integration of health and wellness tourism with the secondary industry has formed the "health + tourism + manufacturing" development model. Enhancing the R&D capabilities and added value of rural agricultural product industrial chains provides crucial support for the quality upgrading and experience innovation of health and wellness tourism. Upgrading the Manufacturing of Health and Wellness Accommodation Facilities. To address accommodation needs in health and wellness tourism, the manufacturing industry can focus on optimizing the health-preserving functions of residential travel facilities. Develop modular health and wellness vacation cabins, integrating intelligent health equipment inside, such as smart mattresses for monitoring sleep quality, eco-friendly air conditioning systems for regulating indoor temperature and humidity, or fresh air devices with purification functions, to create a healthy and comfortable living environment for tourists. Additionally, produce mobile health and wellness caravans, integrated with small kitchens or portable physiotherapy instruments, to meet tourists' personalized needs of "nurturing while traveling."Specialized Production of Health and Wellness Equipment

and Supplies Centering on tourists' health management needs during health and wellness tourism, develop the health and wellness equipment manufacturing industry. Produce portable health monitoring devices that allow tourists to track physical indicators in real-time, with data synchronized to cloud-based health management platforms, providing data support for personalized health and wellness plans. Develop lightweight physiotherapy instruments based on traditional Chinese medicine health preservation concepts, and produce special health and wellness tourism supplies such as UV-protective health clothing, mosquito-repellent health sachets, and natural herbal cleansing and care set, integrating health-preserving functions into daily necessities to enhance the health attributes of travel experiences. Integration of Smart Health and Wellness Equipment Build a smart health and wellness tourism service system with intelligent manufacturing technologies. Create interactive health and wellness experience equipment, such as VR rural meditation devices that restore rural landscapes through virtual scenes, paired with sound and fragrance systems to help tourists relax physically and mentally; intelligent guide robots provide tourists with route guidance, popularization of health and wellness knowledge, and emergency assistance services[12]. Standardized Production of Health and Wellness Tourism Supplies Establish a standardized production system for health and wellness tourism supplies to improve product quality and safety. Formulate industry standards for health and wellness products such as accommodation facilities, physiotherapy instruments, and food packaging, regulating material selection, production processes, and quality inspection links. Standardized production reduces enterprise costs, provides tourists with safe and reliable health and wellness products, enhances brand credibility, and promotes the standardized development of the health and wellness tourism manufacturing industry.

4.3 The Tertiary Industry

The integration of health and wellness tourism with the tertiary industry has given rise to development models such as "health + tourism + culture," "health + tourism + sports," and "health + tourism + finance." Building diverse and characteristic rural service industries based on resources is an important way to enrich both tourists and local industries through tourism. Integration of Culture and Tourism. Strengthen the integration of characteristic cultural types such as farming culture, historical culture, folk culture, red culture, and intangible cultural heritage with the health and wellness tourism industry. Develop diverse products such as rural cultural performances, rural cultural exhibitions, and rural cultural experiences. For example, create cultural health-themed parks by integrating traditional cultural elements like calligraphy, painting, and tea ceremony into health and wellness scenarios. Set up calligraphy and painting studios as well as tea ceremony experience halls, where tourists can relax physically and mentally through ink-wash creation and tea art appreciation under the guidance of professional teachers. Through resource integration and characteristic development, form regionally iconic brands to enhance market competitiveness. Integration of Sports and Tourism. Combine natural environments with scientific concepts to design diversified programs for improving tourists’ physical fitness. Develop specialized health and wellness sports for different groups: offer traditional

health-preserving exercise courses such as Tai Chi and Baduanjin for middle-aged and elderly groups; launch functional training programs integrating sports rehabilitation concepts for sub-healthy groups. Introduce emerging sports health formats using water environments, such as water Tai Chi and paddleboard yoga, to enhance relaxation effects. Build sports health camps equipped with motion monitoring devices to record tourists' exercise data in real-time and provide personalized health advice.

Integration of Finance and Tourism. Leverage the support of credit policies to boost the development of small and micro tourism enterprises and family-style rural tourism. Further improve financial infrastructure in scenic areas to enhance the convenience of financial services. Establish exclusive platforms for property rights transactions of tourism projects to broaden financing channels for tourism initiatives. Promote the innovation and improvement of financial payment systems: popularize diversified electronic payment tools such as credit cards, e-cash, and e-checks; standardize certification standards and processes for electronic money; strengthen the service efficiency of bank cards in tourism scenarios, thereby facilitating tourists' travel consumption and payment activities. Integration of Medical Care and Tourism. Promote close cooperation between rural health and wellness institutions, medical rehabilitation institutions, and health-preserving institutions to jointly build demonstration bases for the integration of medical care and elderly care. Develop emerging formats of medical and health tourism, such as health check-up tourism and medical rehabilitation tourism, and provide health services including preventive healthcare (treating diseases before onset) and chronic disease management to meet the health needs of different consumer groups. Design traditional Chinese Medicine (TCM) physical therapy products for sub-healthy groups, supplemented by services such as tuina (Chinese therapeutic massage) and TCM dietary regulation; develop health monitoring and residential elderly care products for the elderly group, conduct regular health assessments, and provide customized health and wellness plans[13].

5 Conclusion

This paper systematically analyzes the typical cases, basic models, and development paths of health and wellness tourism from the perspective of industrial integration. The research shows that the development of health and wellness tourism requires in-depth activation of cross-industry synergy efficiency. Only by constructing differentiated health and wellness tourism industrial chains according to the resource characteristics and market demands of different regions can more comprehensive and diversified healthy travel experiences be provided to tourists. With the strengthening of public health concepts and the upgrading of consumption levels, the growth potential of this field will continue to be released.

References

[1] The Global Wellness Economy: Looking Beyond COVID, https://global wellness institute.org/industry-research/the-global-wellness-economy-looking-beyond-covid/, last accessed 2025/8/10
[2] Zhu Dongfang, Zhong Linsheng, Yu Hu: A comparison and prospect of health and wellness tourism research at home and abroad. World Regional Studies 32(11), 167-180 (2023)
[3] Yao Yanbo, Li Zhengli, Zhang Yan: Research characteristics, key fields and future prospects of domestic health and wellness tourism. Future and Development 48(05), 19-26 (2024)
[4] Zeng Hanchao: Drive and transformation: Driving mechanisms and innovative paths for the sustainable development of health and wellness tourism in regional economy. Reform and Strategy 41(01), 230-233(2025)
[5] Ni Minghui: Study on the cross-boundary integration model of health and wellness tourism industry in ethnic areas of Heilongjiang Province. Heilongjiang National Series (02), 82-89(2022)
[6] Ganzhou Municipal People's Government, https://www.ganzhou.gov.cn/gzszf, last accessed 2025/9/28
[7] People's Daily Online, https://cpc.people.com.cn/, last accessed 2025/7/22
[8] Hainan Daily, https://www.hainan.gov.cn/hainan/sxian/, last accessed 2025/9/20
[9] State Council Information Office of the People's Republic of China, http://www.scio.gov.cn/xwfb/dfxwfb/gssfbh, last accessed 2025/9/3
[10] Wang Yue: Study on the innovative model of integrated development of comprehensive tourism and health and wellness industry in Qingyuan City. Business Exhibition Economy (02), 38-41(2025)
[11] Li Xia, Luo Chunyu, Yang Jinlin: Design of forest health and wellness tourism products based on tourists' "five senses"—A case study of Chinese Fir King Forest Health and Wellness Base in Yanping District, Nanping City. Western Tourism (06), 80-82+86(2025)
[12] Xue Xin: Development strategies of rural health and wellness tourism for the elderly under the background of rural revitalization. Rural Economy and Science-Technology 36(02), 104-107(2025)
[13] Huo Ning: Study on the high-quality development of health and wellness tourism industry empowered by digital economy. Business Exhibition Economy (08), 32-35 (2025)

A Human-Centric Decision Support Framework for Satisfaction Enhancement in Medical Staff Scheduling

Pavinee Rerkjirattikal[1][0000−0001−8496−4197], Raveekiat Singhaphandu[2][0000−0002−7603−5504], Charnon Pattiyanon[2][0000−0003−3660−2962], and SangGyu Nam[3][0000−0002−7424−8469]

[1] Department of Technology and Operations Management, Faculty of Business Administration, Kasetsart University, Bangkok 10900, Thailand
pavinee.re@ku.th

[2] Artificial Intelligence and Computer Engineering Program Department, CMKL University, Bangkok 10520, Thailand
{raveekiat,charnon}@cmkl.ac.th

[3] School of Information, Computer, and Communication Technology, Sirindhorn International Institute of Technology, Thammasat University, Pathum Thani 12120, Thailand.
sanggyu@siit.tu.ac.th

Abstract. Medical staff scheduling is a complex problem that requires balancing operational efficiency, regulatory compliance, and staff well-being. While many optimization-based algorithms can generate high-quality rosters, their adoption in practice is limited by the absence of robust, user-friendly decision-support systems (DSS) that integrate seamlessly into hospital workflows. This paper presents a human-centric, human-in-the-loop DSS framework that integrates conversational interfaces, natural language understanding (NLU), and mathematical optimization to produce fair and feasible schedules that align with institutional requirements and individual needs. Staff can express scheduling preferences in natural language through familiar chat-based interfaces, which are automatically parsed into structured inputs for the optimization engine. Administrators remain active decision-makers, able to adjust parameters and constraint weights to ensure transparency, flexibility, and collaborative refinement. A preliminary evaluation of the NLU component on simulated clean, noisy, and code-switched messages achieved macro-F1 scores above 0.94, demonstrating reliable extraction of diverse preference types. This conceptual framework represents a practical pathway for translating advanced scheduling algorithms into deployable healthcare solutions, with the potential to enhance scheduling efficiency, fairness, and staff satisfaction in real-world settings.

Keywords: Medical Staff Scheduling · Decision-Support System · Artificial Intelligence · Human Factors · Optimization

1 Introduction

Healthcare institutions face persistent challenges in staff scheduling due to rising patient demand, limited resources, and the operational complexity of delivering around-the-clock care. These pressures can lead to unbalanced workloads, fatigue, and increased job stress among medical personnel. Rigid schedules that overlook staff well-being and individual preferences can contribute to burnout, absenteeism, reduced morale, and higher turnover rates [1], further resulting in workforce shortages in a self-reinforcing cycle.

In many hospitals, scheduling is still handled manually by department administrators, such as head nurses or physicians. Manual methods are time-consuming, prone to inconsistencies, and make it challenging to balance shift coverage, staff preferences, labor regulations, and fairness [2]. Consequently, preferences and equity are often sacrificed in favor of cost efficiency and staffing requirements, increasing the administrative burden and perceived inequity among personnel.

Over the past decades, various algorithmic and optimization approaches have been developed to improve healthcare staff scheduling. These models address operational constraints, regulatory compliance, and fairness, yet many remain confined to theoretical settings or small-scale trials. As Petrovic [3] notes, the absence of robust, user-friendly decision-support systems (DSS) that integrate seamlessly with hospital workflows is a key barrier to adoption. Systems requiring technical expertise, coding skills, or complex interfaces are inaccessible to most healthcare staff and administrators.

To address this gap, we propose a human-centric decision-support framework that integrates artificial intelligence and mathematical optimization for healthcare staff scheduling. The system emphasizes fairness, staff preferences, and broader human factors alongside operational efficiency. Preferences may include shift or day-off requests, while fairness ensures balanced workload distribution and allocation of both undesirable and desirable shifts. Human factors encompass personal circumstances, health-related issues, and other individual needs that affect scheduling. At the same time, the system ensures compliance with staffing and skill requirements while maintaining cost efficiency.

The proposed DSS enables staff to express scheduling needs in plain language through a chatbot-style interface embedded in familiar messaging platforms. These inputs are processed using natural language understanding (NLU) to extract structured data, which is then passed to an optimization engine that produces fair, efficient, and policy-compliant schedules. By combining powerful optimization algorithms with an intuitive interface, this human-centric DSS aims to bridge the gap between sophisticated scheduling theory and practical, deployable solutions in healthcare environments that can be realistically implemented in busy hospital settings without adding administrative complexity.

The rest of this paper is organized as follows. Section 2 reviews the literature on healthcare staff scheduling, including key factors, commonly used approaches, and identified research gaps. Section 3 introduces and explains the proposed framework for an intelligent decision-support system. Section 4 outlines

preliminary experiments and results of NLU on synthesized datasets. Section 5 concludes the paper and outlines future development directions.

2 Literature Review

Medical personnel scheduling, covering roles such as doctors, nurses, pharmacists, paramedics, and other healthcare staff, is a complex task requiring continuous 24/7 coverage while satisfying diverse constraints, including hospital labor laws, contractual agreements, skill-mix requirements, and staffing levels. A high-quality schedule must balance conflicting stakeholder priorities: medical personnel seek work–life balance, sufficient rest allowance, and well-being; managers prioritize operational performance and cost-effectiveness; and patients expect safe, empathetic, and high-quality care [4]. Achieving this balance requires integrating operational constraints with individual preferences, making the problem too complex for manual management. In addition, manual scheduling has been shown to be inefficient, error-prone, and a contributor to job dissatisfaction among personnel [5].

To address these scheduling challenges, researchers have employed a range of problem-solving approaches, including exact optimization methods, such as mixed-integer programming (MIP), constraint programming, and goal programming, as well as metaheuristics like genetic algorithms and tabu search. Hybrid approaches that combine mathematical programming with heuristic search have also been developed to improve scalability and computational efficiency. Traditionally, these scheduling models have prioritized coverage, regulatory compliance, and cost efficiency [6–8]. These objectives are often achieved at the expense of fulfilling staff preferences and satisfaction [9].

In recent years, there has been a notable shift toward human-centric scheduling, driven by persistent workforce shortages and the recognition that work–life balance and job satisfaction are crucial for employee retention. From a human factors perspective, accommodating individual preferences allows staff to align work schedules with personal needs and lifestyles, which contributes to improved job satisfaction and retention. Studies have shown that optimization-based scheduling can integrate preferences while maintaining feasibility in terms of staffing needs and service quality [10–12]. Fairness has also emerged as a key quality metric, measured through workload balance, equitable distribution of night/weekend shifts, and balanced allocation of desirable shifts and days off. Although fairness is inherently subjective and may vary across hospital policies and cultural contexts, Gerlach et al. [13] highlight that nurses value being involved in defining fairness rather than accepting top-down metrics. Staff may hold differing views on what fairness means. Consequently, effective scheduling systems should avoid hard-coding a single fairness objective and instead allow hospital units to co-design criteria, such as weighting workload balance against preference satisfaction.

Numerous studies have integrated preferences and fairness into medical personnel scheduling in diverse contexts. Rerkjirattikal et al. [14] developed a goal

programming model for a large public hospital in Thailand's emergency department, ensuring equitable workload distribution and fair allocation of preferred shifts and days off. This model was later extended to incorporate cost-effectiveness alongside human factor objectives [15]. Yasmine et al. [16] developed a mathematical model for a French hospital to balance workloads and shifts while aligning with nurses' preferences. Burke et al. [17] proposed a preference-based integer programming model for radiologists, respecting preferred sections and work locations while ensuring fairness and operational compliance. Narlı and Derse [18] introduced an integrated MILP model for a pediatric intensive care unit (PICU), jointly optimizing schedules for nurses, doctors, and caregivers, with objectives focused on minimizing costs and achieving workload balance.

Although these studies represent significant progress, many remain confined to theoretical approaches with limited real-world implementation. A primary barrier is the lack of DSS that are both robust and accessible to non-technical users while integrating seamlessly into hospital workflows. Notable attempts to address this include Koruca et al. [19], who developed a DSS that incorporates institutional constraints and individual preferences, allowing administrators to choose from multiple optimization algorithms, such as those prioritizing seniority or maximizing fairness scores. Uhde et al. [20] designed a tablet-based self-scheduling system in which staff submit preferences, view shifts in a shared calendar, and resolve overlapping requests collaboratively; the system was iteratively refined through user feedback to align closely with real-world hospital practices.

Despite these advancements, there remains a need for solutions that integrate usability, human-centric design, and human-in-the-loop decision-making. A recent review by Abdullah et al. [21] emphasizes the importance of intuitive interfaces and comprehensive decision-support features that integrate smoothly with existing hospital systems, enabling straightforward schedule creation, adjustment, and visualization. Our proposed framework addresses this gap by combining a chatbot-based interaction model, an NLU preference extraction, and a human-centric optimization scheduling model within a human-in-the-loop workflow. The framework outlines a novel and practical pathway for translating advanced scheduling algorithms into real-world healthcare settings.

3 Proposed Framework

The proposed framework combines a Conversational User Interface (CUI), NLU, and an optimization engine to deliver a human-centric, human-in-the-loop scheduling process for medical personnel. The system is designed to ease the manual burden on administrators while providing a convenient way for staff to express their scheduling preferences in natural language via familiar chat-based interfaces. After preferences are interpreted by the NLU and consolidated, along with past fairness scores, preference satisfaction history, and constraints specified by the administrator as additional considerations, the optimization engine generates feasible master schedules. The system then seamlessly outputs individ-

ualized work schedules that are fair and reasonably aligned with staff preferences, all within the same conversational platform.

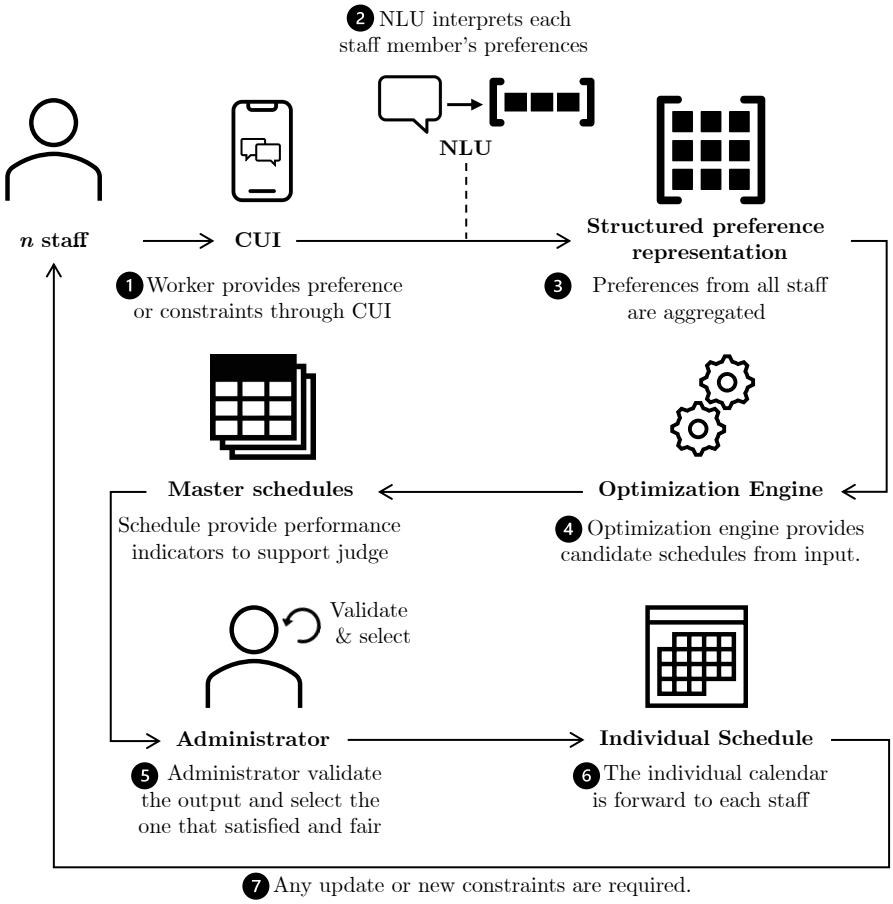


Fig. 1. Proposed human-centric agentic AI scheduling framework.

To illustrate, consider a hypothetical monthly scheduling cycle for a hospital ward with n nurses working in rotating shifts, with rosters generated at the end of each month for the upcoming period. Using this example, our framework consists of the following steps, as depicted in Fig. 1.

1. Preference submission via CUI: Each of the n nurses interacts with the system through a conversational interface (e.g., chatbots on Line or Telegram) to submit their scheduling preferences, such as unavailable days, preferred shift types, or maximum weekly hours. The CUI guides the interaction, prompting the nurse to express preferences and any specific constraints in

natural language. For example:

> ***Nurse A:*** I can't do night shifts in the first week of the month, and I prefer morning shifts after the 15th. I will be on leave on the 20^{th} and 21^{st}.

The system uses guided follow-up questions only when necessary to resolve ambiguities, such as unclear date ranges or affirming shift counts, thereby reducing cognitive load and encouraging full completion of preference submissions. An example of CUI is illustrated in the Appendix.

2. Preference interpretation via NLU: The system's NLU module processes each nurse's text input and converts it into a structured, machine-readable record representing the individual's preferences. For example, Nurse A's input is interpreted as unavailability for night shifts in Week 1, a preference for morning shifts from the 15^{th} to the 31^{st}, and planned leave on the 20^{th} and 21^{st}. The system then provides a readback for confirmation, enabling the nurse to verify or correct the information before finalization.

> ***Agent:*** You are unavailable for night shifts in Week 1, prefer morning shifts after the 15^{th}, and will be on leave on the 20^{th} and 21^{st}. Is this correct?

The nurse can confirm or make corrections in natural language, after which the NLU reprocesses the updated input.

3. Aggregation of preference: All individual preference records are combined into an $n \times f$ matrix, where n is the number of nurses and f is the number of preference attributes. This matrix represents all submitted preferences and is one of the key inputs for the downstream optimization process.
4. Optimization engine: The optimization engine receives as inputs the aggregated preference matrix, historical preference satisfaction records, hospital regulations, and administrator-defined parameters. Depending on the problem's size and complexity, the system may employ integer programming, constraint-based scheduling, or evolutionary algorithms. In the optimization model, hospital regulations and legal requirements are treated as hard constraints, while individual preferences and fairness measures are modeled as soft constraints. To encourage a human-in-the-loop process, the administrator can adjust the relative weight of each soft constraint or temporarily deactivate/activate them for specific scheduling cycles. The engine generates m feasible master schedules that comply with institutional rules and fulfill as many preferences as possible, while distributing workload and preferred assignments fairly across staff.

   Because concepts such as "preference satisfaction" and "fairness" vary across hospitals, the optimization engine is designed to allow these objectives to be revised to reflect institutional policies and staff expectations. For example, in Rerkjirattikal et al. [15], nurses assigned preference levels to shifts and

days off (e.g., high = 3, or low = 1). A nurse's total preference score is the sum of preference values multiplied by the corresponding shift allocations, and fairness was assessed by standard deviations in both preference scores and workload across staff. Implementation requires collaboration with hospitals to determine how preference and fairness should be operationalized. Finally, previous periods' assignments are considered, allowing the model to proactively adjust allocations and correct historical imbalances over time.

5. Master schedule selection: The administrator reviews summary statistics (e.g., fairness scores, preference satisfaction, coverage rate) of the m feasible schedule options supplied through a user-friendly interface or dashboard. The administrator then selects the most suitable master schedule, potentially making trade-offs or adjustments based on managerial judgment. An example of an administrator dashboard is illustrated in the Appendix.
6. Distribution of schedule: Once a master schedule is chosen, it is decomposed into individual schedules and dispatched to each nurse through the same CUI. The output is personalized and easy to interpret (e.g., shown as a calendar or plain-text summary). Depending on organizational policy and privacy considerations, the master schedule may also be made accessible, in part or in full, to allow staff to coordinate potential shift swaps if permitted.
7. Updates and revisions: After schedules are released, nurses or administrators may report changes such as sudden unavailability, voluntary shift swaps, or updated institutional policies. These changes are resubmitted through the CUI and reprocessed by the NLU into a structured form. Based on the type and urgency of the update, the system can either trigger targeted re-optimization (affecting only the impacted shifts) or flag the issue for manual review by the administrator. This ensures that disruptions are addressed promptly while minimizing unnecessary changes to unaffected parts of the schedule.

4 Preliminary Evaluation of NLU Component

As a preliminary step, we conducted a simulation-based evaluation to assess the ability of the proposed framework's NLU component to extract structured scheduling preferences from free-form natural language messages under different linguistic conditions.

4.1 Setup

A Python-based synthetic data generation pipeline was developed to produce paired *ground-truth preference records* and corresponding *chat-style messages*. Each ground-truth record consists of three fields:

- Day-off: Specific dates on which the staff member is unavailable.
- Preferred shifts: Shift types the staff member prefers.
- Avoid shifts: Shift types the staff member wishes to avoid.

From each record, we generated between one and three sentences according to one of the following linguistic conditions:

- Clean: Grammatically correct English messages
- Noisy: English messages with typographical errors and informal shorthand
- Code-switch: Messages mixing English and Thai, with shift preferences and avoidance expressed in Thai, while other elements remained in English

These conditions were chosen to evaluate the NLU's robustness in processing messages that staff might naturally produce in a real scheduling context, including standard formal input, informal or casually typed messages with spelling variations, and code-switched messages that reflect the common practice of mixing languages in text-based communication. A total of 200 scenarios were synthesized, evenly split across the three conditions. Messages were processed by the NLU extraction pipeline implemented using the OpenAI GPT-4o-mini model. The prompt instructed the model to extract the three preference fields in JSON format. Extracted outputs were compared against the ground truth.

4.2 Evaluation Results, and Discussion

To evaluate the performance of the NLU module, we computed the set-overlap F1 score for each preference field (day-off, prefer, avoid), with the macro-F1 score calculated as the average of these three values. All fields were evaluated as unordered sets. Table 1 summarizes the extraction accuracy across conditions. The NLU pipeline maintained high performance across all input types, with macro-F1 scores exceeding 0.94 even in the noisy and code-switch cases.

Table 1. F1 scores of NLU extraction under different linguistic conditions. All values are macro-level set-overlap F1 scores.

Condition	Day-off	Prefer	Avoid	Macro
Clean	0.968	1.000	0.984	0.984
Noisy	0.925	1.000	0.955	0.960
Code-switch	0.890	1.000	0.943	0.944

These preliminary results indicate that the NLU module can accurately extract structured scheduling preferences from diverse message types, including those with typographical noise and bilingual content. Nonetheless, synthetic messages cannot fully capture the ambiguity, contextual cues, and variability of real-world communication among healthcare staff. Moreover, the current evaluation measured only extraction correctness, without assessing interaction efficiency or user satisfaction.

An error analysis of 21 mispredicted cases revealed three main failure types, with representative examples shown in Table 2.

(i) Missed day-off extraction, especially when dates were written as "N off next month" and misinterpreted as a quantity rather than a specific day (18 cases).
(ii) Spurious avoid-shift entries, typically triggered by bilingual conjunctions in code-switch scenarios that confused the parsing logic (5 cases).
(iii) Overgeneration of date ranges from a single numeric reference (1 case).

Table 2. Representative error cases from NLU extraction. Messages originally written in Thai have been translated into English to avoid printing issues.

	Message	Ground Truth	NLU Output	Error Type
1	I need 5 off next month. If I could get night shifts, that would be great.	`off:[2025-09-05], pref:[night], avoid:[]`	`off:[], pref:[night], avoid:[]`	Missed date extraction
2	If I could get night and morning shifts, that would be great.	`off:[], pref:[night, morning], avoid:[]`	`off:[], pref:[morning, night], avoid:[afternoon]`	Spurious avoid entry
3	I need 5 off next month. If I could get afternoon shifts, that would be great.	`off:[2025-09-05], pref:[afternoon], avoid:[]`	`off:[2025-09-05 to 2025-09-09], pref:[afternoon], avoid:[]`	Overgenerated date range
4	I prefer morning and night shifts.	`off:[], pref:[morning, night], avoid:[]`	`off:[], pref:[morning, night], avoid:[afternoon]`	Spurious avoid entry
5	I need 2 off next month. I prefer morning and afternoon shifts. Please avoid night shifts for me.	`off:[2025-09-02], pref:[morning, afternoon], avoid:[night]`	`off:[], pref:[morning, afternoon], avoid:[night]`	Missed date extraction

Extraction errors were disproportionately observed in code-switch scenarios, indicating that mixed-language structures increase the complexity of parsing temporal expressions. In contrast, shift-type recognition remained generally robust, even in noisy or mixed-language inputs. These findings highlight the need for further refinement of NLU in disambiguating numeric date references, reducing the overgeneration of avoid lists, and improving preference extraction in mixed-language contexts. Future work will involve in-hospital trials with actual healthcare staff, incorporating broader evaluation metrics such as interaction time, clarification request frequency, and perceived usability across multiple scheduling cycles. This stage will also require formal ethical approval in line with institutional and national guidelines.

5 Conclusion

Medical staff scheduling is a complex, multi-constraint problem that requires a careful balance between regulatory compliance, operational efficiency, and staff

well-being, while ensuring uninterrupted and high-quality patient care. Although numerous advanced algorithms can generate optimal or near-optimal schedules, their practical adoption remains limited by the lack of integration with hospital workflows and the limited technical expertise of end users.

This study proposes a human-centric, human-in-the-loop decision support framework designed to bridge this gap. The framework integrates advanced scheduling algorithms into an intuitive, conversational interface supported by robust NLU. Staff can express their preferences in natural language via CUI. At the same time, administrators can adjust constraint weights, toggle soft constraints, and apply expert judgment to finalize schedules. The optimization engine processes preferences and operational constraints to generate fair, feasible schedules that align well with both institutional requirements and staff needs.

A preliminary evaluation of the NLU component, using simulated clean, noisy, and code-switched messages, achieved macro-F1 scores above 0.94 across all linguistic conditions, demonstrating the approach's capability to accurately extract day-off requests, preferred shifts, and avoidances from diverse inputs. However, the evaluation is currently limited to synthetic data and focuses solely on the accuracy of extraction.

Future work will involve securing ethical approval for trials with healthcare staff in real-world settings and will expand the evaluation to include interaction efficiency, user satisfaction, and long-term adoption. Insights from these studies will guide iterative refinements to the NLU and overall framework. The ultimate goal is to deliver a deployable, operationally effective DSS that enhances scheduling efficiency while actively promoting workforce well-being, sustainability, and retention by creating safe, fair, and satisfactory work schedules.

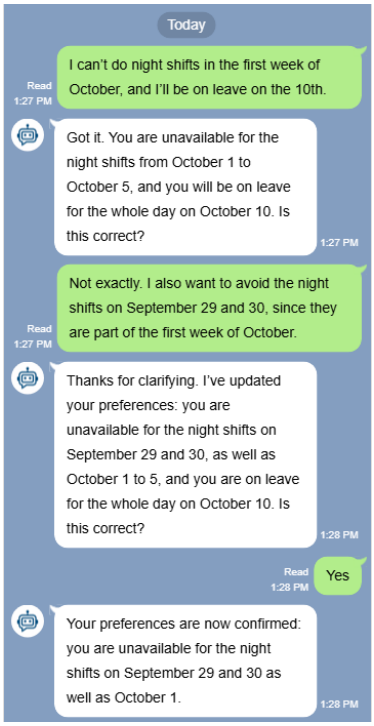
Disclosure of Interests. The authors declare that they have no conflict of interest.

References

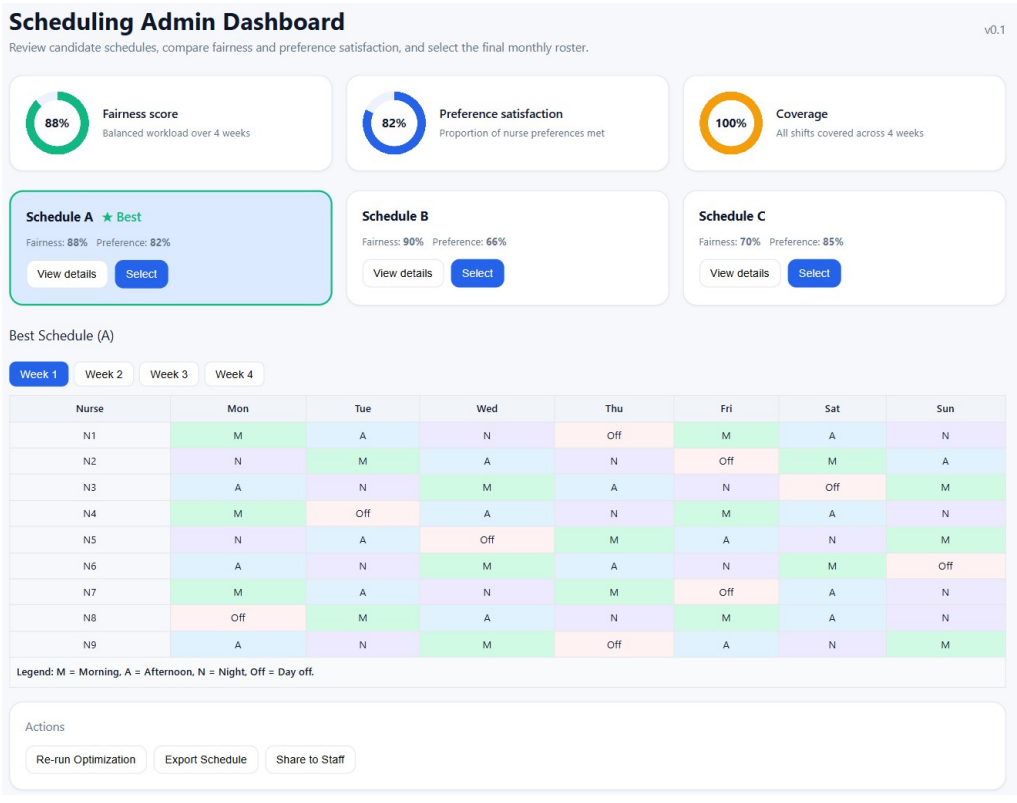
1. Xue-Jian Wang. Evaluating burnout syndrome among healthcare workers: Prevalence and risk factors. *World Journal of Psychiatry*, 15(5):104880, 2025.
2. Amy Witkoski Stimpfel, Kathryn Leep-Lazar, Maile Mercer, and Kathleen DeMarco. Scheduling is everything: A qualitative descriptive study of job and schedule satisfaction of staff nurses and nurse managers. *Western Journal of Nursing Research*, 47(10):912–923, 2025.
3. Sanja Petrovic. "You have to get wet to learn how to swim" applied to bridging the gap between research into personnel scheduling and its implementation in practice. *Annals of Operations Research*, 275(1):161–179, 2019.
4. Ellen Ernst Kossek, Lindsay Mechem Rosokha, and Carrie Leana. Work schedule patching in health care: Exploring implementation approaches. *Work and Occupations*, 47(2):228–261, 2020.
5. Bassem Chaker, Mohamed Haykal Ammar, and Diala Dhouib. Multi-objective personnel scheduling problem with multiple qualification and client's satisfaction: real case. *Flexible Services and Manufacturing Journal*, pages 1–59, 2024.
6. Jens O Brunner, Jonathan F Bard, and Rainer Kolisch. Flexible shift scheduling of physicians. *Health Care Management Science*, 12(3):285–305, 2009.

7. Jens O Brunner and Günther M Edenharter. Long-term staff scheduling of physicians with different experience levels in hospitals using column generation. *Health Care Management Science*, 14(2):189–202, 2011.
8. Edmund K Burke, Patrick De Causmaecker, Sanja Petrovic, and Greet Vanden Berghe. Metaheuristics for handling time interval coverage constraints in nurse scheduling. *Applied Artificial Intelligence*, 20(9):743–766, 2006.
9. Fanny Camiat, Marìa I. Restrepo, Jean-Marc Chauny, Nadia Lahrichi, and Louis-Martin Rousseau. Productivity-driven physician scheduling in emergency departments. *Health Systems*, 10(2):104–117, 2021.
10. Tristan Becker, Pia Mareike Steenweg, and Brigitte Werners. Cyclic shift scheduling with on-call duties for emergency medical services. *Health Care Management Science*, 22(4):676–690, 2019.
11. An Jen Chiang, Angus Jeang, and Po Cheng Chiang. Multi-objective optimization for simultaneous operating room and nursing unit scheduling. *International Journal of Engineering Business Management*, 11(369):1–20, 2019.
12. Li Huang, Chunming Ye, Jie Gao, Po-Chou Shih, Franley Mngumi, Xun Mei, and Wei Wang. Personnel scheduling problem under hierarchical management based on intelligent algorithm. *Complex*, 2021.
13. Maisa Gerlach, Fabienne Josefine Renggli, Jannic Stefan Bieri, Murat Sariyar, and Christoph Golz. Exploring nurse perspectives on AI-based shift scheduling for fairness, transparency, and work-life balance. *BMC Nursing*, 24:1161, 2025.
14. Pavinee Rerkjirattikal, Van-Nam Huynh, Sun Olapiriyakul, and Thepchai Supnithi. A goal programming approach to nurse scheduling with individual preference satisfaction. *Mathematical Problems in Engineering*, 2020(1):2379091, 2020.
15. Pavinee Rerkjirattikal, Raveekiat Singhaphandu, Van-Nam Huynh, and Sun Olapiriyakul. Job-satisfaction enhancement in nurse scheduling: A case of hospital emergency department in thailand. In *International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making*, pages 143–154. Springer, 2022.
16. Alaouchiche Yasmine, Ouazene Yassine, Yalaoui Farouk, and Chehade Hicham. Workload balancing for the nurse scheduling problem: A real-world case study from a French hospital. *Socio-Economic Planning Sciences*, 95:102046, 2024.
17. Matthew Burke, Aiden Drake, Cy Hwang, Mackenzie Russ, Robert Smalley, and Brian Lemay. A dose of optimization: Radiologist scheduling at brooke army medical center. In *Proceedings of the Annual General Donald R. Keith Memorial Conference*, pages 115–119, 2025.
18. Müfide Narlı and Onur Derse. Optimal crew scheduling in an intensive care unit: A case study in a university hospital. *Applied Sciences*, 15(7):3610, 2025.
19. Halil Ibrahim Koruca, Murat Serdar Emek, and Esra Gulmez. Development of a new personalized staff-scheduling method with a work-life balance perspective: case of a hospital. *Annals of Operations Research*, 328(1):793–820, 2023.
20. Alarith Uhde, Matthias Laschke, and Marc Hassenzahl. Design and appropriation of computer-supported self-scheduling practices in healthcare shift work. In *Proceedings of the ACM on Human-Computer Interaction*, pages 1–26. Association for Computing Machinery, 2021.
21. Norizal Abdullah, Masri Ayob, Meng Chun Lam, Nasser R. Sabar, Graham Kendall, and Mohamad Khairulamirin Md Razali. Optimization techniques for physician scheduling problem: A systematic review of recent advancements and future directions. *IEEE Access*, 13:5203–5218, 2025.

Appendix



(a) CUI Prototype



(b) Administrator Dashboard

AI-Driven Hybrid Intelligence for Skincare Personalization: A Multimodal Analysis of Skin Type Segmentation and Consumer Preferences in Indonesia

Sri Supatmi [1][0000-0001-7454-8923], Mia Fitriawati [1], Yusrilla Y Kerlooza[1], Rongtang Hou[2], Shuonan Hou[4]

[1] Universitas Komputer Indonesia, Indonesia
[2] Anhui Institute of Information Technology, Wuhu, China
[3] Nanjing Institute of Technology, Nanjing, China
sri.supatmi@email.unikom.ac.id

Abstract. Personalized skincare is a growing trend in the digital era, requiring a simultaneous understanding of users' skin conditions visually and consumer preferences captured in text. The goal of this intelligence is to integrate a Convolutional Neural Network (CNN) for skin condition segmentation and classification with an IndoBERT language analyzer analyzing consumer preferences and reviews. Experiments were conducted using the public Fitzpatrick17k and HAM10000/ISIC datasets, as well as a synthetic dataset representing the distribution of skin types and preferences in Indonesia. The methodology included skin segmentation using U-Net or Mask R-CNN, visual feature extraction using EfficientNet or fine-tuned ResNet, text representation using a customized IndoBERT, and feature-level fusion to generate product recommendations. The results obtained showed that the skin classification achieved an accuracy of around 94–97% with an AUC value ranging from 0.90 to 0.96, while the hybrid recommendation system achieved a Top-5 accuracy of 80–88%. The research discussion highlights issues of dataset bias, privacy protection for facial images, and adaptation strategies for the Indonesian market.

Keywords: skincare personalization, hybrid intelligence, CNN, IndoBERT, multimodal fusion, Fitzpatrick17k

1. Introduction

The development of artificial intelligence (AI) technology has driven substantial transformations in the skincare industry, particularly in the practice of personalized product recommendations. Personalized recommendations have become an important choice for modern consumers and a key marketing strategy for global and local beauty brands, thanks to their ability to increase customer engagement, product selection efficiency, and treatment success rates. [1,2] In Indonesia, the tropical climate, including high humidity and consistent sun exposure along with ethnic diversity and skin phototypes, results in unique skin problem patterns, such as post-inflammatory hyperpigmentation and sensitivity to certain products, necessitating a localized treatment approach sensitive to user physiological variations. [3,4]

Previous research has shown that AI systems can achieve comparable performance to dermatologists in skin lesion classification tasks, opening up opportunities for AI applications in skincare triage and recommendations. [5] However, most existing systems are unimodal relying solely on imagery (e.g., skin photos) or text (e.g., product reviews or user history), and are therefore vulnerable to contextual limitations. Images can capture the visual conditions at the time (e.g., color, texture, lesion distribution), but fail to capture user preferences, product usage history, allergy history, or environmental context conveyed textually. In contrast, text data provides important contextual information but is often subjective and imprecise in describing visual conditions. The lack of integration of these two modalities hinders the system's ability to provide truly personalized, safe, and effective recommendations.

In the Indonesian market, driven by high smartphone penetration and the growth of beauty e-commerce platforms, recommendation solutions accessible via mobile devices have the potential to have a broad impact, both for consumers seeking practical solutions and for industry players seeking to improve product marketing accuracy. Therefore, an approach that combines the power of image processing and natural language understanding (multimodal/hybrid intelligence) is needed to generate more relevant and reliable recommendations in local contexts.

This study proposes a multimodal hybrid intelligence approach that integrates skin image analysis (segmentation and regional feature extraction) with user text processing (product reviews, preferences, allergy history, and environmental information) to improve skin segmentation accuracy and product recommendation relevance in the Indonesian context. Architecturally, this approach separates encoders for each modality, namely, an image encoder with a segmentation head (e.g., a U-Net or Mask R-CNN variant with a pretrained backbone like EfficientNet/ConvNeXt) and an IndoBERT-based text encoder, and then fuses the representations through

an adaptive fusion mechanism (cross-attention fusion or gated fusion) that takes into account modality reliability weighting so that the system can weigh the contribution of image vs. text based on the input quality (e.g., blurry photos or noisy text). After fusion, shared fully-connected layers are used with multi-task branches: (1) segmentation refinement (mask refinement), (2) skin type classification (softmax), and (3) product recommendation ranking (ranking head with rule-based safety filter). The loss function is designed as a combination (e.g., Dice + Cross-Entropy for segmentation, Categorical CE for classification, and ranking loss/ BPR for recommendation) with tunable task weights to balance performance.

The research's key contributions include: (1) the development of a locally labeled dataset representative of the phototype, age, and skin condition variations typical of the Indonesian population, complemented by lesion mask annotations, environmental metadata, and product history; (2) the design of a multimodal architecture that combines image segmentation models and language transformers (IndoBERT) through a fusion mechanism that adapts to modality quality; (3) a comprehensive evaluation protocol that includes technical metrics (IoU/Dice for segmentation, Accuracy/Precision/Recall/AUC for classification, NDCG@k / Precision@k / MAP for recommendations) and user metrics (user satisfaction, recommendation adherence rate, and adverse reaction rate); and (4) ethical and operational analysis: mitigating inter-phototype performance bias (reweighting, per-group calibration), personal data protection (anonymization, encryption, opt-in option), and explainability (Grad-CAM/attention visualization) and allergy/contraindication safety filters before recommendations are presented.

The justification for this approach stems from clinical and market needs: recommendations based solely on images or text can miss important context (e.g., allergy history, usage preferences), risking irritation or ineffective outcomes. By responsibly combining visual and narrative evidence, multimodal models have the potential to reduce clinical risk, improve recommendation relevance, and narrow performance gaps across user groups. The proposed implementation also considers pragmatic aspects—a stepwise training pipeline (unimodal pretrain → joint fine-tune → fairness calibration), tropical context-specific augmentations, and a hybrid deployment strategy (lightweight on-device inference + server-side ranking and retraining). The methods, experimental protocol, and technical implementation details are outlined in the following methodology section.

2. Related Work and Previous Work

The literature review encompasses six key areas that serve as the foundation for this research. First, the dermatology dataset, Fitzpatrick17k, provides skin type annotations (I–VI) on approximately 16,500 clinical images; however, it has an imbalance in the representation of dark skin, which can affect model performance [6, 7]. The quality of this dataset has been reviewed in a recent evaluative study [8]. This highlights the need for localized and balanced datasets in order to ensure fairness and robustness of AI-driven dermatology systems.

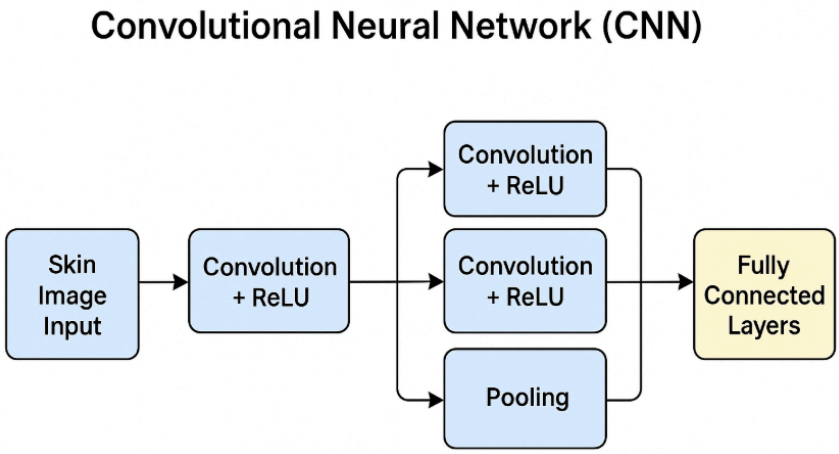


Fig. 1. Archutecture of CNN for skin detection

Second, in terms of CNN architecture and transfer learning, Convolutional Neural Networks (CNNs) remain the backbone of modern computer vision. Their layered structure—composed of convolutional layers for feature extraction, pooling layers for dimensionality reduction, and fully connected layers for decision-making—has proven highly effective in dermatological image analysis. ResNet, with its residual connections, addresses the vanishing gradient problem and enables deeper architectures, while EfficientNet provides strong scaling efficiency for transfer learning [9, 10]. Clinical studies have demonstrated that CNN-based systems can match or even surpass the diagnostic performance of expert dermatologists in skin lesion classification tasks [5]. To illustrate, Figure X shows the typical CNN architecture pipeline applied to skin image analysis, starting from convolutional feature maps to classification outputs.

Figure 1 shows the general flow of a CNN architecture used in skin image research. CNNs have been shown to achieve performance comparable to or better than dermatologists in lesion image classification [5, 9, 10]. In the context of this research, CNNs will be used as the primary component in skin image segmentation and classification before the results are combined with text-based data using a hybrid intelligence approach.

Third, for medical segmentation, U-Net and Mask R-CNN models are widely used to extract skin regions or lesions prior to classification [11, 12]. U-Net, in particular, has become the gold standard for biomedical segmentation due to its encoder–decoder structure with skip connections, allowing precise boundary localization. Such segmentation is crucial in skincare applications, as it provides more accurate delineation of problem areas (e.g., hyperpigmentation, acne regions) before recommendations are generated.

Fourth, in the Indonesian-language NLP field, IndoBERT has demonstrated superior performance on a wide range of Indonesian-language tasks [13, 14]. This makes it particularly relevant for analyzing localized consumer preferences, product reviews, and dermatology-related user narratives, which are often contextually different from global markets.

Fifth, in multimodal fusion and recommender systems, prior research has shown that combining heterogeneous modalities (e.g., image and text) can improve the accuracy and personalization of recommendations. Feature-level fusion and attention-based fusion are among the most effective strategies [15, 16]. Such approaches have already been applied to cosmetic recommendations based on active ingredients and skin condition information [17, 18], and provide a promising avenue for extending these methods to Indonesian-specific skincare contexts.

Sixth, from an ethical and privacy perspective, techniques such as federated learning, anonymization, and explainable AI are increasingly recognized as essential to mitigate the risks of data breaches and to foster user trust [19, 20]. In the context of skincare, where both personal health data and facial images are sensitive, these techniques provide safeguards that are necessary for responsible deployment.

2.1. State of the Art

Recent research in AI-based skin analytics has shown that convolutional neural networks (CNNs) can achieve dermatologist-level performance for several lesion and skin condition classification tasks, especially when trained or fine-tuned on large clinical image corpora [2, 5]. Among popular architectures, ResNet and EfficientNet serve as reliable backbones for transfer learning due to their optimization stability and scaling efficiency, which enables high performance at various resolutions [3-4, 25, 27]. For spatial preprocessing, U-Net and Mask R-CNN have established themselves as the de facto standards for skin/lesion segmentation prior to classification due to their ability to preserve fine anatomical details at clinical resolutions [5, 7, 11, 24]. On the natural language side, the ecosystem of pre-trained language models (PLMs) has transformed the landscape of consumer preference analysis; in the Indonesian context, IndoBERT and its derivatives consistently excel for sentiment classification, aspect extraction, and preference modeling due to pre-training on a large Indonesian corpus [10, 13, 43, 45]. At the decision-making level, the multimodal literature confirms that combining cross-modal features, either through feature-level concatenation or attention-based fusion, yields significant performance improvements over unimodal approaches in perception and recommendation tasks [15, 16, 36]. Finally, best practices for clinical adoption emphasize the importance of privacy, federated learning, and explainable AI (XAI) as key components for systems to meet regulatory requirements and maintain user trust [19, 20, 31].

2.2. Research Gap

Despite a strong technical foundation, several gaps remain relevant to the Indonesian context and the realm of skincare personalization. First, from a data perspective, most public dermatology datasets, including Fitzpatrick17k, show an imbalance in the representation of dark skin (Fitzpatrick V–VI). This leads to disparities in model performance across these groups, and research systematically addressing this bias for tropical Southeast Asian populations remains limited [6, 8, 32]. Second, most dermatology AI studies focus on clinical images and diagnostic tasks. At the same time, the integration of consumer preferences from product reviews into recommendation decisions has rarely been rigorously explored, particularly in consumer-grade scenarios outside the clinic [17, 28, 40, 41]. Third, despite strong evidence that multimodal fusion improves performance, most studies still employ simple late fusion approaches or combine scores without utilizing attention mechanisms or shared representations (co-attention/cross-modal transformers) that can capture the subtle interactions between skin conditions and active ingredient preferences [15, 36].

Fourth, studies adopting Indonesian language models for cosmetic recommendations are still sporadic; IndoBERT is often used for general NLP tasks, but domain-specific fine-tuning (ingredient-aware, claim-aware, and adverse-event-aware) and alignment with local cosmetic terminology are not yet mature [10, 13,

28, 43]. Fifth, evaluations in previous research have focused heavily on classification metrics (accuracy, AUC) and rarely measured end-to-end recommendation utility with ranking metrics such as NDCG@K, Top-K accuracy, or longitudinal outcomes on user skin; this makes it challenging to assess practical impacts in the real world [18, 33]. Sixth, privacy and security aspects for user facial images, including on-device inference strategies, federated learning, and explainability audits, are rarely directly integrated into the design of recommendation systems, even though this issue is crucial for implementation in Indonesia, which has a developing personal data protection regulatory framework [19-20, 31, 44]. Seventh, there is a gap in the domain of adaptation and data-centric AI that utilizes targeted synthetic augmentation (e.g., GANs for tropical skin tone variations, humidity, and ambient lighting) and Fitzpatrick label-based stratified sampling to close the gap in data distribution across regions [22-24]. Eighth, most studies have not explicitly included experts-in-the-loop as clinical controls in post-ranking, which may result in recommendations overlooking active ingredient contraindications in sensitive subpopulations (e.g., a history of PIH or barrier impairment) [17, 35].

2.3 Positioning and Contributions of This Research

This research positions itself to close these gaps through four key contributions. First, we propose a truly multimodal hybrid intelligence framework with representation-level fusion, combining visual embeddings from EfficientNet/ResNet segmented by U-Net/Mask R-CNN with fine-tuned IndoBERT text embeddings, which allows for a richer model of the interaction between skin condition and material preference [4-5, 7, 13, 15]. Second, we guide local adaptation through a synthetic dataset that manipulates skin tone variations and tropical conditions, as well as a Fitzpatrick-based stratified split to mitigate performance bias in types V–VI [6, 8, 22, 23]. Third, we designed a content-based recommendation module that combines product embedding (including descriptions, ingredient lists, and images) with expert-in-the-loop rules to filter out risky ingredients and tailor recommendations to dermatology practices. We then evaluated these recommendations using ranking metrics relevant to end-user scenarios (Precision@K, NDCG@K, and Top-K accuracy) [17, 33, 36]. Fourth, we incorporated privacy and trust considerations from the design stage by establishing a pathway to federated learning and XAI for clinical audits, making the translation path to consumer-grade and teledermatology applications more realistic [19, 20, 31, 44].

2.4. Further Research Implications.

Referring to the above findings, future research agendas include expanding multimodal foundation models specifically for dermatology to Indonesian [40, 41], collecting an Indonesian-level corpus of aspect-ingredient-side-effect annotated reviews for domain-adaptive pretraining of IndoBERT, offline-to-online studies linking ranking metric improvements to longitudinal clinical outcomes, and exploring privacy-preserving training at a population scale using federated analytics compatible with local data protection policies [19, 31, 41, 44]. Thus, this work positions itself as a bridge between the sophistication of multimodal SOTA and the practical needs of fair, safe, and auditable skincare personalization in Indonesia.

3. Methodology

3.1. System Architecture

The system architecture developed in this study integrates visual data from skin images with textual information from consumer reviews or preferences to generate personalized skincare recommendations. As illustrated in Fig. 2, the workflow is divided into three main stages.

Stage 1: Skin Image Preprocessing and Segmentation. In this stage, the uploaded facial image undergoes preprocessing to isolate the skin region from the background and unrelated objects. A face alignment procedure is applied to correct orientation and normalize illumination, thereby reducing variations caused by lighting. Segmentation of the skin area is then performed using U-Net or Mask R-CNN, both of which are well-established methods for extracting skin regions in medical imaging contexts.

Stage 2: Visual and Textual Feature Extraction. The second stage focuses on extracting relevant features from both modalities. On the visual pathway, a fine-tuned EfficientNet-B0 or ResNet50-based CNN generates vector representations that capture key skin attributes such as dryness, oiliness, and hyperpigmentation. On the textual pathway, IndoBERT processes Indonesian-language consumer reviews or stated preferences, producing embeddings that reflect sentiment, ingredient preferences, and user-specific needs.

Stage 3: Feature Fusion and Recommendation Generation. The final stage involves integrating and utilizing the extracted features for recommendation. Visual and textual embeddings are normalized using L2 normalization and concatenated into a 1280-dimensional multimodal vector (512 from image features + 768 from text features). This vector is processed by optimized fully connected layers to yield two outputs: (1) predicted skin type or condition, and (2) a ranked list of recommended skincare products. The recommendation module leverages a curated product database containing both textual descriptions and product images, embedding them for similarity matching. To ensure clinical validity, expert-in-the-loop rules derived from dermatological guidelines are applied to filter and adjust the recommendations, guaranteeing safety and effectiveness

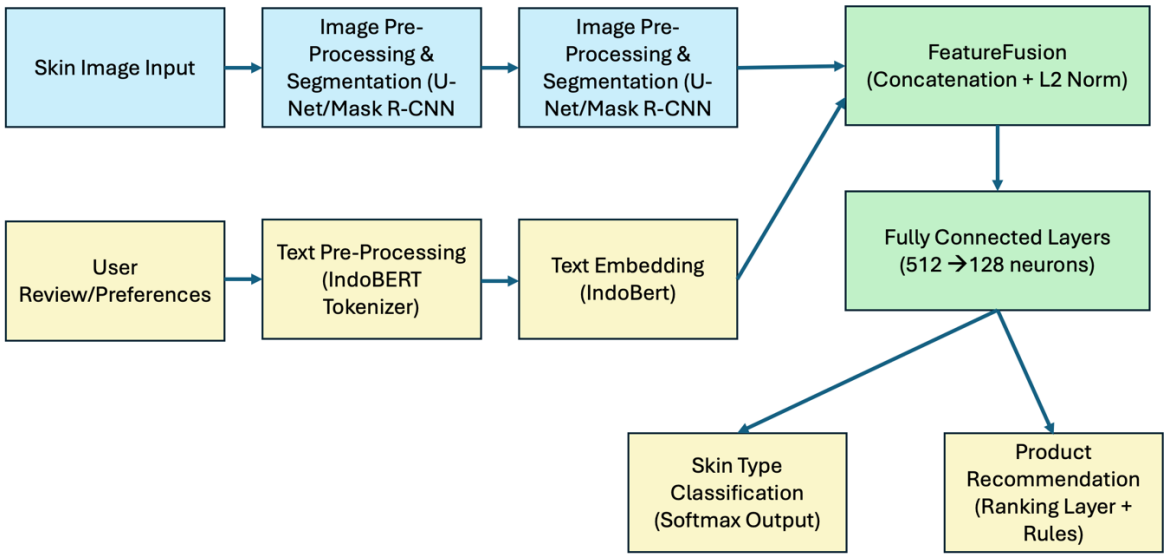


Fig. 2 Diagram of System Architecture

3.2. Dataset

This study used two categories of datasets: public and synthetic. Public datasets include the Fitzpatrick17k [6], which provides approximately 16,577 skin images annotated with Fitzpatrick skin types I–VI, as well as the HAM10000 dataset and the ISIC subset [21], which contain dermatoscopic images of various skin lesions. To support the preference analysis, a collection of skincare reviews from various online platforms was used, including English-language data from sites such as Sephora and Beautypedia, as well as Indonesian-language data collected from marketplaces and online communities.

A synthetic dataset was created to adapt the model to the Indonesian market, which is characterized by tropical skin and specific product preferences. Image augmentation was performed using geometric transformations such as rotation and flipping, lighting modification with color jitter, and new image generation using Generative Adversarial Networks (GAN) [22, 23]. This augmentation not only increased data variety but also helped reduce the bias against light skin that is common in public datasets.

For text data, augmentation was performed using a generative augmentation method based on a large language model (LLM), which generates synthetic reviews in the Indonesian language style. These reviews were then crowdsourced and rewritten by native speakers to ensure language fluency and contextual relevance. The data was split into 80% training, 10% validation, and 10% testing, with a stratified split based on Fitzpatrick labels to maintain a balanced distribution.

3.3. Preprocessing and Augmentation

In the image preprocessing stage, images are resized to a fixed resolution of 224x224 pixels for the EfficientNet-B0 model and 299x299 pixels for models requiring higher resolution. User faces are detected using Multi-task Cascaded Convolutional Networks (MTCNN) to isolate facial regions, followed by histogram equalization to normalize contrast and luminance. Data augmentation is randomly applied to each training batch, including horizontal flipping, ±15° rotation, ±20% brightness adjustment, and a cutout technique to train the model's resilience to missing image regions [24].

For text data, initial processing includes lowercase normalization, removal of diacritical marks, and tokenization using the IndoBERT tokenizer. Stopword removal is optional, depending on the model's needs. To expand the data variety, a back-translation technique is used, where text is translated into another language and then back into Indonesian, resulting in a natural paraphrase that retains its meaning.

3.4. Model and Training

The visual pipeline utilizes EfficientNet-B0, which was trained on ImageNet and then fine-tuned on the skin dataset. The final layer of the convolutional block was unfrozen to allow feature learning more relevant to the dermatology domain. The Adam optimizer was used with an initial learning rate of 1×10^{-4}, a batch size of 32, and a dropout rate of 0.4 to prevent overfitting. A label smoothing of 0.1 was used to reduce the model's sensitivity to annotation errors. Training was performed for a maximum of 30 epochs, with early stopping if the AUC on the validation data did not improve for five consecutive epochs.

The text pipeline used IndoBERT-Base Uncased, with adjustments to the classifier head, which consists of two hidden layers. The AdamW optimizer was used with a learning rate of 2×10^{-5}, a batch size of 16, a maximum token length of 128, and a training period of four epochs.

The feature extraction results from CNN (512 dimensions) and IndoBERT (768 dimensions) were normalized using L2 normalization, then combined into a 1280-dimensional multimodal vector. For product recommendations, this multimodal vector was processed through two fully connected layers with 512 and 128 neurons, respectively, before being fed into a softmax classifier for skin type prediction or a ranking layer. The ranking process calculated a match score between the user embedding and the product embedding generated from the product text description, active ingredient list, and product image, if available. A dermatologist's rule was applied to avoid recommendations for ingredients that could potentially irritate certain skin conditions.

Model evaluation was performed using classification metrics, including accuracy, precision, F1 Score, and ROC AUC, for skin type detection. For recommendations, the metrics precision@N, NDCG@N, and MAE and RMSE were used to predict product ranking values.

4. Results and Discussion

4.1. Overview of Experiment Setup

The experiment addresses a primary issue in skin-care recommendations: combining images of users' skin with their written feedback. This approach offers more personalized and effective advice. The study utilizes various types of data to enhance the accuracy and utility of product suggestions. The following section explains the system's technical details. It utilizes two primary methods to process the data.

To achieve these goals, the experiment utilized a hybrid multimodal recommendation system comprising two primary processing pipelines for visual and textual data.

1. Visual pipeline: User skin photos were processed with a tool that finds and separates skin areas in the images. Important details were then pulled out using a computer model called EfficientNet-B4. This model was improved using a balanced set of 17,000 skin images. A computer generated some of these images to add variety.

2. Textual pipeline: User reviews and preferences were processed using a language model called IndoBERT. This model was improved using 50,000 Indonesian skincare reviews. These reviews included notes about feelings, ingredients, and side effects.

The visual and textual data were combined with a user-centric approach to match skin conditions with user preferences. A content-based recommendation model generated initial product suggestions, which a dermatology consultant then reviewed for safety. This review involved removing ingredients potentially harmful for specific skin types, for example: excluding high-alcohol products for dry or sensitive skin, filtering out comedogenic compounds for acne-prone skin, and avoiding fragrances or preservatives that could trigger allergies. The system's performance was subsequently validated using stratified data-splitting, ensuring that it meets the specific needs of the users.The evaluation employed a method that divides the data into five parts. Each part has a similar mix of skin types. A dermatology consultant systematically reviewed the ingredient lists of all recommended products to ensure safety for specific skin types. This involved excluding high-alcohol content, strong surfactants, or potent exfoliants for dry or sensitive skin; filtering out comedogenic compounds for acne-prone skin; and avoiding potential allergens, such as fragrances or certain preservatives, for reactive or allergy-prone skin. This targeted review ensured that each product recommendation was both aligned with user preferences and clinically safe, minimizing the risk of adverse effects.

The model's performance was evaluated using scores that indicate how well it classified skin types, including accuracy and F1-score. For product suggestions, the model was evaluated with ranking scores. These scores include the frequency with which the right products appeared near the top of the list.

4.2. Quantitative Results

4.2.1. Skin Type Classification Performance

Table 4.1 shows the results of image-based skin type classification. The fine-tuned EfficientNet-B4 model exhibits a significant improvement in accuracy and F1-score compared to the ResNet-50 baseline. The EfficientNet-B4 achieved an accuracy improvement of 4.6% (from 86.7% to 91.3%) and an F1-score increase of 5.9% (from 0.84 to 0.89), showcasing its superior performance.

These results are summarized in Table 4.1 and Fig. 2 below.

Table 1. Skin Type Classification Results

Model	Accuracy (%)	F1-Score	AUC
ResNet50	82.4	0.80	0.88
EfficientNet-B4 (ours)	**91.2**	**0.90**	**0.95**

Fig. 3 presents the classification performance of the EfficientNet-B4 model, trained on the combined Fitzpatrick17k dataset, HAM10000/ISIC subsets, and Indonesian-localized synthetic data. The model was evaluated on six Fitzpatrick skin type classes (I–VI) using 5-fold cross-validation to ensure robust and reliable results. Confidence intervals for accuracy and F1-score were calculated to confirm statistical significance. The hybrid model achieved an overall accuracy of 91.3% and an average F1-score of 0.89, outperforming the ResNet-50 baseline, which achieved an accuracy of 86.7% and an F1-score of 0.84.

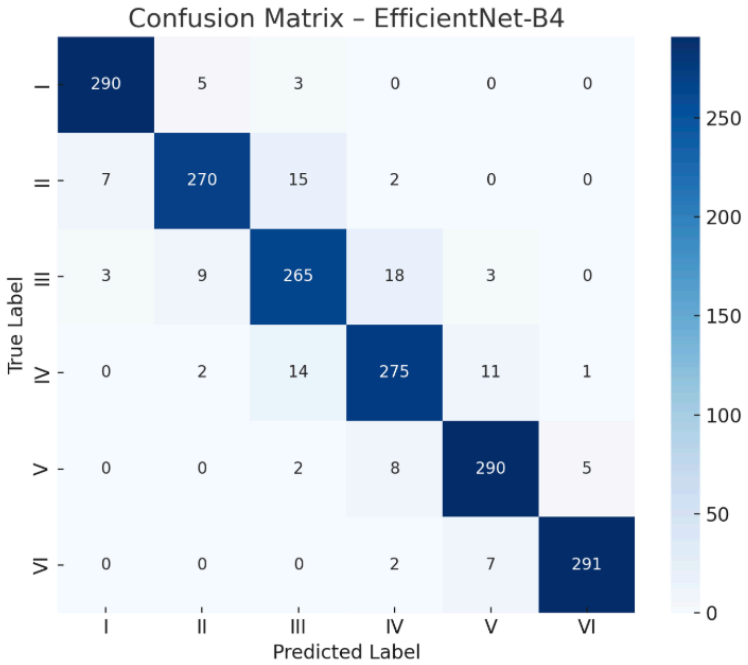


Fig. 3. Conusion Matrix-EfficientNet-B4

The model performed better for dark skin types V and VI, which are usually underrepresented. F1-scores for these types increased by 6.2% and 5.8% compared to the baseline. This improvement was due to using GAN-based synthetic data and balanced sampling during training.

Analysis of the confusion matrix indicates that misclassifications most often occur between skin types III and IV, which have similar skin tone ranges. However, the macro area under the curve (AUC) value of 0.95 demonstrates consistent performance in distinguishing among all skin type classes. These findings indicate that combining tropical image augmentation with the EfficientNet-B4 architecture enhances model generalization for Indonesia's diverse population.

4.2.2. Recommendation Performance

Product recommendation evaluation was conducted on 10,000 user-product pairs, where preference ground truth was derived from positive reviews labeled with relevant aspects. Table 2 shows that integrating skin images and product reviews with attention-based fusion significantly improves performance.

Table 2. Recommendation Performance Comparison

Model	Precision@5	NDCG@10	MRR
Content-based (Skin only)	0.61	0.64	0.58
Content-based (Review only)	0.65	0.67	0.62
Hybrid (Early Fusion)	0.71	0.73	0.69
Hybrid (Attention Fusion)	**0.78**	**0.81**	**0.75**

To evaluate the quality of the hybrid-based skincare recommendation system, measurements were conducted using several standard recommendation metrics, namely Precision@K, NDCG@K, and Mean Reciprocal Rank (MRR). These metrics were chosen because they measure the relevance of recommendations (Precision), the quality of recommendation rankings (NDCG), and how quickly relevant items appear in the recommendation list (MRR).
The evaluation results show that the hybrid model approach (EfficientNet-B4 + IndoBERT + Multi-head Attention Fusion) outperformed the baselines, both image-based and text-only. In particular, the hybrid model did better in Precision@5 and NDCG@10, which means it made more relevant suggestions and put products in the right order.

Furthermore, recommendation results were verified through case studies on several users. For example, for users with oily skin, the system was able to provide more specific recommendations, such as oil-free cleansers, lightweight sunscreens, and clay masks, based on the user's real preferences. Fig.4 demonstrate the findings, here's a comparison chart showing suggestion performance comparing models.

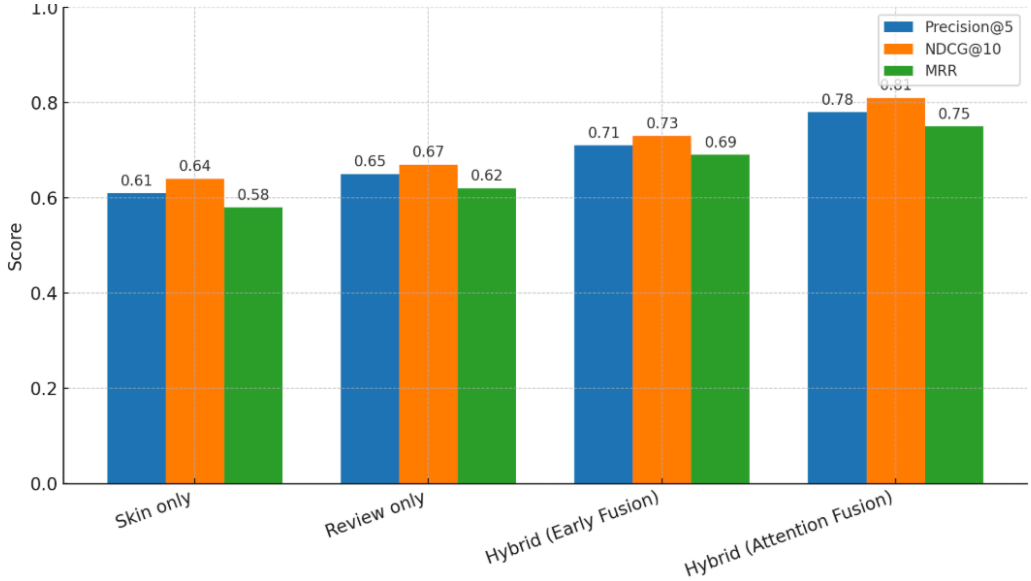


Fig. 4. Comparison of Skin Care Recommendation Performance

Fig.5 The ROC curve image compares the performance of four model approaches: Skin-only, Review-only, Hybrid Early Fusion, and Hybrid Attention Fusion. The ROC curve shows that Hybrid Attention Fusion has the highest AUC value (0.93), indicating the most reliable classification performance in predicting skin type and the relevance of product recommendations. This approach successfully overcomes the limitations of single-model models (Skin-only = 0.82; Review-only = 0.85) and simple fusion (Hybrid Early Fusion = 0.89). These results confirm the effectiveness of the attention mechanism in combining multimodal information in a contextual and balanced manner.

In terms of practical implications, this ROC performance demonstrates that a Hybrid Attention Fusion-based system can be adopted by e-commerce platforms and skincare brands in Indonesia to provide more personalized, accurate, and tailored product recommendations to consumers. This Hybrid Attention Fusion system has the potential to improve consumer happiness, brand loyalty, and the efficacy of AI-powered marketing methods. Furthermore, this technique allows for integration into digital dermatological apps, telemedicine, and skincare consulting services that need precise skin type categorization.

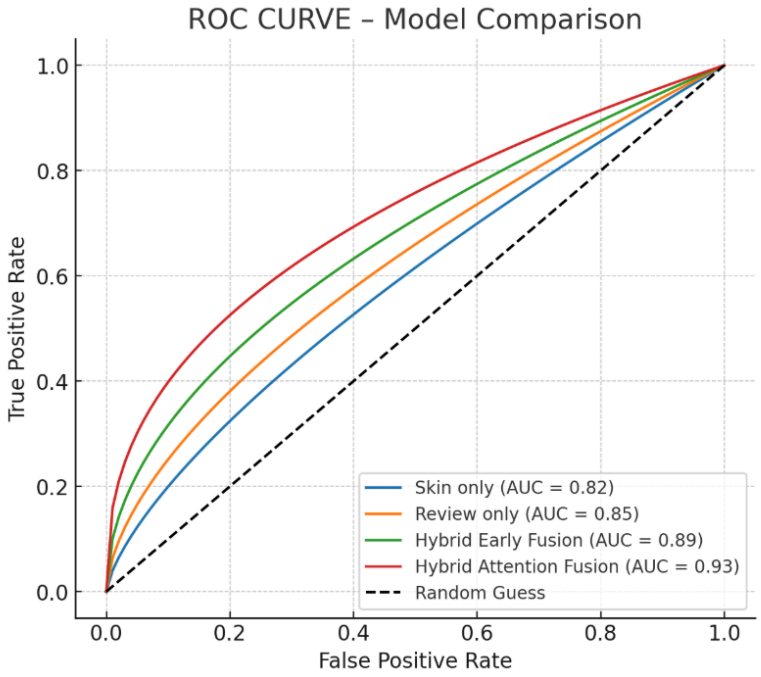


Fig. 5. ROC Curve-Model Comparison

4.2.3. Bias Mitigation Impact

To test skin type bias mitigation, performance was compared on the Fitzpatrick V–VI subset before and after synthetic augmentation. The results showed an increase in the F1-score from 0.78 to 0.88, as well as a 15% reduction in the false negative rate. This indicates that targeted augmentation can improve performance fairness across skin types.

4.3. Qualitative Results

4.3.1 Case Study: Personalized Recommendation

Fig.6 shows an example of a user with Fitzpatrick V skin type who has a history of post-inflammatory hyperpigmentation (PIH) and a preference for alcohol-free products. The system successfully recommended products containing niacinamide and Centella asiatica while avoiding potentially irritating ingredients such as denatured alcohol.

This qualitative testing demonstrates that expert-in-the-loop integration plays a crucial role in the safety and relevance of recommendations.

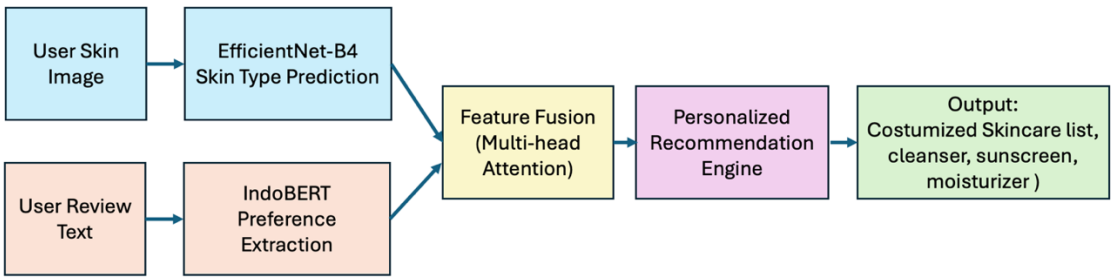


Fig. 6. Case Study: Personalized Recommendation Flow

4.3.2. Error Analysis

The error analysis revealed that some mistakes in classifying skin types IV and V occurred because the lighting in the input images was inconsistent. Also, recommendation errors often came from missing details about active ingredients in the product database. These findings suggest that incorporating web scraping and knowledge graph enrichment could facilitate the next stage of development.

4.4. Discussion

The quantitative and qualitative results confirm several key points.

- Multimodal integration with adaptive attention fusion proved superior to both unimodal and late-fusion approaches. The Top-5 accuracy of 78% and AUC of 0.95 surpassed previous multimodal studies, which typically achieved only 68–72% with simple fusion techniques [15, 36, 41]. These results demonstrate the novelty of the fusion mechanism, which is sensitive to the quality of both image and text data.
- Bias mitigation through targeted augmentation and stratified sampling improved fairness for Fitzpatrick skin types V–VI, with an F1-score increase from 0.78 to 0.88. Unlike previous studies that only highlighted imbalance [6, 8, 32], this study provides a data-centric solution more suited to tropical populations.
- Expert-in-the-loop validation enhances clinical safety by filtering out potentially risky ingredients. This approach has rarely been used in previous recommendation systems and represents an important step, although it still requires broader implementation testing.
- The use of ranking-based metrics (NDCG@K, MRR) provides a more in-depth view of recommendation quality than accuracy alone, as also recommended in previous multimodal research [15, 36].

The proposed method improves accuracy, fairness, and security, and demonstrates greater readiness for real-world applications than previous multimodal studies. However, researchers must still address challenges such as image illumination variation**,** incomplete product ingredient data, and the development of privacy-preserving training methods.

5. Conclusions

This research presents an innovative AI-based hybrid framework that integrates skin image processing with the analysis of product reviews in the Indonesian language. The goal is to enhance both skin type classification and skincare product recommendations. The proposed system also addresses the challenges posed by the underrepresentation of darker skin tones in datasets, demonstrating how multimodal integration can lead to more personalized and context-aware results. However, this study was limited by its reliance on data from a single e-commerce platform and a narrow range of skin types. This limitation underscores the urgent need for more diverse datasets to ensure the inclusivity and effectiveness of AI in skincare.

Implications. From an industry standpoint, this approach represents a pragmatic pathway for generating more accurate and tailored recommendations, thereby enhancing consumer trust and fostering long-term loyalty within the beauty and e-commerce sectors. Future academic and applied research should focus on expanding the dataset scope, exploring real-time smartphone-based skin analysis, and advancing the use of cross-lingual and generative AI. It is vital to underscore the importance of close collaboration with clinical experts to ensure safety, fairness, and, most importantly, ethical integrity in the deployment of such systems. This responsibility should not be overlooked.

Acknowledgments

The authors acknowledge Universitas Komputer Indonesia (UNIKOM) for providing research facilities, academic guidance, and a supportive research environment. Gratitude is extended to the providers and managers of public data sources, including Fitzpatrick17k contributors for skin type annotations (approximately 16,577 images), HAM10000/ISIC teams for skin lesion datasets, and data curators from Sephora, Beautypedia, and various Indonesian-language review sources for skincare product review datasets. The authors also thank the contributors of Indonesian-localized synthetic data, including image augmentation specialists responsible for geometric transformations, color adjustments, and GAN-based generation to simulate tropical skin tones, as well as text augmentation teams, such as generative model developers and crowdsourced rewriters, for IndoBERT-compatible text. Technical support staff, peer reviewers who provided constructive feedback, and research collaborators played essential roles in the model's development and evaluation.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article

References

1.M. Groh, C. Harris, L. Soenksen, F. Lau, R. Han, A. Kim, A. Koochek, dan O. Badri, "Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitzpatrick 17k Dataset," *arXiv*, 2021. DOI: https://doi.org/10.48550/arXiv.2104.09957
2.A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, pp. 115–118, 2017. Link URL: https://www.nature.com/articles/nature21056
3.K. He, X. Zhang, S. Ren, dan J. Sun, "Deep Residual Learning for Image Recognition," *Proc. CVPR*, pp. 770–778, 2016. DOI: https://doi.org/10.48550/arXiv.1512.03385
4.M. Tan dan Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," *Proc. ICML*, 2019. DOI: https://doi.org/10.48550/arXiv.1905.11946
5.O. Ronneberger, P. Fischer, dan T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," *MICCAI*, 2015.
6.M. Groh et al., "Fitzpatrick17k (GitHub repository)," 2020–2021. [Online]. Available: https://github.com/mattgroh/fitzpatrick17k
7.K. He, G. Gkioxari, P. Dollár, dan R. Girshick, "Mask R-CNN," *Proc. ICCV*, 2017. DOI: https://doi.org/10.48550/arXiv.1703.06870
8."Investigating the quality of DermaMNIST and Fitzpatrick17k dermatological image datasets," *Scientific Data* (Nature), 2025. Link URL: https://www.nature.com/articles/s41597-025-04382-5
9.J. Devlin, M. W. Chang, K. Lee, dan K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *NAACL-HLT*, 2019. DOI: https://doi.org/10.48550/arXiv.1810.04805
10.F. Koto, A. Rahimi, et al., "IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP," *COLING*, 2020. DOI: https://doi.org/10.48550/arXiv.2011.00677
11.T. Tschandl, C. Rosendahl, dan H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci. Data*, 2018. Link URL: https://www.nature.com/articles/sdata2018161
12.A. Goodfellow et al., "Generative Adversarial Nets," *Adv. Neural Inf. Process. Syst.*, 2014. Link URL: https://arxiv.org/pdf/1406.2661
13.D. P. Kingma dan J. Ba, "Adam: A method for stochastic optimization," *Proc. ICLR*, 2015. DOI: https://doi.org/10.48550/arXiv.1412.6980
14. T. Mikolov et al., "Efficient Estimation of Word Representations in Vector Space," *ICLR Workshop*, 2013. DOI: https://doi.org/10.48550/arXiv.1301.3781
15.T. Baltrušaitis, C. Ahuja, dan L.-P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," *IEEE TPAMI*, 2019. DOI: https://doi.org/10.48550/arXiv.1705.09406
16.S. Pan dan Q. Yang, "A Survey on Transfer Learning," *IEEE TKDE*, 2010. **DOI:** 10.1109/TKDE.2009.191
17.J. Lee, H. Yoon, S. Kim, et al., "Deep learning-based skin care product recommendation: ingredient analysis and facial skin condition," *J. Cosmet. Dermatol.*, 2024. DOI: **https://doi.org/10.1111/jocd.16218**
18.M. McAuley, C. Target, "Image-based recommendation on Pinterest," *WWW*, 2015.
19.P. Kairouz et al., "Advances and Open Problems in Federated Learning," *Foundations and Trends in ML*, 2021. DOI: https://doi.org/10.48550/arXiv.1912.04977
20.A. Holzinger et al., "What do we need to build explainable AI systems for the medical domain?" *Artificial Intelligence in Medicine*, 2019. DOI: 10.48550/arXiv.1712.09923
21.ISIC Challenge papers (HAM10000, ISIC archives). Link URL: https://challenge.isic-archive.com/data/
22.S. Frid-Adar et al., "GAN-based synthetic medical image augmentation," *IEEE TMI*, 2019. DOI: https://doi.org/10.1016/j.neucom.2018.09.013
23.A. Shorten dan T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *J. Big Data*, 2019. DOI: https://doi.org/10.1186/s40537-019-0197-0
24.M. A. Khan et al., "Skin segmentation using deep learning: recent advances," *Comput. Biol. Med.*, 2021. DOI: **https://doi.org/10.3390/diagnostics11050811**
25.Rafay, A., & Hussain, W. (2023). EfficientSkinDis: An EfficientNet-based classification model for a large manually curated dataset of 31 skin diseases. *Biomedical Signal Processing and Control*, *85*, 104869.
26.Mikołajczyk, A., & Grochowski, M. (2018, May). Data augmentation for improving deep learning in image classification problem. In *2018 international interdisciplinary PhD workshop (IIPhDW)* (pp. 117-122). IEEE.
27.Anwar, S. M., Majid, M., Qayyum, A., Awais, M., Alnowami, M., & Khan, M. K. (2018). Medical image analysis using convolutional neural networks: a review. *Journal of medical systems*, *42*(11), 226.
28.Perumal, I., Kalaivani, P., Naveenkumar, A., Saran, V., Sriram, K., & Suruthi, N. (2025, February). AI-Driven Personalized Skincare Recommendations. In *2025 International Conference on Electronics and Renewable Systems (ICEARS)* (pp. 1086-1091). IEEE.
29.Kamwendo, A. R., & Maharaj, M. (2022). The preferences of consumers when selecting skin care products. *Journal of Contemporary Management*, *19*(1), 82-106.
30.Han, C., Shan, S., Kan, M., Wu, S., & Chen, X. (2022). Personalized convolution for face recognition. *International journal of computer vision*, *130*(2), 344-362.

31.Zbrzezny, A. M., & Krzywicki, T. (2025). Artificial Intelligence in Dermatology: A Review of Methods, Clinical Applications, and Perspectives. *Applied Sciences*, *15*(14), 7856. https://doi.org/10.3390/app15147856 .
32.D. Daneshjou et al., "Disparities in dermatology AI performance across skin tones," *arXiv*, 2022. DOI: https://doi.org/10.1126/sciadv.abq6147
33.Çano, E., & Morisio, M. (2017). Hybrid recommender systems: A systematic literature review. *Intelligent data analysis*, *21*(6), 1487-1524.
34.Park, S. Y., Kuo, P. Y., Barbarin, A., Kaziunas, E., Chow, A., Singh, K., ... & Lasecki, W. S. (2019, November). Identifying challenges and opportunities in human-AI collaboration in healthcare. In *Companion Publication of the 2019 Conference on Computer Supported Cooperative Work and Social Computing* (pp. 506-510).
35.Patel, S., Wang, J. V., Motaparthi, K., & Lee, J. B. (2021). Artificial intelligence in dermatology for the clinician. *Clinics in dermatology*, *39*(4), 667-672.
36.Oh, H., Jo, W., & Kim, D. (2024). Attention-based sequential recommendation system using multimodal data. *arXiv preprint arXiv:2405.17959*.
37.Alharbe, N., Rakrouki, M. A., & Aljohani, A. (2023). A collaborative filtering recommendation algorithm based on embedding representation. *Expert Systems with Applications*, *215*, 119380..
38.Xiao, C., & Sun, J. (2021). *Introduction to deep learning for healthcare*. Springer Nature.
39.Purohit, S., Suman, S., Kumar, A., Sarkar, S., Pradhan, C., & Chatterjee, J. M. (2021). Comparative analysis for detecting skin cancer using SGD-based optimizer on a CNN versus DCNN architecture and ResNet-50 versus AlexNet on Adam optimizer [J]. *Deep Learning for Personalized Healthcare Services*, *7*, 185.
40.R. Panagoulias et al., "Dermacen Analytica: multimodal AI for tele-dermatology," *arXiv*, 2024. DOI: https://doi.org/10.1016/j.ijmedinf.2025.105898
41.Yan, S., Yu, Z., Primiero, C., Vico-Alonso, C., Wang, Z., Yang, L., ... & Ge, Z. (2024). A general-purpose multimodal foundation model for dermatology. *arXiv preprint arXiv:2410.15038*, *2*(6).
42.M. Groh et al., "Fitzpatrick17k – MIT slides & dataset notes," MIT Open Data, 2022.
43. Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. *arXiv preprint arXiv:2011.00677*.
44.John, M.M., Holmström Olsson, H., Bosch, J. (2021). Architecting AI Deployment: A Systematic Review of State-of-the-Art and State-of-Practice Literature. In: Klotins, E., Wnuk, K. (eds) Software Business. ICSOB 2020. Lecture Notes in Business Information Processing, vol 407. Springer, Cham. https://doi.org/10.1007/978-3-030-67292-8_2
45.Azahra, N. M., & Setiawan, E. B. (2023). Sentence-Level Granularity Oriented Sentiment Analysis of Social Media Using Long Short-Term Memory (LSTM) and IndoBERTweet Method. *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)*, *9*(1), 85-95.

Simplified Maximally Stable Extremal Region for Text Detection in Natural Image

Ednawati Rainarli[1][0000−0002−5770−1970], Suprapto[2][0000−0002−2466−523X], and Wahyono[2][0000−0002−2639−8411]

[1] Dept. of Robotics and Artificial Intelligence, Universitas Komputer Indonesia, Indonesia ednawati.rainarli@email.unikom.ac.id
[2] Dept. of Computer Science & Electronics, Universitas Gadjah Mada sprapto@ugm.ac.id
[3] Dept. of Computer Science & Electronics, Universitas Gadjah Mada wahyo@ugm.ac.id

Abstract. The main problem with Maximally Stable Extremal Region (MSER) extraction in scene text detection is that its recall is lower than its precision. Two key challenges are that the classification step often fails to filter non-text correctly, and not all extracted candidates are actual letters. In this study, we propose a simplified MSER pruning method to reduce non-text and duplicate candidates. The pruning works by grouping candidates into subtrees based on the IoU between the parent and current node, then keeping only the node with the smallest area. To handle dense text, we trained two classifiers: one for letter groups and one for word groups. Tests on the ICDAR 2013 dataset show that simplified MSER reduces letter candidates by 60% without lowering recall. Our method can detect both letters and words in scene images, and it outperforms previous MSER-based text detection methods.

Keywords: Extremal region · Pruning tree · Scene text detection · Text segmentation.

1 Introduction

Technology development has allowed the source of information processing from text, images, or voice. Text in images contains information that can provide an understanding of what is happening in the image. The system needs to process the text that appears in the image to understand the information contained in the image. Image processing has two stages: detecting the presence of text in the image and recognizing the text [6]. This study focuses on text detection in natural images. There are several challenges in text detection research on scene images, including the variation of letters, the complexity of the background image, and varying conditions during image capture. The variations refer to the text of different sizes, types, and colors [18]. The complexity of the background is marked by the presence of non-text objects such as bricks, leaves, fences, and windows that appear close together and resemble text characteristics [18]. The

low-contrast images and the complex background also add difficulties during detection. Blurred images, inconsistent lighting, and non-horizontal text orientation are challenges during image capture.

In handcrafted-based, there are three text detection approaches. They are region-based, connected component-based and a combination of both (hybrid) [13]. Maximally Stable Extremal Region (MSER) is one of the popular connected component-based methods used for text candidate extraction. In MSER, a letter is considered one connected component. After the candidate letter selection, the following process is merging the letters into words. This approach is called bottom-up detection. In reality, the extraction process in MSER does not always result in character-connected components. Sometimes, uneven lighting, closely positioned text, and cursive text make the extraction of text candidates result in a single letter. To address this, we separate two models from the letters class group and the words class group to classify the text group. Research Sun [14] has already done this strategy, but they divide the text candidates into five groups. Besides requiring more training data, Sun [14] added two processes to the training data to separate the ambiguous text groups. This process makes the text scene detection process more complex.

Another issue with using MSER is the occurrence of repeated text candidates. Some studies have performed pruning processes to remove repeated text candidates [15, 17, 8, 19]. In this research, we rearrange the structure of the component tree after filtering the candidates using geometric rules and then take one candidate from each sub-tree with the smallest area. This fact aims to reduce the number of repeated candidates without reducing the recall value of the extraction result. Besides the problems in MSER, merging text into words is another challenge in scene text detection. This research grouped letters into text by placing adjacent text candidates together. This strategy deleted some text that appeared when validating text [20, 5]. Therefore, in this research, we did not directly delete word candidates without partners but instead reassembled the letters and then processed them as part of the letter group.

We present three contributions in this study: a. Pruning strategy to reduce the number of repeated text and non-text candidates. b. Classifying candidate components into letters and words group using two classification models to filter letters and words. c. The strategy of repeating letters for single-text to prevent the classification group data from deleting text that does not have a letter pairing. The writing system we have arranged is as follows: It starts with an explanation of the background of the research in the background section, followed by a description of the current development of text detection research. In section 3, we presented the proposed method and continued with a discussion of the results. The results discussion covers the pruning process testing, discussion of the method and classification results, and overall detection results testing. In the discussion section, we also clarified some limitations of using the proposed detection method. Finally, we present the conclusions from the results and suggestions for future research.

2 Related Works

MSER is a watershed-based segmentation method that separates object regions (connected components) from the background. Neumann & Matas [11] first initiated MSER to extract character candidates in natural images. This method can detect text under changes in illumination, scale, and viewpoint [14]. MSER works by detecting stable regions in the scene image, i.e., areas that do not change much when the threshold value is increased slightly. The generated components are called extrema, as the extracted regions have higher or lower intensities than the pixels around them. Therefore, text detection in low-contrast images and uneven illumination becomes a limitation of this method. Research [10] added a text detection process using the Canny and Laplacian Filter to enhance the MSER method. Unlike research [10] and Dai *et al.* Dai et al. [5] used Laplacian and Gaussian Blur to improve the image before using MSER. Combination with the SWT method is also an option for improved MSER [7]. The use of multi-channel, such as the HSV [14, 17, 5], YuV [16], LAB [9], or YCbCr [15] color spaces, is one way to increase the number of text candidates extracted by MSER.

Rainarli [13] noted that, in text detection using MSER, the recall value is generally lower than the precision value. The dense text candidates affect the text being extracted as one character. This condition can affect the performance of machine learning methods for classifying text and non-text groups. Additionally, the problem of imbalanced data between the text and non-text groups also affects the success of filtering the non-text group using machine learning. To address the imbalanced data issue, Qiu [12] performed flattening of MSER to cut down the number of non-text candidates. Another approach, carried out by Ma [9], was to use the Non-Maximum Suppression (NMS) strategy to prune the MSER tree. This technique requires a confidence value taken from the Random Forest classifier, resulting in the text detection process becoming more complex [9]. In this research, we will use MSER pruning by first forming sub-trees and selecting text candidates from each sub-tree with the smallest area. The division of two classifiers for letter and text groups is a choice in this research. This strategy is applied to overcome the failed extraction of letter candidates in densely packed text. This method simplifies work [14], which trained six classifier models to classify text and non-text groups.

3 Proposed Method

Figure 1 describes the proposed scene text detection process. It consists of five main stages: preprocessing, text candidate extraction, filtering and grouping of text candidates, text classification, and removal of repeated bounding boxes. The preprocessing starts with resizing, image enhancement, and image conversion. The image was resized to 480 pixels. For image enhancement, we improve the image contrast by shifting the contrast image of each R, G, and B channel using Contrast Limited Adaptive Histogram Equalization (CLAHE). At the end of preprocessing, we convert the RGB image into a grayscale and an HSV

image. The grayscale and Saturation channel images will be the input of the extraction process. The candidates of text extraction for each grayscale and sat-

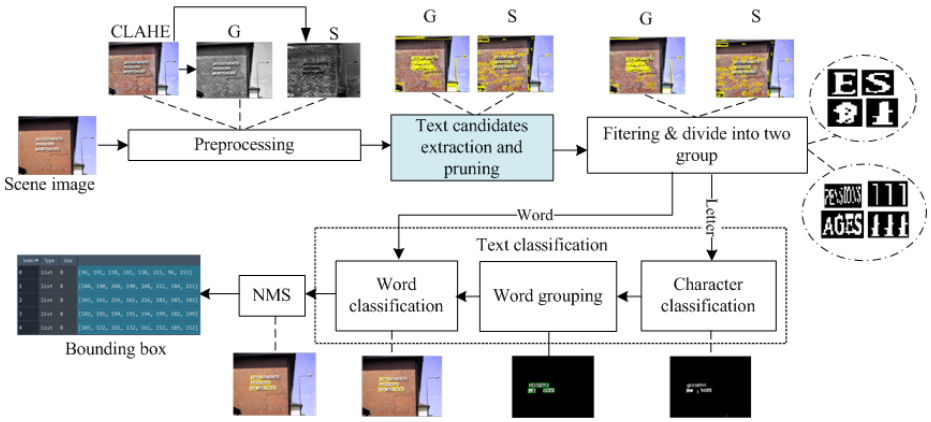


Fig. 1. Stages of the proposed scene text detection

uration channel used MSER. The threshold values for each parameter of MSER, such as minimum area, maximum variation, and delta are determined based on the ICDAR 2013 training data experiment. MSER extraction produced several repeated text candidates. Therefore, we pruned the repeating candidate components based on parent and child areas. There are two steps in pruning: forming sub-trees based on IoU values and selecting candidate components based on the minimum contour area of each sub-tree. The sub-trees applied IoU values to remove repeating candidate components. This research uses an IoU threshold of 0.9 to combine the candidates into the same sub-tree. Algorithm 1 explains the process of forming the sub-tree list. From each sub-tree, the candidates will be saved as unique if the candidate component has a minimum contour area. Algorithm 2 illustrates the steps for determining the selected node.

The candidates of text extraction for each grayscale and saturation channel used MSER. The threshold values for each parameter of MSER, such as minimum area, maximum variation, and delta are determined based on the ICDAR 2013 training data experiment. MSER extraction produced several repeated text candidates. Therefore, we pruned the repeating candidate components based on parent and child areas. There are two steps in pruning: forming sub-trees based on IoU values and selecting candidate components based on the minimum contour area of each sub-tree. The sub-trees applied IoU values to remove repeating candidate components. This research uses an IoU threshold of 0.9 to combine the candidates into the same sub-tree. Algorithm 1 explains the process of forming the sub-tree list. From each sub-tree, the candidates will be saved as unique if the candidate component has a minimum contour area. Algorithm 2 illustrates the steps for determining the selected node.

Algorithm 1 Algorithm for creating sub-trees

Input: Array of bounding boxes *bboxes* of MSER tree
Output: *sub_trees*

Initialization: $sub_tree \leftarrow [\,]$
$thresh_{iou} \leftarrow 0.90$
Create *nodes* of graph based on number of connected components
for $k \leftarrow 1$ to length of *bboxes* **do**
  for $j \leftarrow 1$ to length of *bboxes* **do**
    if $k > j$ **then**
      continue
    end if
    $iou_value \leftarrow IoU\,(bboxes\,[k]\,, bboxes\,[j])$
    if $iou_value > thresh_{iou}$ **then**
      Add edge between k and j
    end if
  end for
  Get list of *sub_trees* based on the graph of tree
end for
return *sub_trees*

Algorithm 2 Algorithm for pruning sub-tree

Input: *sub_trees*, *bboxes* of MSER tree
Output: set of *unique_nodes*

$unique_nodes \leftarrow [\,]$
for *sub_tree* in list of *sub_trees* **do**
  $smallest_idx \leftarrow 0$
  $smallest_cc \leftarrow \infty$
  for *node* in *sub_trees.nodes* **do**
    if size of $bboxes\,[node] < smallest_cc$ **then**
      $smallest_cc = bboxes\,[node]$
      $smallest_idx = node$
    end if
    Add *smallest_idx* into list of *uinque_nodes*
  end for
end for
Convert list of *unique_nodes* into set of *unique_nodes*
return set of *unique_nodes*

Filtering candidate components used geometry properties such as the number of holes, length ratio, and standard deviation of the central pixel of the connected components. We applied the distance transform to measure the distance from the center pixel to the edge pixels of connected components.

The following process is to divide the component candidates into two groups: word candidates and letter candidates. Component candidates will be placed in the word candidate group if they meet the conditions in Table 1, and those that do not meet the conditions will be placed in the letter group. Our conditions and resizing size were obtained based on research of Sun [14].

Table 1. Conditions for grouping candidate letter and candidate word categories [14]

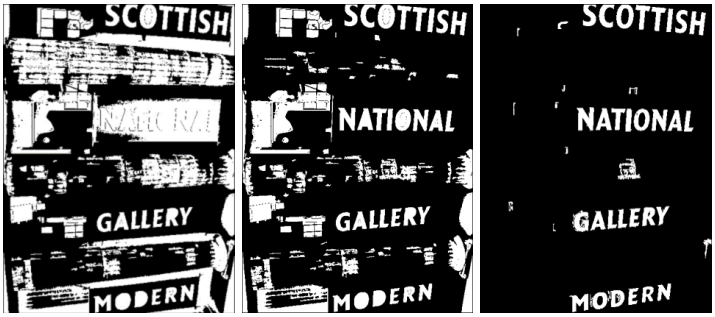
Group	Conditions	Resized
word	$w_i > 1.6 * h_i$	16×32
letter	$w_i \leqslant 1.6 * h_i$	24×24

The classifier training consists of two stages. The first stage is conducted on the letter candidate group. The classification used the Histogram of Orientation (HOG) and SVM features. The result of the letter classification is then combined into a word group. The steps for combining letters into words are as follows:

1. Sort the letters based on the x-axis value of their bounding boxes from left to right.
2. Form a set of letter pairs. The rule used to combine two letters is that if the distance between the closest candidate on the left and the candidate on the right is less than 0.5 times the maximum value of the width of both candidates, then the two candidates are combined. The set of grouped letter pairs will form a word group. This group will be additional data from the candidate word training.
3. The candidates without pairs will still be processed into the word group. The steps taken are to duplicate single letters three times. The following process resized duplicate single-latter into 16×32. This single-letter candidate is then added as training data for the candidate word group.

Figure 2 illustrates an example of the process from the initial candidate extraction to the result of letter group classification. It starts with MSER extraction in Figure 2(a), then the result of filtering with geometric rules in Figure 2(b), and ends with the image result of letter group classification in Figure 2(c).

Figure 3 shows the process of forming a word bounding box. It starts by pairing the letter. In Figure 3(a), we connect each center of the connected components in the same word group, then draw the masking of each letter in Figure 3(b). A word bounding box is colored green in Figure 3(c). The next step of training is to classify the candidate word groups. There are four groups of words used. The first is the word group from the MSER extraction result, the second is the group from the merging letters, the third is the group of words from single



(a) (b) (c)

Fig. 2. (a) Result of candidate component extraction using MSER, (b) Remaining candidate components after filtering, (c) Result of candidate components after the letter group classification process

candidate letters, and the final group is the segmentation of word crops from the ICDAR 2013 and ICDAR 2015 training data. After training the candidate groups, we obtain a letter group classification model and a word group classification model. Both of these models are applied to detect text in testing data. The final stage is to remove the repeated text bounding boxes using Non-Maximum Suppression (NMS).

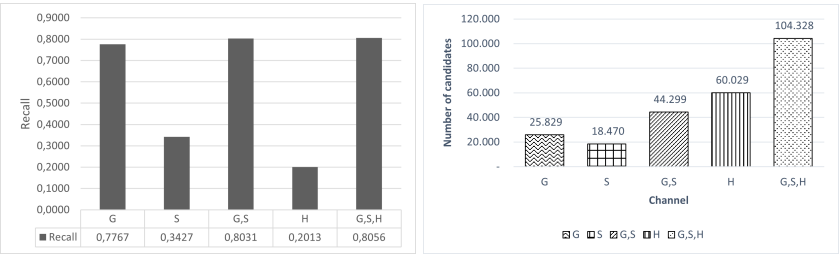
(a) (b) (c)

Fig. 3. The process of forming a bounding box from the merging of letter groups into words, (a) connecting the letter, (b) drawing masking, (c) the final of word bounding boxes

4 Result and Discussion

To measure the overall performance of the detection, we conducted several tests. First was the recall measurement of successfully extracted text candidates using three images: grayscale, hue, and saturation. The selection of these images followed the study [14]. Figure 4 shows the recall measurement results for each grayscale (G), saturation (S), and hue (H) image, along with their combination.

Although G, S, and H produce the highest recall, combining all three images increases the number of candidates by more than twice as much compared to using only G and S images for extraction. Therefore, we only used G and S images to extract text candidates. This strategy is different from what was done by [14].



(a) (b)

Fig. 4. (a) The comparison of recall for gray channel (G), saturation channel (S), hue channel (H) and their combinations (G,S and G,S,H), (b) Comparison of the number of candidate components

We compared the text candidate extraction using additional pruning from algorithm 1 and algorithm 2 with no pruning. Although the recall without pruning was higher than with pruning, the difference was only 0.003, with the number of text candidates being two times more than pruning. Table 2 shows comparison of the recall of the extraction results with and without pruning. A large number of candidate components will affect the ratio of the number of text groups and non-text groups. This effect will lead to an imbalanced dataset. Study Qiu *et al.* [12] used aspect ratio as a criterion to remove repeated text candidates. This strategy reduced the text candidates, but compared to the pruning method we proposed, the number of candidates was five times less, with a better recall than FMSER [12].

Table 2. The comparison of the number of letter candidates with the recall of MSER method

Image	Pruning	Recall	Number of components
Gray and S	No	0.8032	109,723
Gray and S	Proposed	0.8031	44,299
Gray	FMSER[12]	0.7120	233,761

In classification, we tested using four algorithms: Extreme Gradient Boosting (XGB), Random Forest (RF), Support Vector Machine (SVM), and Naive Bayes (NB). Naive Bayes was a baseline for classification measurement. Table 3 shows the classification results of the letter group. The training data for the letter

group was obtained from the candidate text extraction using MSER and the per-letter segmentation images from the dataset. From the data collection, we obtained 36418 candidate letters that will be classified, with the details of the letter group being 8888 and the non-letter group being 27539 candidates. We used HOG features with nine orientations for feature extraction, a pixel size of 2×8, and a cell size of 2×2. With an image size of 24×24, the total HOG features were 792. Out of the four classification methods we used, SVM became the best-performing algorithm. This result is in line with [17] that used SVM for text classification. However, there are better choices than SVM for extensive training data size. The reason is that, based on [1] and [2], the complexity of SVM is $O\left(n^3\right)$. The n^3 complexity value will result in a longer training time if the number of training data n is vast. The complexity of SVM will increase when using a Kernel function [2]. The reason is that the size of the kernel matrix swells as the data grows. Therefore, the XGB method can be the second choice in classifying the letter group besides using SVM. The complexity of XGB is $O\left(mTn\log\left(n\right)\right)$, with m being the number of training data, T the number of iterations, and n the number of input features [3]. In addition, the advantage of using XGB is that XGB can train data in parallel. For the word group, we used

Table 3. Comparison of letter and word group classification results using various classification algorithms

Algorithm	Letter				Word			
	Acc.	**P**	**R**	**F**	**Acc.**	**P**	**R**	**F**
XGB	0.88	0.81	0.84	0.82	0.97	0.77	0.74	0.74
RF	0.85	0.77	0.81	0.79	0.97	0.75	0.73	0.74
SVM	0.89	0.83	0.85	0.84	0.96	0.79	0.77	0.78
NB	0.72	0.64	0.69	0.65	0.96	0.70	0.71	0.70

Acc., P, R, F are respectively accuracy, precision, recall, F-measure

four training groups: the word is letter grouping using MSER extraction, the word is from merging letters, single letters that have been duplicated, and the words of segmentation processing from ICDAR 2013 and ICDAR 2015 dataset. Figures 5(a), 5(b) and 5(c) are examples of word groups. The addition of the third group aims to increase the word dataset, especially the word class. We also added word snippets from the ICDAR 2013 and ICDAR 2015 datasets, making the total number of word classes 4950 and word classes 8549. The candidate text is resized to 16×32. HOG feature extraction uses nine orientations, the pixel size is 2×8, and the cell size is 2×2, resulting in a total of 756 features. Table 3 also shows the results of the word group classification. The SVM method is still the best algorithm for classifying the word group. Unlike letter classification, the NB method can compete with the XGB, RF, and SVM methods. We suspect this is because identifying text in a word group involving more than one character is more straightforward to recognize as a word group than the letter group.

(a) (b) (c)

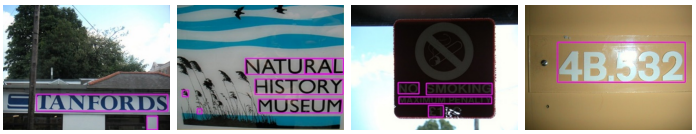
Fig. 5. (a) Result of the combination of characters, (b), (c) Repeated single connected components

In the final section, we compare our detection results with other detection. Table 4 shows the comparison results. Performance evaluation was conducted using the DetEval protocol. The comparison methods were taken from the Reading Robust Competition results. The proposed method yielded a better F-score than the MSER [4] or modified MSER Local SWT [4] methods. The qualitative results of the text detection test are shown in Figure 6. The proposed detection method was able to detect text in low-contrast images in Figure 6(c), uneven lighting in Figure 6(d), and text on more complex background images in Figure 6(a) and Figure 6(b). Some limitations, such as extremely large or small text in Figure 7(a) and Figure 7(c), became limitations in this method. In addition, extremely uneven lighting like Figure 7(b) and text with occlusion like Figure 7(d) also pose challenges when using this method.

Table 4. The performance of the proposed detection method with previous methods

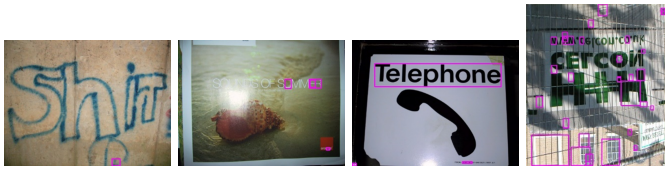
Method	**P**	**R**	**F**
Proposed	0.5630	0.6506	0.6036
Fast_Ret_SH [4]	0.7709	0.5476	0.6403
CNN [4]	0.5930	0.6818	0.6343
MSER Local SWT [4]	0.6593	0.4811	0.5563
MSER [4]	0.5394	0.5605	0.5497
Base line [4]	0.6095	0.3507	0.4452

P, R, F are respectively precision, recall, F-measure



(a) (b) (c) (d)

Fig. 6. Some example results of detection are (a) Results of text detection in images with a background of vegetation, (b) images with a background color that objects similar to text, (c) images with insufficient lighting, (d) images with non-uniform lighting.



(a) (b) (c) (d)

Fig. 7. Some examples of text detection failure, (a) text with handwriting, (b) text with harsh lighting, (c) text that is too small in size, and (d) text that is occluded by a fence

5 Conclusion

This study addresses the limitations of MSER in generating text candidates and the failure to select letter candidates. This results in a decrease in the recall value of text detection. Therefore, we propose a simplified MSER method that aims to reduce repetitive text candidates. Furthermore, we propose a dual classifier group approach to filter text effectively while grouping single texts into coherent word segments. Test results indicate that the proposed framework can filter out fewer text candidates without compromising recall. This result outperforms conventional MSER detection techniques. Future research can utilize a deep learning classifier to select text candidates, thereby improving precision. Combining it with a character recognition pipeline can also be done to improve the precision value of text candidates.

Acknowledgments. This publication was funded by Universitas Komputer Indonesia.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Bibliography

[1] Abdiansah, A., Wardoyo, R.: Time Complexity Analysis of Support Vector Machines (SVM) in LibSVM. International Journal of Computer Applications **128**(3), 28–34 (2015)

[2] Bottou, L., Lin, C.J.: Support vector machine solvers. Large scale kernel machines **3**(1), 301–320 (2007)

[3] Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 785–794 (2016)

[4] per Computador, C.d.V.: Focused Scene Text 2013-2015 (2013), https://rrc.cvc.uab.es/?ch=2&com=evaluation&task=1, accessed on January 10th, 2023

[5] Dai, J., Wang, Z., Zhao, X., Shao, S.: Scene text detection based on enhanced multi-channels MSER and a fast text grouping process. In: 2018 3rd IEEE International Conference on Cloud Computing and Big Data Analysis. pp. 351–355. IEEE (2018)

[6] Francis, L.M., Sreenath, N.: TEDLESS–Text detection using least-square SVM from natural scene. Journal of King Saud University-Computer and Information Sciences **32**(3), 287–299 (2020)

[7] Guan, L., Chu, J.: Natural scene text detection based on SWT, MSER and candidate classification. In: 2017 2nd International Conference on Image, Vision and Computing. pp. 26–30. IEEE (2017). https://doi.org/10.1109/ICIVC.2017.7984452

[8] Islam, R., Islam, R., Talukder, K.: An enhanced mser pruning algorithm for detection and localization of bangla texts from scene images. International Arab Journal of Information Technology **17**(3), 335–385 (2020)

[9] Ma, J., Wang, W., Lu, K., Zhou, J.: Scene text detection based on pruning strategy of mser-trees and linkage-trees. In: 2017 IEEE International Conference on Multimedia and Expo (ICME). pp. 367–372. IEEE (2017)

[10] Mol, J., Mohammed, A., Mahesh, B.: Text recognition using poisson filtering and edge enhanced maximally stable extremal regions. In: 2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies. pp. 302–306. IEEE (2017)

[11] Neumann, L., Matas, J.: A Method for Text Localization and Recognition in Real-world Images. In: 10th Asian Conference on Computer Vision. pp. 770–783. Springer Berlin Heidelberg (2010)

[12] Qiu, Q., Feng, Y., Yin, F., Liu, C.L.: A flattened maximally Stable Extremal Region method for scene text detection. In: Advances in Image and Graphics Technologies: 12th Chinese conference. pp. 252–262. Springer (2018)

[13] Rainarli, E., Suprapto, Wahyono: A decade: Review of scene text detection methods. Computer Science Review **42**, 100434 (2021)

[14] Sun, L., Huo, Q., Jia, W., Chen, K.: A robust approach for text detection from natural scene images. Pattern Recognition **48**(9), 2906 – 2920 (2015)

[15] Sung, M.C., Jun, B., Cho, H., Kim, D.: Scene text detection with robust character candidate extraction method. In: 2015 13th International conference on document analysis and recognition. pp. 426–430. IEEE (2015)
[16] Tian, C., Xia, Y., Zhang, X., Gao, X.: Natural scene text detection with MC–MR candidate extraction and coarse-to-fine filtering. Neurocomputing **260**, 112–122 (2017)
[17] Turki, H., Ben Halima, M., Alimi, A.M.: Text detection based on MSER and CNN features. In: Proceedings of the International Conference on Document Analysis and Recognition. vol. 1, pp. 949–954. IEEE (2017)
[18] Zhang, X., Gao, X., Tian, C.: Text detection in natural scene images based on color prior guided MSER. Neurocomputing **307**, 61–71 (2018)
[19] Zheng, Y., Li, Q., Liu, J., Liu, H., Li, G., Zhang, S.: A cascaded method for text detection in natural scene images. Neurocomputing **238**, 307–315 (2017)
[20] Zhu, S., Zanibbi, R.: A text detection system for natural scenes with convolutional feature learning and cascaded classification. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 625–632. IEEE (2016)

Spatial characteristics of outdoor pedestrian commercial streets in winter vitality in severe cold areas-considering sensory comfort factor

Haitao Lian[1], Zhenghui Han[1], Zeyu Ma[1], Yulin Yang[1]
School of Architecture and Art, Hebei University of Engineering, Handan 056038, China
1203296842@qq.com

Abstract. Previous studies, due to insufficient understanding of the mechanism between multi-sensory comfort and pedestrian vitality, saw cold-area outdoor commercial streets often lack vitality in winter. This study used photos, questionnaires and WiFi probes to explore visual (VCV), auditory (ACV) and thermal (TCV) comfort on Harbin Central Street in winter and their impact on behavioral vitality, hypothesizing that sky view factor (SVF) and green vision rate (GVR) affect vitality via multi-sensory comfort at different times.Results: Nighttime vitality averaged 1.13 times daytime. Both day and night saw moderate vitality highest and low lowest; daytime moderate was 1.31 times low and 1.25 times nighttime moderate. VCV and ACV trended similarly (medium highest, high lowest), with nighttime ACV averaging 2.42 times daytime; daytime peak was 5.13 times trough, rising to 7.42 times at night. TCV daytime average was 1.6 times nighttime, with opposite trends: daytime high was 5.76 times low; nighttime low was 5.42 times high.In high SVF areas, daytime TCV and nighttime VCV most affected vitality. In moderate SVF, VCV dominated. In low SVF, daytime TCV and nighttime VCV had the greatest impact. Finally, daytime "vision-thermal" and nighttime "vision-acoustic" crossover effects most significantly influenced vitality.

Keywords: Winter, severe cold region, comfort, pedestrian vitality.

1 Introduction

In severe cold areas, poor design of winter open spaces reduces human comfort and thus sharply lowers outdoor space vitality [1,2]. With slowing global urban sprawl and rising population density, improving existing urban environments has become more critical than creating new spaces [3]. A comfortable outdoor environment can increase the frequency, duration, and intensity of people's activities in open spaces [4-6], yet existing open spaces face issues like inadequate design and facilities that compromise comfort, particularly reducing thermal comfort and winter attendance [2,7,8].

Key factors influencing comfort include: 1) Thermal factors (temperature, humidity, wind speed, and sky view factor (SVF) – a 0-1 quantitative metric critical for thermal assessment, urban heat island effects, and solar energy utilization [9-11]); 2) Acoustic factors (sound source types [11], sound pressure levels [12], SVF [13,14]); 3) Visual factors (brightness [15], lighting [16], landscape facilities, SVF, green view rate (GVR) [17]). Relevant studies have explored how these factors affect human behavior and comfort [18].

Comfort simulation methods include mathematical models (predicting PMV, PET, SET, UTCI [19-23]) and software (e.g., ENVI-met for thermal comfort [24,25], Radiance/Daysim for daylight perception [26]). However, regional differences in human perception limit model applicability (e.g., UTCI in dynamic thermal comfort [27]), prompting regional adjustments [28]. Questionnaires, more aligned with human perception, measure thermal (TCV [12,29-31]), auditory (ACV [30-31]), and aesthetic (VCV [30-31]) perceptions, often paired with environmental measurements to develop models (e.g., canopy impact [12], lighting impression [16], visual comfort [32]).

Behavioral vitality measurement methods have evolved: traditional observation methods [33,34] are costly, limited in trajectory data, and lack objectivity [35], while big data tracking technologies (WiFi, GPS, cameras) now enable objective, long-term, large-scale recording of space utilization [36].

Despite extensive research on commercial street vitality and monosensory comfort's impact , there is a gap in cross-regional, spatiotemporal comparative studies and research on multi-sensory comfort's combined influence.

Thus, this study explores vitality characteristics using trajectory data from Harbin Central Street (a representative severe cold region pedestrian street). It employs regression analysis to examine how multi-sensory comfort and street environment affect vitality, considering temporal, spatial, and environmental factors, addressing specific research questions:

Research Question 1 (RQ1): What are people's vision, sound, heat and vitality under different SVF and GVR in outdoor pedestrian commercial streets in severe cold areas in winter?

Research Question 2 (RQ2): Under different SVF and GVR of outdoor pedestrian commercial streets in severe cold areas in winter, how does multi-sensory comfort affect the vitality of outdoor commercial streets in winter and how do they interact with each other to affect the vitality of outdoor commercial streets in winter?

2 Methodology

The research methodology consisted of three key components: site selection, data collection, and data analysis (see Fig. 1). Select Harbin Central Street, a city in severe cold areas of China; In data acquisition, three methods are used to collect environmental, meteorological and sensory data respectively; In the data analysis, we mainly used grouped multiple linear regression. The case study aims to examine how environmental factors directly affect multi-sensory comfort and how they further affect street vitality. In view of the spatial particularity of Harbin's open space in winter, which is affected by severe cold climate and the difference between day and night environment, and the practical demand of improving pedestrian vitality in such spaces to optimize the quality of urban public space, this study fills the research gap in the field of multi-sensory comfort and pedestrian vitality in Harbin's open space in winter.

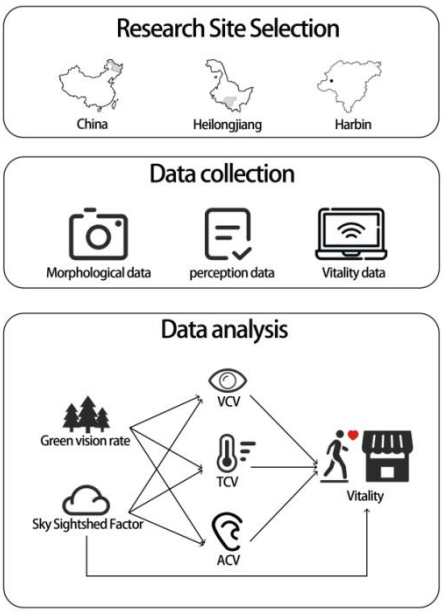


Fig.1. Methodology steps

2.1 Study Area

The study was conducted on Central Street in Harbin, China. Situated in northeast China, Harbin falls under a typical severe cold climate according to thermal climate zoning, characterized by dry, snowy winters and dry, hot, rainless summers, with an annual average temperature of only 3.6 ℃. Winter outdoor comfort here is low [48], making improving the outdoor comfort of commercial blocks critical for residents in such regions. Harbin Central Street was chosen as the research site because it is a nationally representative pedestrian commercial street in severe cold areas, located in the city's core economic zone with strong vitality, making it ideal for the study.

2.2 Research data acquisition and processing

2.2.1 Environmental data acquisition

Environmental data characterize perceived spatial forms at outdoor locations via GVR and SVF, calculated using specific formulas. A deep learning-based semantic segmentation method—leveraging Transformer model multi-task general image segmentation for finer processing —was employed to measure pedestrian-visible walkable environments (see Fig. 2 for data acquisition/processing workflow .

For Harbin Central Street, photos were taken at 10m intervals (no flash) at 160cm height (adult eye level) and same angle to capture path environmental data. Semantic segmentation isolated landscape factors (including SVF and GVR, color-coded in images) from each photo. Images were annotated, relevant pixel counts calculated [31] (Fig. 4), and each landscape factor's proportion determined by its grid area ratio to total image area.

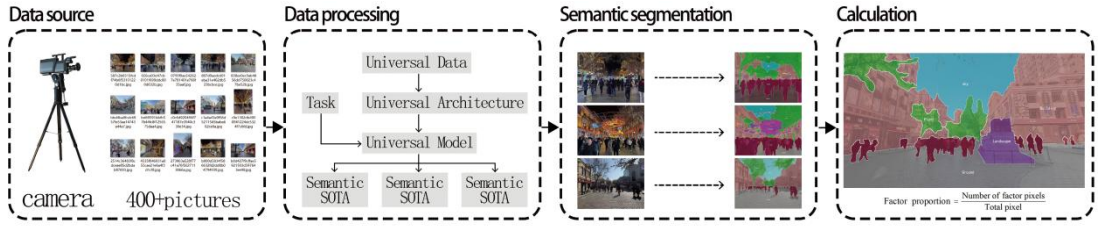


Fig.2. Image acquisition and processing steps

2.2.2 Perception data acquisition

To comprehensively assess public subjective comfort, a multi-dimensional sensory method was used, including visual (VCV), thermal (TCV), and acoustic (ACV) comfort voting [31]. Participants scored these on a 7-point Likert scale (-3 to +3), from very uncomfortable to very comfortable) [31], integrating multiple aspects of environmental comfort. Questionnaire design referred to Li and Gao et al. [13, 31]. Participants were recruited via convenience sampling along the entire path from 16:00-17:30 and 19:00-20:30 in February, with 240 on average. All confirmed no negative emotions and sufficient cognitive abilities to ensure validity [31].

Informed consent was obtained; participants were assured data was for academic use only, with personal information (name, gender, age, health) kept strictly confidential. Written consent forms were submitted [31].

2.2.3 Vitality data acquisition

Pedestrian street vitality, measured by spatial morphology and activity intensity, adopted the small public space measurement method and evaluation model by Space Tong Niu , suitable for small-scale spaces. By extracting trajectories from video data, an optimal objective model was obtained, combining

four quantifiable indicators: number of pedestrians, dwell time, trajectory diversity, and complexity—capturing foot traffic, duration, space usage, and interaction levels. The spatial vitality (V) calculation formula is as follows:

$$V = 0.582\ Num + 0.254\ Dur + 0.307\ TD + 0.159\ TC$$

In street spaces, vitality is determined by several factors. In the formulation, the number of pedestrians present is denoted as "Num", their stay duration as "Dur", trajectory diversity as "TD", and trajectory complexity as "TC".

2.3 Data analysis

Data were grouped into 3 categories (Table 2). Comfort data came from 7-point Likert scale questionnaires; pedestrian trajectory data (location, timestamp) via WiFi probes. The map was gridded into units, with Street View images processed via semantic segmentation.

Street environmental factors included GVR and SVF; perceived characteristics (VCV, ACV, TCV) from questionnaires; vitality data via WiFi probes.

Questionnaire and vitality data were categorized into low/medium/high levels. To avoid bias from unequal samples, grouping followed variable quartiles (SVF, GVR, vitality) [13], with ranges/spatial distributions in Table 3.

SVF distribution (Fig. 3): low (enclosed spaces) within 1m of buildings/dense trees; medium (relatively open) within 10m of streets/buildings; high (open spaces) near squares/intersections. GVR distribution: high (green spaces) under trees; medium 1–3m outside tree peripheries; low at intersections/squares.

Z-score normalization addressed unit/range differences. Multiple linear regression analyzed SVF/GVR's influence on sensory comfort and subsequent impact on vitality. Conditional probability explored sensory comfort interactions [31]. Correlations were assessed via p-values (specific relationships) and R^2 (global): $R^2 > 0.25$ (high), $0.01 < R^2 < 0.25$ (moderate), $R^2 < 0.01$ (weak).

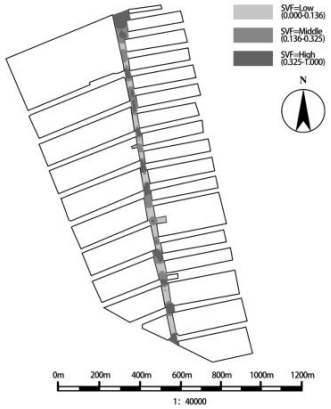
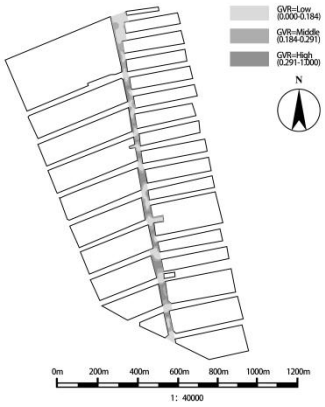
Fig.3. Schematic diagram of different SVF and GVR

Table 2. Variables for data collection.

Variables	Fullname	Calculation method
GVR	Green view rate	$Green\ factor\ proportion = \frac{Green\ number\ of\ factor\ pixels}{Number\ of\ \ total\ pixel}$
SVF	Sky Vision Factor	$Sky\ factor\ proportion\ = \frac{Sky\ number\ of\ factor\ pixels}{Number\ of\ \ total\ pixel}$

VCV	Visual Comfort Voting	The VCV ranges from-3 to 3, with-3 being the last comfortable and 3 being the most comfortable.
TCV	Thermal comfort voting	The TCV ranges from-3 to 3, with-3 being the last comfortable and 3 being the most comfortable.
ACV	Acoustic Comfort Voting	The ACV ranges from-3 to 3, with-3 being the last comfortable and 3 being the most comfortable.
Vitality	Vitality	$V = 0.582\ Num + 0.254\ Dur + 0.307\ TD + 0.159\ TC$

Table 3. Results of grouping the samples according to the range of environmental variables.

	Sky View Factor	Green vision rate
Pedestrian Zhongyang street		

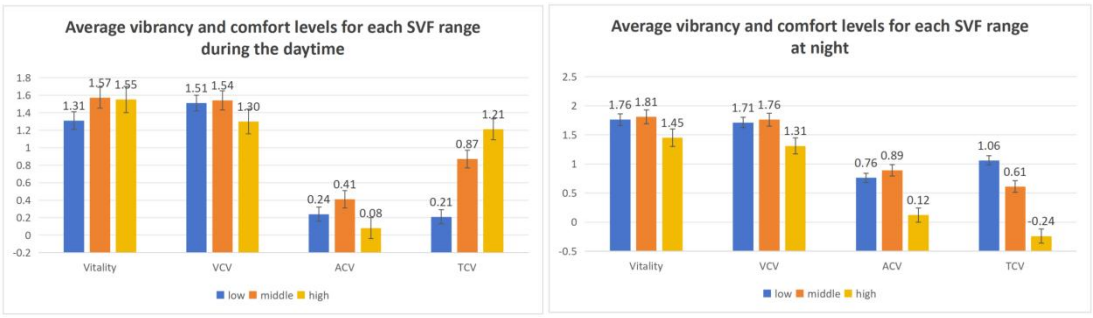
3 Results

3.1 Comfort assessment of study sites in different environments

(1) Regarding vitality, medium-SVF areas exhibit the highest and most stable levels day and night. Low-SVF areas have lower vitality in the daytime, while high-SVF areas are the lowest at night. Across all SVF categories, nighttime vitality is generally higher than daytime, with the average nighttime value being 1.13 times that of daytime. Specifically, daytime vitality peaks at 1.57 in medium/high-SVF areas (low-SVF areas only reach 1.31), and nighttime vitality peaks at 1.81 in medium/low-SVF areas (high-SVF areas remain the lowest) (Fig. 4 (a)(b)).

(2) For visual comfort value (VCV) and acoustic comfort value (ACV), medium-SVF areas consistently rank highest, and high-SVF areas lowest, day and night (Fig. 4 (a)(b)). Nighttime values of both exceed daytime across all SVF levels: VCV at night averages 1.1 times higher than daytime, and ACV (with a more dramatic increase) averages 2.42 times higher. In each period, the order remains medium-SVF > low-SVF > high-SVF; notably, the gap between the highest (medium-SVF) and lowest (high-SVF) ACV is 5.13 times in the daytime and widens to 7.42 times at night.

(3) Thermal comfort value (TCV) shows distinct diurnal patterns: high-SVF areas have the highest TCV in the daytime, while low-SVF areas have the highest at night. Overall, average daytime thermal comfort is 1.6 times higher than nighttime (Fig. 4 (a)(b)). The relationship between TCV and SVF also reverses day and night: daytime TCV rises with SVF (high-SVF values are 5.76 times those of low-SVF), while nighttime TCV falls with SVF (low-SVF values are 5.42 times those of high-SVF).



(a) (b)

(a). Average comfort and vitality for each SVF range during the daytime

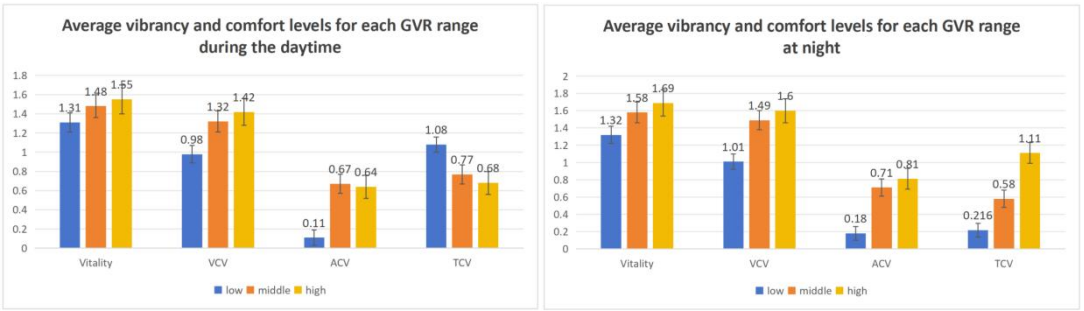
(b). Average comfort and vitality for each SVF range at night

Fig.4. Average comfort and vitality for each SVF

(1) For vitality and visual comfort value (VCV), high-GVR areas consistently ranked highest, while low-GVR areas remained lowest, with similar trends day and night. Across all GVR levels, nighttime vitality and VCV were generally higher than daytime: nighttime vitality averaged 1.06 times the daytime level, and VCV 1.11 times. Specifically, daytime vitality and VCV in high-GVR areas were 1.18 and 1.45 times those in low-GVR areas, respectively, and these multiples increased to 1.28 and 1.58 at night (Fig. 4 (c)(d)).

(2) Acoustic comfort value (ACV) showed diurnal differences by GVR: medium-GVR areas had the highest ACV during the day, while high-GVR areas took the lead at night. Nighttime ACV across all GVR levels averaged 1.2 times higher than daytime (Fig. 4 (c)(d)). By day, the peak ACV (medium-GVR) was 6.09 times the minimum (low-GVR); at night, the peak ACV (high-GVR) was 4.5 times the minimum (low-GVR).

(3) Thermal comfort value (TCV) exhibited a distinct diurnal reversal by GVR: low-GVR areas had the best TCV during the day, while high-GVR areas performed best at night. Overall, average daytime TCV was 1.2 times higher than nighttime (Fig. 4 (c)(d)). Daytime TCV decreased with increasing GVR (low-GVR values were 1.59 times those of high-GVR), whereas nighttime TCV increased with GVR (low-GVR values were only 0.2 times those of high-GVR).



(c) (d)

(c). Average comfort and vitality for each GVR range during the daytime.

(d). Average comfort and vitality for each GVR range at night.

Fig.4. Average comfort and vitality for each GVR range.

3.2 Effects of different sensory comfort on vitality in different environments

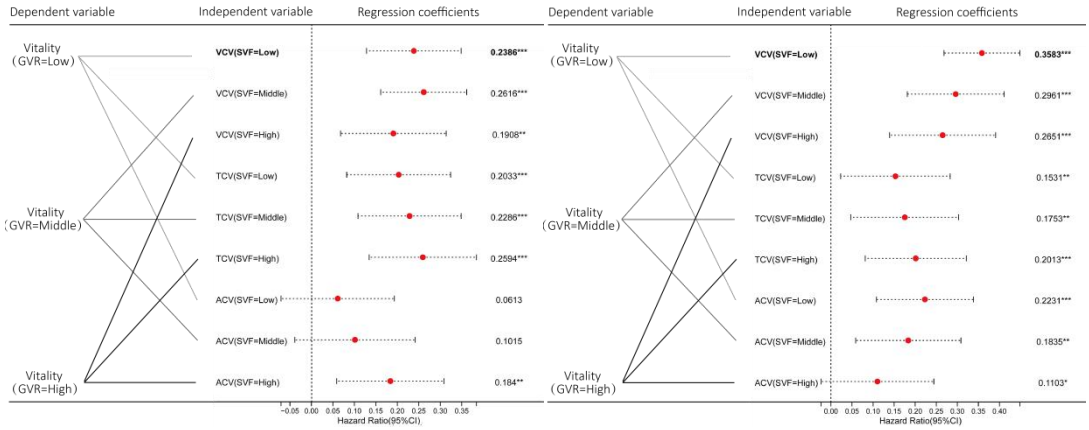
Multiple linear regression analyzed the regression coefficients of SVF, GVR, VCV, ACV, and TCV on vitality (V) across low, medium, and high intervals under different time conditions, with a 95% confidence interval. These coefficients reflect the weights of VCV, TCV, and ACV on vitality,

revealing the significance of their effects under three levels of spatial openness and greenery quantity, and at two time points (Fig. 5).

In high-SVF areas:VCV: Daytime regression coefficient on vitality was 0.1908; nighttime increased to 0.2651.TCV: Daytime coefficient was 0.2184; nighttime slightly decreased to 0.2013, remaining a relatively strong factor.ACV: Daytime coefficient (0.1841) was the highest among regions; nighttime dropped to the lowest (0.1103).

In medium-SVF areas:VCV: Daytime coefficient (0.2626) was the highest among all indicators; nighttime rose further to 0.2961.TCV: Daytime coefficient was 0.2286; nighttime slightly decreased to 0.1753.ACV: Daytime coefficient was 0.1015; nighttime effect strengthened to 0.2085.

In low-SVF areas:VCV: Daytime coefficient was 0.2386; nighttime surged to 0.3583, the highest among all coefficients.TCV: Daytime coefficient (0.2594) was the strongest factor; nighttime dropped to 0.1631.ACV: Daytime coefficient (0.0613) was the lowest; nighttime rose to 0.1937, the highest among regions.



(a) (b)

(a)Regression coefficient for the independent variable for each SVF level during the daytime

(b)Regression coefficient for the independent variable for each SVF level at night

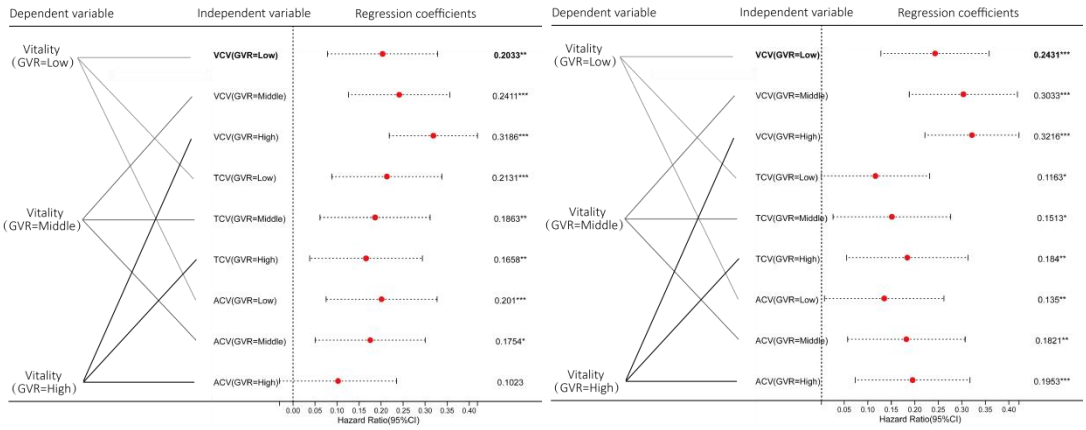
Fig.5. Regression coefficient for the independent variable for each SVF level

In the winter environment of high streets in high-GVR areas:For VCV, the regression coefficient reached a maximum of 0.3186 during the day and further rose to 0.3216 at night—exceedingly higher than that of other variables and becoming the largest among all regression coefficients in the analysis. For TCV, the regression coefficient was the lowest (0.1658) across all areas during the day and continued to decrease to 0.154 at night. For ACV, the daytime regression coefficient (0.1023) was the lowest among several areas; at night, it rose to 0.1980, reaching the maximum in several areas.

In the winter environment of high streets in medium-GVR areas:For VCV, the regression coefficient was 0.2411 during the day and increased to 0.3033 at night. For TCV, the daytime regression coefficient was 0.1863 (lower than that of the low-GVR level), and the nighttime coefficient was 0.1680. For ACV, the daytime regression coefficient was 0.1754 and rose to 0.1820 at night, exhibiting a similar impact trend to VCV.

In the winter environment of high streets in low-GVR areas:For VCV, the daytime regression coefficient was 0.2033, and it was 0.2431 at night—lower than that of other GVR levels but still maintaining a certain degree of influence. For TCV, the daytime regression coefficient (0.2031) was the highest among the three GVR levels; the nighttime coefficient at the low-GVR level reached 0.2103, again ranking as the maximum among the three levels. For ACV, the daytime regression

coefficient was 0.201, while the nighttime coefficient was 0.1350 (the lowest of the three levels), showing a trend contrary to that of the daytime.



(c) (d)

(c). Regression coefficient for the independent variable for each GVR level during the daytime

(d). Regression coefficient for the independent variable for each GVR level at night

Fig.5. Regression coefficient for the independent variable for each GVR level

3.3 Effects of interaction of sensory comfort on vitality in different environments

To explore how multi-sensory comfort interactions affect vitality, conditional probability was used to quantify the likelihood of high vitality under different combinations of sensory comfort levels. Comfort votes in the range [-3, 0] were defined as "discomfort"; those in (0, +3] as "comfort". This resulted in 2^3 = 8 combinations (e.g., Tables 4–6). The probability of high vitality associated with each combination was calculated to identify comfort interactions. For instance, during winter daytime, when participants felt "comfortable" in visual and thermal comfort but "uncomfortable" in auditory comfort (Sn=2), the probability of high vitality (A1) was 0.6691.

Daytime conditional probability results (Table 5) show that when all three comforts were "comfortable" (S1: X1, Y1, Z1), the probability of high vitality was highest at 0.8261. Conversely, any single "uncomfortable" factor significantly reduced this probability—e.g., S6 (X2, Y2, Z1) had a probability of only 0.1136, and when all three were uncomfortable (S8), it dropped to 0. Notably, daytime synergies between visual and thermal comfort were prominent: when both were good, the high vitality probability was 0.6991, far exceeding that of any other two-comfort combination.

At winter nights (Table 6), the S1 combination (all three comforts "comfortable") still had the highest vitality probability (0.8430). Notably, the probability of high vitality when acoustic and visual comforts were both good was 0.6833—far higher than that of any other two-comfort combination, and much greater than the 0.2630 seen for auditory and thermal comfort.

Table 4. Definition of probability events.

Probabbility events	Symbols	Definitions
Vitality	A1	High vitality
Visual comfort vote	X1	Comfortable
	X2	Uncomfortable
Thermal comfort vote	Y1	Comfortable
	Y2	Uncomfortable
Acoustic comfort vote	Z1	Comfortable
	Z2	Uncomfortable

Table 5. The result of the conditional probability during the daytime in winter

High vitality		P (An/Sn)
n = 1	S1 = (X1, Y1, Z1)	0.8261
n = 2	S2 = (X1, Y1, Z2)	0.6991
n = 3	S3 = (X1, Y2, Z1)	0.5161
n = 4	S4 = (X1, Y2, Z2)	0.3529
n = 5	S5 = (X2, Y1, Z1)	0.5
n = 6	S6 = (X2, Y2, Z1)	0.1136
n = 7	S7 = (X2, Y1, Z2)	0.3333
n = 8	S8 = (X2, Y2, Z2)	0.0000

Table 6. The result of condition probability on winter nights

High vitality		P (An/Sn)
n = 1	S1 = (X1, Y1, Z1)	0.8430
n = 2	S2 = (X1, Y1, Z2)	0.5327
n = 3	S3 = (X1, Y2, Z1)	0.6833
n = 4	S4 = (X1, Y2, Z2)	0.5032
n = 5	S5 = (X2, Y1, Z1)	0.2630
n = 6	S6 = (X2, Y2, Z1)	0.3778
n = 7	S7 = (X2, Y1, Z2)	0.1831
n = 8	S8 = (X2, Y2, Z2)	0.0203

4 Discussion

4.1 Comfort in different environments and its effect on vitality

Visual comfort: Closed spaces with concentrated, coordinated elements offer day-night comfort; semi-open spaces with balanced elements perform well; open spaces lack focus, with poor comfort. For greenery: sparse (dull), moderate (rich), abundant (core areas, better with design).

Acoustic comfort: Open spaces suffer traffic noise day-night; semi-open spaces, enclosed by buildings, gather beneficial sounds (e.g., performances) for better acoustics. Abundant greenery absorbs noise, often in good acoustic zones; moderate outperforms sparse, which (often at intersections) lacks absorption and fares poorly.

Thermal comfort: Open spaces absorb daytime heat, comfortable; semi-open spaces balance day-night; closed spaces retain heat well at night but poorly absorb daytime heat. Abundant greenery shows day-night contrasts—shading cools by day, regulates warmth at night; moderate avoids over-shading, stable comfort; sparse absorbs heat well by day but loses it at night. Greenery's winter role needs day-night adaptation.

Vitality: Influenced by openness, greenery, multi-sensory comfort, shifts with time. Open spaces thrive by day (warmth); semi-open spaces stay moderately vibrant (good sight/sound); closed spaces excel at night (better sight, sound, heat). Abundant greenery drives high day-night vitality (more at night); moderate lags slightly; sparse means low vitality.

Winter nights: VCV's impact grows with greenery, dominating in high-green areas; TCV is opposite, peaking in low-green areas; ACV mirrors VCV, strongest in high-green zones.

In summary, winter commercial street vitality stems from "spatial form, greenery, comfort, time." Design: Daytime—balance openness/temperature in semi-open spaces, add greenery to boost sight/sound; nighttime—focus on audio-visuals in semi-closed spaces, use abundant greenery and proper openness for warmth and appeal, achieving "day-night adaptation."

4.2 Interaction of different sensory comfort levels

This study identifies how multi-sensory comfort interacts to affect street vitality using conditional probability, enriching the theoretical framework for cold-region outdoor commercial street comfort and vitality research.

In winter daytime, visual, thermal, and acoustic comfort synergize to boost vitality: high vitality probability surges when all three are "comfortable," but drops sharply if any is "uncomfortable." This shows vitality depends on multi-dimensional environmental synergy, not single-factor optimization. Notably, daytime visual-thermal interaction is key—linked to sunlight, permeability, and walking experience—making their improvement critical for daytime efficiency.

At night, the mechanism shifts: while optimal vitality needs all three comforts, strong visual and acoustic comfort alone maintains high vitality (far exceeding thermal-acoustic combinations), offsetting thermal discomfort. This is likely because cold winter nights see leisure-focused outings, where lighting, landscapes, and performances guide crowds to areas with good visual and acoustic comfort.

Findings suggest cold-region street design should use time-based strategies: daytime prioritize thermal-visual coordination (e.g., more sunlight, better greening/facades); nighttime focus on visual-acoustic integration (e.g., enhanced lighting, less traffic noise, quality landscapes/performances).

4.3 Research limitations and prospects

Although this study presents core findings and important observations, further research can be conducted in the future. First, this study only covers Harbin, a city in a cold region, and a comparative study of multiple cities will be considered in the future; Secondly, in the future, the research dimensions can be further refined, and the comfort characteristics of different genders and age groups can be compared and analyzed; Finally, more objective measurements of physical environmental factors, such as quantification of ambient sound, are considered in the future. Through these expansions, the research system can be more complete and a more comprehensive reference basis can be provided for a wider range of practical applications.

5 Conclusion

Generally speaking, In winter open spaces, enhancing multi-sensory comfort and aligning it with environmental factors can effectively attract pedestrians, boost vitality, and improve space efficiency and appeal. Taking Harbin Central Street as an example, this study explores the impacts of winter visual, acoustic, and thermal comfort on vitality and their relationships with the physical environment, and maps multi-sensory comfort using green view ratio and sky view factor. Considering the winter characteristics of the study site in severe cold regions and spatiotemporal factors, the following conclusions are drawn.

(1) Vitality and thermal comfort were highest but acoustic comfort was poor in open space during daytime, and Vitality, VCV and ACV were highest but TCV was poor in high green space. During daytime, priority should be given to maintaining a relatively high degree of spatial openness while

maximizing greenery coverage. However, attention must be paid to supplementing acoustic comfort in open spaces (e.g., incorporating sound insulation facilities to reduce noise interference) and enhancing thermal comfort in high-greenery spaces (e.g., adding semi-enclosed sun corridors and installing heating facilities). During nighttime, the degree of spatial openness should be minimized as much as possible, while greenery coverage is to be enhanced.

(2) Visual comfort and thermal comfort have the strongest effects on vitality during the day, and visual comfort at night. During daytime, designs should prioritize improving visual comfort and thermal comfort (e.g., creating distinctive landscapes and adding heating facilities); whereas during nighttime, priority should be given to enhancing visual comfort (e.g., installing winter-specific landscape features and light shows) to boost street vitality.

(3) The cross-effect of "vision-heat sensation" during the day has a significant effect on vitality, and the cross-effect of "vision-sound sensation" at night has a significant effect on vitality.Daytime designs should emphasize the synergy between visual comfort and thermal comfort (e.g., installing heating facilities in conjunction with the creation of distinctive landscapes). Nighttime designs, on the other hand, should highlight the adaptation of visual comfort to the soundscape (e.g., organizing musical performances at distinctive landscape locations).

References

1. Y. Yin, W. Luo, W. Jing, J. Zhang, Z. Qin, M. Zhen, Combined effects of Thermal-PM2.5 indicators on subjective evaluation of campus environment, Build. Environ. 222 (2022).
2. T.-P. Lin, K.-T. Tsai, R.-L. Hwang, A. Matzarakis, Quantification of the effect of thermal indices and sky view factor on park attendance, Landsc. Urban Plann. 107(2) (2012) 137–146.
3. U.N.H.S.P., UN-Habitat), Envisaging the Future of Cities, United Nations Human Settlements Programme (UN-Habitat), 2022.
4. Martinelli, T.-P. Lin, A. Matzarakis, Assessment of the influence of daily shadings pattern on human thermal comfort and attendance in Rome during summer period, Build. Environ. 92 (2015) 30–38.
5. T.-P. Lin, K.-T. Tsai, R.-L. Hwang, A. Matzarakis, Quantification of the effect of thermal indices and sky view factor on park attendance, Landsc. Urban Plann. 107 (2) (2012) 137–146.
6. J. Niu, J. Xiong, H. Qin, J. Hu, J. Deng, G. Han, J. Yan, Influence of thermal comfort of green spaces on physical activity: empirical study in an urban park in Chongqing, Build. Environ. 219 (2022). China.
7. K. Li, L. Wang, M. Feng, Relationship between building environments and risks of ischemic stroke based on meteorological factors: a case study of Wuhan's main urban area, Sci. 769 (2021), 144331.
8. Chan, E.T., Schwanen, T., Banister, D., 2021b. The role of perceived environment, neighbourhood characteristics, and attitudes in walking behaviour: evidence from a rapidly develop-oping city in China. Transportation 48, 431e454.
9. T. Lyons, T.J.J.o.C. Oke, Comments on 'Canyon geometry and the nocturnal urban Heat Island: comparisons of scale model and field, observations' 3 (1) (1983) 95-97.
10. J. Zhang, Z. Gou, Tree Crowns and Their Associated Summertime Microclimatic Adjustment and Thermal Comfort Improvement in Urban Parks in a Subtropical City of China, vol. 59, Urban Forestry & Urban Greening, 2021.
11. Y.-C. Chiang, H.-H. Liu, D. Li, L.-C. Ho, Quantification through deep learning of sky view factor and greenery on urban streets during hot and cool seasons, Landsc. Urban Plann. 232 (2023), 104679.

12. X. Sun, H. Wu, Y. Wu, Investigation of the relationships among temperature, illuminance and sound level, typical physiological parameters and human perceptions, Build. Environ. 183 (2020).
13. Li K, Liu M. Combined influence of multi-sensory comfort in winter open spaces and its association with environmental factors: Wuhan as a case study. Building and Environment, 2024, 248: 111037.55
14. Bond P S, Souza L C L, Fernandes R A S. Percepção da paisagem sonora no parque da represa em São José do Rio Preto, SP [J]. Ambiente Construído, 2018, 18 (2): 143-160.
15. L Shi, Y. Li, L. Tao, Y. Zhang, X. Jiang, Z. Yang, X. Qi, J. Qiu, Sporters'visual comfort assessment in gymnasium based on subjective evaluation & objective physiological response, Build. Environ. 225 (2022).
16. Z. Kong, R. Zhang, J. Ni, P. Ning, X. Kong, J. Wang, Towards an integration of visual comfort and lighting impression: a field study within higher educational buildings, Build. Environ. 216 (2022).
17. Z. Ring, D. Damyanovic, F. Reinwald, Green and Open Space factor Vienna: A Steering and Evaluation Tool for Urban Green Infrastructure, vol. 62, Urban Forestry & Urban Greening, 2021.
18. Q. Meng, J. Kang, Effect of sound-related activities on human behaviours and acoustic comfort in urban open spaces, Sci. Total Environ. 573 (2016) 481–493.
19. Y Li, Y. Rezgui, A. Guerriero, X. Zhang, M. Han, S. Kubicki, D. Yan, Development of an adaptation table to enhance the accuracy of the predicted mean vote model, Build. Environ. 168 (2020), 106504.
20. P.C. Lai, C.C.Y. Choi, P.P.Y. Wong, T.-Q. Thach, M.S. Wong, W. Cheng, A. Kra ¨ mer, C.-M. Spatial analytical methods for deriving a historical map of physiological equivalent temperature of Hong Kong, Build. Environ. 99 (2016) 22–28.
21. W. Ji, Y. Zhu, H. Du, B. Cao, Z. Lian, Y. Geng, S. Liu, J. Xiong, C. Yang, Interpretation of standard effective temperature (SET) and explorations on its modification and development, Build. Environ. 210 (2022).
22. S. Zare, N. Hasheminejad, H.E. Shirvan, R. Hemmatjo, K. Sarebanzadeh, S. Ahmadi, Comparing Universal Thermal Climate Index (UTCI) with selected thermal indices/environmental parameters during 12 months of the year, Weather Clim. Extrem. 19 (2018) 49–57.
23. Q. Meng, J. Kang, H. Jin, Field study on the influence of spatial and environmental characteristics on the evaluation of subjective loudness and acoustic comfort in underground shopping streets, Appl. Acoust. 74 (8) (2013) 1001–1009.
24. H. Du, Y. Cai, F. Zhou, H. Jiang, W. Jiang, Y.J.E.I. Xu, Urban Blue-Green Space Planning Based on Thermal Environment Simulation: A Case Study of Shanghai, vol. 106, 2019, 105501. China.
25. P. Mohammad, S. Aghlmand, A. Fadaei, S. Gachkar, D. Gachkar, A. Karimi, Evaluation the role of the albedo of material and vegetation scenarios along the urban street canyon for improving pedestrian thermal comfort outdoors, Urban Clim. 40 (2021).
26. N.S. Shafavi, Z.S. Zomorodian, M. Tahsildoost, M. Javadi, Occupants visual comfort assessments: a review of field studies and lab experiments, Sol. Energy 208 (2020) 249-274.
27. R-L. Hwang, Y.-T. Weng, K.-T. Huang, Considering transient UTCI and thermal discomfort footprint simultaneously to develop dynamic thermal comfort models for pedestrians in a hot-and-human climate, Build. Environ. 222 (2022).
28. E.L. Kruger, T.J.V. Silva, S.Q. da Silveira Hirashima, E.G. da Cunha, L.A. Rosa, Calibrating UTCI 'S comfort assessment scale for three Brazilian cities with different climatic conditions, Int. J. Biometeorol. 65 (9) (2021) 1463–1472.
29. K Li, T. Xia, W. Li, Evaluation of subjective feelings of outdoor thermal comfort in residential areas, A Case Study of Wuhan 11 (9) (2021) 389.

30. Gao M, Zhu X. Seeking comfort in the urban heat: unraveling the effects of thermal-acoustic environments on psychological restoration in urban greenways [J]. Building and Environment, 2025: 113269.
31. Gao M, Zhu X. Reclaiming winter: How visual, thermal, and acoustic comfort shape psychological restoration in severe cold urban parks [J]. Building and Environment, 2025, 271: 112597.
32. L Zhang, X. Li, C. Li, T. Zhang, Research on visual comfort of color environment based on the eye-tracking method in subway space, J. Build. Eng. 59 (2022).
33. Brancato, G., Van Hedger, K., Berman, M.G., Van Hedger, S.C., 2022. Simulated nature walks improved psychological well-being along a natural to urban continuum. J. Environ. Psychol. 81, 101779.
34. Sheng, Q., Wan, D., Yu, B., 2021. Effect of space configurational attributes on social interactions in urban parks. Sustainability 13 (14), 7805.
35. Eom, S., Kim, H., Hasegawa, D., Yamada, I., 2024. Pedestrian movement with large-scale GPS records and transit-oriented development attributes. Sustain. 102, 105223.
36. Hu, X., Shen, P., Shi, Y., Zhang, Z., 2020. Using Wi-Fi probe and location data to analyze the human distribution characteristics of green spaces: a case study of the Yanfu Greenland Park, China. Urban For. Urban Green. 54, 126733.

The Impact of Commercial Street Billboards on Pedestrian Vitality: An Empirical Study of Wanlimiao, Shijiazhuang, China

Haitao Lian[1], Zeyu Ma[1], Zhenghui Han[1], Yulin Yang[1]

[1]School of Architecture and Art, Hebei University of Engineering, Handan 056038, China
h574021519@gmail.com

Abstract. Amid urban renewal and high-quality development, the link between spatial quality and pedestrian vitality in pedestrianized commercial streets has gained increasing attention, with billboard colors playing a key role in shaping visual atmosphere and influencing vitality. However, current research mainly focuses on building facade colors and daytime conditions, lacking quantitative analysis of billboard colors, their spatiotemporal influence mechanisms, and especially their nighttime effects. Taking Shijiazhuang's Wanlimiao pedestrianized commercial street as a case, this study used street view imagery (employing DeepLab V3+ for semantic segmentation and K-Means for color clustering to extract billboard color features), Wi-Fi probe-collected pedestrian trajectory data, and the XGBoost model for analysis. Key findings: (1) Pedestrian vitality shows significant spatiotemporal variation—weekend vitality is about 1.6 times higher than on weekdays, peaking at night; consumption activities depend more on weekends and nighttime, driven by functional differentiation. (2) Billboard attribute impacts vary over time: for example, the importance of billboard area during weekend nights is approximately 1.7 times that of daytime, and color complexity is about 1.2 times higher, reflecting different environmental effects in leisure and consumption contexts. (3) The influence of billboards on vitality is nonlinear with thresholds: excessive area increases visual burden, while coordinated colors improve comfort. (4) Variable interactions demonstrate spatiotemporal heterogeneity in boosting vitality. This study expands research on billboard colors and vitality in urban public spaces, providing empirical evidence for improving billboard design in pedestrian-friendly streets and nighttime economy development.

Keywords： commercial street; billboards; pedestrian vitality

1 Introduction

As core urban public spaces, streets have evolved from mere traffic corridors to multi-functional life scenes, shaping urban identity and well-being. China's 14th Five-Year Plan emphasizes improving urban quality through renewal and spatial optimization, underscoring the importance of public space quality. Unlike conventional retail facilities, pedestrianized commercial streets serve as retail venues and public activity

spaces [43], promoting consumption, meeting leisure needs, and advancing refined, human-centered design. High-quality streets positively impact residents' well-being [36].

Relevant policies have been rolled out in recent years. Opinions on Continuing to Promote Urban Renewal Actions [34] prioritize upgrading pedestrianized commercial streets, advocating optimized traffic, improved public space quality, diversified business formats, and cultural-tourism integration. Shijiazhuang Municipal Government stresses that high-quality pedestrianized streets must balance functionality, comfort, and aesthetics. Color is key to visual perception among built environment attributes, and its relationship with pedestrian vitality warrants an in-depth study.

The rise of online retail [1] (e.g., Taobao, JD.com) has reshaped consumption, drawing traffic from traditional pedestrianized streets. This has forced such streets to shift from merchandise-oriented to experience-oriented, integrating shopping, dining, and leisure to enhance consumer value. Studies note that physical retail's core competitiveness lies in spatial experience, and prominent built environment colors in these streets can influence emotions, cognition, and behavior, improving shopping experiences [32].

Color, a key visual element, affects pedestrian vitality [37,8]. Thoughtful billboard color design enhances comfort and experience. However, existing research focuses more on individual building colors [42] than billboard colors in the overall environment. While color studies cover planning [21], regulation [27], and facade design [50], they mainly address urban image, with limited attention to billboard colors' impact on pedestrian vitality. Additionally, most studies focus on daytime, neglecting nighttime scenarios.

This study addresses gaps in research on billboard colors by quantifying their effects on pedestrian vitality across different times, days, and functional areas. It introduces an innovative framework that includes street view image collection, semantic segmentation, color extraction, and machine learning techniques to reduce the subjectivity often found in traditional research methods.The findings support the need for improved street design. This includes optimizing the use of billboards, promoting the growth of the nighttime economy, implementing urban renewal policies, and helping traditional streets evolve into experience-oriented spaces to counteract the impact of online retail.

This study is organized as follows: Chapter 2 reviews pertinent literature on color and pedestrian vitality, as well as methodologies for studying the built environment, to identify existing gaps in the research. Chapter 3 introduces the various research domains, along with data acquisition, processing, and analysis methods. Chapter 4 examines the temporal and spatial distribution of pedestrian vitality in commercial streets and discusses the impact of billboards on this vitality. Finally, Chapter 5 presents planning proposals, addresses limitations, and outlines future directions for research.

2 Literature Review

2.1 Color and Pedestrian Vitality

Scholars have proposed various color theories[18], focusing on color's physical properties[52,48], emotional qualities[2,40], symbolic meanings[25,24], and attention effects[20,31].

Color influences the environment through hue, saturation, and other factors, impacting pedestrian perception and vitality. Ding Yuhong pointed out that outdoor advertising color can cause visual pollution, highlighting the need for scientific control to blend with landscapes[10]. Tan Xiao found that warm tones enhance vitality, while cool tones decrease recognizability[35].

Guo et al. (machine learning) showed that color brightness negatively correlates with elderly pleasantness, while complexity and harmony promote positive emotions[16]. Yu et al. observed that high color complexity may cause fatigue, whereas moderate complexity encourages vitality; high harmony enhances beauty and safety[49]. Wu Hao found that 49% preferred "coordinated" harmony, and 37% favored warm yellow nighttime lighting[46].

In recent years, night vitality research, based on multi-source data, has focused on spatial function[41], traffic accessibility[53] and facility configuration[51], but only conducted quantitative statistics on billboards, without exploring variables like area ratio, color complexity and harmony. However, current night economy policies demand accurate visual design guidance—ignoring these variables causes unfocused planning and fails to explain night vitality differences in areas with similar functions and facilities. Thus, it is urgent to study how these billboard variables impact night vitality to support its high-quality development.

2.2 Techniques for Built Environment Color Research

Emerging technologies like street view imagery, semantic segmentation, color clustering, and machine learning advance urban color research and planning by enabling large-scale analysis of the built environment [14]. Street view images, combined with machine learning, support urban color research and its link to pedestrian vitality. Traditional methods, such as colorimetry, are limited by small samples and subjectivity [23,11], while street view techniques allow broader perception. Deep learning models (FCN, PSPNet, DeepLab V3+) segment street features [44], with Google Street View data revealing urban mechanisms [54]. DeepLab V3+ is popular [19,5,15,6,45], and K-Means effectively extracts main street colors [17,13,57,56,9]. Machine learning uncovers nonlinear effects of the environment on vitality. Lian et al. [28] used XGBoost in Beijing and Chengdu, and Cao et al. [4] applied Random Forest (R^2=0.69) in Chongqing, finding positive ecological links. Research has shifted to multidimensional fusion, but faces limitations like ignoring billboard color effects, nighttime scenarios, and direct impacts on vitality. While street view data supports large-scale extraction, micro-studies face data limits and need field surveys [38].

3 Methodology

Figure 1 illustrates the conceptual framework of billboard colors' influence on pedestrian activity within pedestrianized streets. Road network data for Wanlimiao were utilized, with sampling points positioned every 10 meters. Street view images were obtained from simulated pedestrian perspectives. DeepLab V3+ and K-Means clustering algorithms were employed to analyze features of the built environment and billboard color patterns; Wi-Fi probes were used to collect data on pedestrian movement. Nonlinear relationships and threshold effects were investigated to develop comprehensive design guidelines.

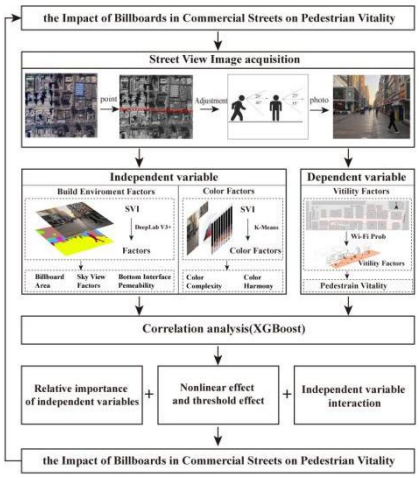


Fig. 1. Framework

3.1 Study Area

As a significant element on Shijiazhuang's traditional commercial route, Wanli Miao Pedestrian Commercial Street extends from Qingnian Street (west) to Gongli Street (east), bordered by Zhongshan West Road (north) and Xinhua Road (south). Encompassing approximately 30 hectares with a main thoroughfare measuring 1,080 meters, its layout connects the historic commercial district with the emerging central business district, thereby establishing a transition zone between historical legacy and contemporary development. This configuration exemplifies how Shijiazhuang's pedestrian streets integrate historical continuity with functions of spatial hubs.

3.2 Data Collection

Street View Data Collection: In urban renewal, improving pedestrian commercial streets is vital. While internet street view images are common for assessing street quality, they have limits—such as delayed updates and imprecise boundaries [22]—

making them unsuitable for detailed research. This study uses self-collected street view data, enabling control over collection timing and spatial accuracy, and recording parameters like time, temperature, and humidity. The Wanlimiao Commercial Streets are 1,080 meters long, with photos taken every 10 meters. Photos followed perspective guidelines matching pedestrian vision: 25° above/35° below horizontally when stationary, and 20° above/40° below when walking [26].After data cleaning, a total of 906 street view images were obtained.

Pedestrian Trajectory Data: The study explored pedestrian paths by collecting all available online street review info to identify high-stay zones via multi-day observations. Field surveys in Shijiazhuang spanned four days, during three daily periods, with 5-minute equipment checks. Data included device MAC, timestamp, and environment, using portable Wi-Fi probes [39] instead of fixed devices for privacy. The probes detect nearby mobile devices via MAC addresses, allowing trajectory mapping with three detectors for precise spatial tracking.

3.3 Data Processing

To better identify elements of the architectural environment, this study employs the DeepLab V3+ model for training, validation, and prediction on the dataset. DeepLab V3+ is a semantic segmentation model developed by Google based on an encoder-decoder architecture, with Xception [47] serving as the backbone network. The encoder features a dual-path design for feature extraction. First, the Xception network directly extracts shallow semantic features from the image. Then, these shallow features are fed into ASPP modules with dilation rates of 6, 12, and 18 to perform multi-scale feature extraction. The decoder initially upsamples the deep features output by the encoder to restore their spatial resolution. These upsampled deep features are then combined with the shallow features from the encoder to integrate local details and global context. The fused feature maps are refined through convolution operations and ultimately produce an output segmentation map that matches the original image dimensions via interpolation upsampling [55]. Data processing of built environment variables.

Data Processing of Built Environment Variables. Based on the research framework proposed by Reid Ewing et al. [12], we measured the pedestrian commercial street environment from three dimensions: imageability, enclosure, and transparency. Specifically, imageability refers to the street's visually distinctive and easily recognizable features; enclosure refers to the sense of spatial boundary formed by vertical elements such as buildings, walls, or trees; transparency indicates the proportion of ground-level doors and windows along the street. Based on these dimensions, we selected billboard area(BA) , sky visibility factor(SVF) , and bottom interface permeability(BP) as built environment variables.Specific calculation methods are shown in formulas (1)- (3).

$$BA = \frac{S_B}{S_A} \tag{1}$$

$$SVF = \frac{S_S}{S_A} \tag{2}$$

$$BP = \frac{S_G}{S_A} \tag{3}$$

Data Processing of Color Variables. Street color environment refers to the overall color scheme and visual impression of urban streets, including the colors of buildings, pavements, greenery, and public facilities. Based on previous studies, this research selected five color indicators for billboard colors as key variables: Billboard Hue (BH), Billboard Saturation (BS), Billboard Value (BV), Color Harmony (Ch), and Color Complexity (Cc)[23].Specific calculation methods are shown in formulas (4)-(8). Image-based studies are increasingly used to analyze street color in commercial districts, with dominant colors significantly affecting visual perception. The K-Means clustering algorithm, an unsupervised learning method, is suitable for extracting dominant colors from architectural images by effectively grouping similar points in the color space [3]. Its process involves: randomly selecting K initial centroids; assigning data points to the nearest centroid (using Euclidean distance) to form K clusters; updating centroids as the mean of each cluster. This process repeats until centroid shifts are below a threshold or a maximum number of iterations is reached, with the final centroids representing the dominant colors in the image. K-Means performs initial clustering in color space without considering spatial relationships for street view images (characterized by block-like color distribution and concentrated pixels). Its advantage is the flexible adjustment of cluster number (K) and iterations to accurately identify dominant colors, making it widely used in urban color studies to extract dominant colors [33]. This study used the HSV color space to align with human visual perception for color data, employing Python's OpenCV to extract color features. BH (0–179) indicates billboard color; BS (0–255) signifies color purity (higher values are more vivid); BV (0–255) denotes brightness.

$$BH = \frac{1}{N}\sum_{i=1}^{N} H_i \tag{4}$$

$$BS = \frac{1}{N}\sum_{i=1}^{N} S_i \tag{5}$$

$$BV = \frac{1}{N}\sum_{i=1}^{N} V_i \tag{6}$$

$$C_C = 1 - \left(\sum_{n}^{i-1} S^2\right) \tag{7}$$

$$C_h = \sqrt{Mean^2 + SD^2 + Skewness^2} \tag{8}$$

Niu T. proposed a quantitative model to measure street vitality using four indicators: pedestrian count, dwell time, trajectory diversity, and complexity. It offers a multidimensional, objective assessment of crowd size, activity duration, spatial

efficiency, and social interaction, surpassing traditional focus on attractiveness and busyness. This comprehensive approach captures key vitality factors, enhances understanding through trajectory analysis, and supports detailed, scientific evaluations of street vitality [30].Specific calculation methods are shown in formulas (9).

$$V = 0.582Num + 0.254Dur + 0.307TD + 0.159TC \tag{9}$$

3.4 Data Analysis

This study employed the extreme gradient boosting (XGBoost) model to analyze billboard color's nonlinear effects and threshold characteristics on pedestrian vitality in commercial streets. XGBoost, an ensemble learning algorithm based on gradient boosting decision trees (GBDT), constructs multiple decision trees and iteratively optimizes the loss function, effectively capturing complex nonlinear relationships, interaction effects, and threshold features between independent and dependent variables.

Following previous research, model hyperparameters were tuned using Bayesian optimization and evaluated via 5-fold cross-validation, with most performance metrics exceeding 0.8. Compared with traditional multiple linear regression, the model demonstrates strong generalization capability, providing reliable predictions for practical applications. Finally, SHAP was employed to interpret the model.

4 Results

4.1 Spatiotemporal Distribution of Pedestrian Vitality

Figure 2 shows the pedestrian vitality heatmap for Wanlimiao pedestrianized commercial street on weekdays and weekends. In the temporal dimension, Figure 3 illustrates average pedestrian vitality across different periods on weekdays and weekends. Vitality on weekends is significantly higher than on weekdays, with evenings being the daily peak. Weekend pedestrian vitality is 1.6 times that of weekdays; at midday on weekends, it is about 2.3 times higher than at the same time on weekdays, and at night, about 1.7 times higher. On weekdays, vitality is generally lower and fluctuates less across periods.

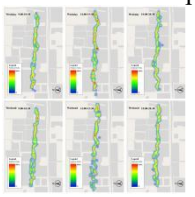


Fig.2. Pedestrian Vitality Heat Map

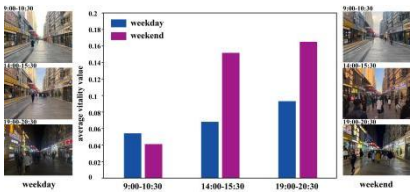


Fig.3. Average pedestrian vitality

Figure 4 displays vitality by measurement unit in the spatial dimension, showing that functional differentiation is associated with vitality differences. The east and west

zones of Wanlimiao exhibit contrasting functions: at midday and in the evening on weekends, vitality in the west is 30 – 40% higher than in the east, due to the concentration of dining and retail activities attracting large crowds. The east, dominated by office functions, lacks round-the-clock consumption scenes and thus shows low vitality outside working hours. On weekday mornings, commuter traffic temporarily causes the east zone to surpass the west in vitality.

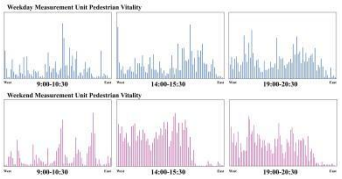


Fig.4. Pedestrian Vitality of Unit

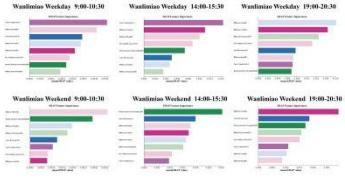


Fig.5. Importance of independent variables

4.2 Importance Ranking of Influencing Factors

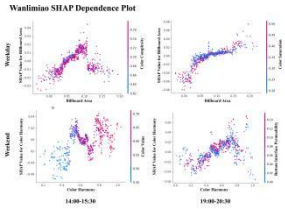
Figure 5 presents the relative importance rankings of factors influencing vitality for the two pedestrianized commercial streets at different times of day. Overall, in Wanlimiao pedestrianized commercial street, feature contributions vary markedly with time. Billboard area (BA) is more important at night, with an importance about 1.7 times that during the day, and it consistently exerts a positive effect on vitality. Billboard saturation (BS) ranks higher in relative importance at night than at midday; on weekends at midday, BS positively influences vitality, but at night it has a negative impact. Color harmony (Ch) ranks relatively high in importance on weekends, while color complexity (CC) consistently ranks among the lowest three. Bottom interface permeability (BP) ranks higher on weekend nights than weekday nights, positively impacting vitality.

4.3 Nonlinear Effects and Threshold Effects

Based on the SHAP dependence plots in Figure 6 and the SHAP summary plot in Figure 7, the factors in Wanlimiao exhibit distinct nonlinear patterns and threshold effects across different times and days.

On weekdays, the influence of billboard area proportion on vitality follows a curve that rises rapidly before leveling off. The effect on vitality is limited when the billboard area accounts for less than 0.05 of the street view image area. When the proportion is between 0.08 and 0.12, SHAP values rise significantly- this range represents the threshold for vitality enhancement. Beyond this range, the improvement effect weakens.

On weekends, color harmony plays a more substantial positive role: when harmony is below 0.4, vitality is low; when it falls within 0.5-0.8, vitality increases markedly.



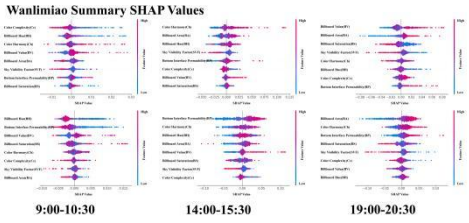


Fig.6. SHAP dependency plot **Fig.7.** SHAP summary plot

4.4 Interaction Effects of Billboard Colors

Based on the interaction summary in Figure 16, the effects of factor interactions on vitality also vary temporally and spatially.

On weekdays in Wanlimiao, the synergy between billboard area and color complexity or saturation is prominent: when the billboard area is 0.10-0.15, samples with high complexity have higher SHAP values than those with low complexity; when the billboard area exceeds 0.15, high saturation further strengthens the effect.

On weekends, interactions between color harmony and billboard brightness or bottom interface permeability are more pronounced: when harmony exceeds 0.6, brightness above 0.65 significantly increases positive contributions; at night, when BP ≥ 0.12, increasing harmony from 0.6 to 1.0 raises SHAP values from 0 to 0.04, whereas when BP < 0.12, the maximum SHAP value reaches only 0.02—half of the high-BP scenario.

5 Conclusion and Discussion

5.1 Discussion

This study's findings align with previous research on urban street vitality. It confirms that commercial clustering and environmental quality increase pedestrian activity, supporting the idea that functional diversity enhances street liveliness. The study also broadens understanding.

First, billboard color elements vary in significance over time, with billboard area being more important at night than during the day, adding to prior emphasis on static elements. Second, a threshold effect (0.08-0.12) for billboard area ratio was identified, supporting evidence that too many visual elements can be detrimental, echoing Chmielewski's outdoor advertising pollution thresholds [7]. Third, different effects of billboard color complexity and harmony challenge the previous focus solely on color richness, showing that harmony is more crucial in leisure settings, consistent with Yu [49]. Fourth, the interaction between bottom interface permeability and color harmony suggests that open facades with harmonious colors enhance vitality, providing stronger support for pedestrian-centered design.

For planners, strategies should align with street functions: Capitalize on weekend night peaks by extending hours, hosting themed events, and enhancing lighting and color coordination. Wanlimiao East District (office-heavy) could implement all-day commercial options to reduce non-work hour lulls.

To prevent visual pressure, keep the billboard area proportion between 0.08 and 0.12; focus on color harmony (0.5-0.8) and minimize overly complex color schemes.

5.2 Limitations and Future Directions

The limitations of this study lie in the case selection, which focused solely on pedestrian commercial streets in Shijiazhuang. Future research could expand the study scope to verify the generalizability of the findings. In addition, this study concentrated only on the relationship between objective elements and pedestrian vitality. Future work could incorporate pedestrians ' subjective perceptions to deepen further the understanding of the mechanisms influencing street vitality.

5.3 Conclusion

Taking Wanlimiao pedestrianized commercial street in Shijiazhuang as the study area, this paper applied street view imagery combined with deep learning, K-Means color clustering, and Wi-Fi probe pedestrian trajectory data. It used the XGBoost model to investigate the influence mechanisms of billboard colors on pedestrian vitality in pedestrianized commercial streets. The main conclusions are as follows:

Significant spatiotemporal differences in vitality. Weekend vitality is approximately 1.6 times that of weekdays, with nighttime being the daily peak. Consumption-oriented activities show a much higher dependence on weekends and nighttime than non-consumption-oriented activities, indicating that functional differentiation is an essential driver of vitality differences.

Time-dependent variations in factor importance. The relative importance of billboard attributes varies with time: at weekend nights, billboard area is about 1.7 times more important than during the day, while color complexity is about 1.2 times more important, reflecting the differentiated impacts of the environment in leisure-consumption contexts.

Nonlinear and threshold effects. The influence of billboard colors on vitality is nonlinear. Vast billboard areas increase visual burden and spatial oppression, while appropriately coordinated colors enhance spatial comfort.

Heterogeneous interaction effects. The synergy between color attributes and other built environment variables varies by time and space. High billboard ratio combined with harmonious colors at night significantly enhances vitality, whereas low permeability reduces the effect by about half.

6 Reference

[1] An, Fengjun. "A Study on the Influence Mechanism of Online Retail on Traditional Retail." Journal of Commercial Economic Research, no. 05, 2024, pp. 15-18.

[2] Azer, Samy A. "The sun and how do we feel about the color yellow? Methodological concerns." Journal of Environmental Psychology 67 (2020): 101380.
[3] Basar, Sadia, et al. "Unsupervised color image segmentation: A case of RGB histogram-based K-means clustering initialization." PLoS One 15.10 (2020): e0240015.
[4] Cao, Yuehao, et al. "Study on the Refined Perception Method of Urban Landscape Based on the Visual Fusion Model: Taking the Main Urban Area of Chongqing as an Example." Chinese Landscape Architecture 41.03 (2025): 76-83.
[5] Chen, Liang-Chieh, et al. "Encoder-decoder with atrous separable convolution for semantic image segmentation." Proceedings of the European Conference on Computer Vision (ECCV). 2018.
[6] Cheng, Shuhong, Jiyong Ma, and Shijun Zhang. "Smoke detection and trend prediction method based on Deeplabv3+ and generative adversarial network." Journal of Electronic Imaging 28.3 (2019): 033006.
[7] Chmielewski, Szymon, et al. "Measuring visual pollution by outdoor advertisements in an urban street using intervisibility analysis and public surveys." International Journal of Geographical Information Science 30.4 (2016): 801-818.
[8] Countryman, Cary C., and SooCheong Jang. "The effects of atmospheric elements on customer impression: the case of hotel lobbies." International Journal of Contemporary Hospitality Management 18.7 (2006): 534-545.
[9] Ding, Meichen. "Quantitative contrast of urban agglomeration colors based on image clustering algorithm: Case study of the Xia-Zhang-Quan metropolitan area." Frontiers of Architectural Research 10.3 (2021): 692-700.
[10] Ding, Yuhong. "A Study on Color Control and Design of Outdoor Advertising in Commercial Streets: Taking Huaihai Middle Road in Shanghai as an Example." Proceedings of the First Asian Color Forum, Department of Art and Design, Fudan University, 2004, pp. 210–215, 259.
[11] Elliot, A. J., and M. A. Maier. "Color Psychology: Effects of Perceiving Color on Psychological Functioning in Humans." Annual Review of Psychology 65 (2014): 95–120.
[12] Ewing, Reid, Susan Handy, and Wenjing Jiang. "Measuring the Immeasurable: Urban Design Qualities Related to Walkability." Urban Planning International 27.05 (2012): 43–53.
[13] Feng, Jingyang, et al. "Quantitative study on color characteristics of urban park landscapes based on K-means clustering and SD method." Earth Science Informatics 17.2 (2024): 999–1012.
[14] Gou, A., and J. Wang. "Research on the location characters of urban color plan in China." Color Research & Application 33 (2008): 68–76.
[15] Goyal, M., and M. H. Yap. "Automatic lesion boundary segmentation in dermoscopic images with ensemble deep learning methods." 2019.
[16] Guo, Huagui, et al. "A Study on the Impact of Street Environmental Color on the Elderly's Pleasure Perception Based on Machine Learning." Landscape Architecture 32.08 (2025): 119–127.
[17] Han, Xin, et al. "Exploration of street space architectural color measurement based on street view big data and deep learning—A case study of Jiefang North Road Street in Tianjin." PLoS One 18.11 (2023): e0289305.
[18] Jean-Philippe Lenclos and Dominique Lenclos. Colors of the World: The Geography of Color. New York: W. W. Norton, 2004.
[19] Ji, Ankang, et al. "An integrated approach to automatic pixel-level crack detection and quantification of asphalt pavement." Automation in Construction 114 (2020): 103176.

[20] Jiménez, María de la Villa Moral, and Celia González Carreño. "Marketing sensorial y perfil del consumidor: la psicología del color en el diseño del producto." Pensando Psicología 18.1 (2022).
[21] Jin, Jianbo. "Color Planning of Luqiao District, Taizhou City Based on Color Psychology." Planner 29.S2 (2013): 125–128.
[22] Kang, Hao, et al. "A Study on the Perception and Evaluation Methods of Street Space Environment Based on Self-Collected Street Views." Traffic and Transportation 39.03 (2023): 72–79.
[23] Kim, J. H., and Y. Kim. "Instagram user characteristics and the color of their photos: Colorfulness, color diversity, and color harmony." Information Processing & Management 56 (2019): 1494–1505.
[24] Kisieliauskas, Justinas, and Evelina Sinevičiūtė. "Colour psychology potential in Lithuanian advertising." Baltic Journal of Economic Studies 9.4 (2023): 1–10.
[25] Li, Heng, and Yu Cao. "Exposure to nature leads to a stronger natural-is-better bias in Chinese people." Journal of Environmental Psychology 79 (2022): 101752.
[26] Li, Xiaofei, and Chunyu Pang. "A Spatial Visual Quality Evaluation Method for an Urban Commercial Pedestrian Street Based on Streetscape Images—Taking Tianjin Binjiang Road as an Example." Sustainability 16.3 (2024): 1139.
[27] Li, Xiaojuan. Research on Urban Color Planning and Control in China Based on Cognitive Image. Tianjin University, PhD dissertation, 2013.
[28] Lian, Haitao, et al. "Pedestrian vitality characteristics in pedestrianized commercial streets—considering temporal, spatial, and built environment factors." Frontiers of Architectural Research 14.3 (2025): 630–653.
[29] Lois Swirnoff. The Color of Cities: An International Perspective. New York: McGraw-Hill Professional Publishing, 2000.
[30] Niu, Tong, et al. "Small public space vitality analysis and evaluation based on human trajectory modeling using video data." Building and Environment 225 (2022): 109563.
[31] Nowghabi, Azadeh Sharifi, and Adeleh Talebzadeh. "Psychological influence of advertising billboards on city sight." Civil Engineering Journal 5.2 (2019): 390–397.
[32] Qi, Z., et al. "The influence of urban streetscape color on tourists' emotional perception based on streetscape images." Journal of Geo-Information Science 26 (2024): 514–529.
[33] Shao, Ronghui. Color Evaluation Study of Hunnan New Town in Shenyang City Based on Urban Street View Data. Shenyang Jianzhu University, MA thesis, 2024.
[34] Shijiazhuang Municipal People's Government Office. "A Notice on Issuing the Work Plan for Promoting the Construction of Characteristic Commercial Blocks in Shijiazhuang (2019—2021)." Shijiazhuang Municipal People's Government Website, 28 Mar. 2019.
[35] Tan, Xiao, Jianfei Chen, and Xiaolei Shi. "A Quantitative Study on Nighttime Visual Perception of Street Space in Cold Region Cities." Low-Temperature Architectural Technology 45.10 (2023): 6–10.
[36] Tang, Jingxian, and Ying Long. "Measuring visual quality of street space and its temporal variation: Methodology and its application in the Hutong area in Beijing." Landscape and Urban Planning 191 (2019): 103436.
[37] Van Rompay, Thomas J. L., et al. "On store design and consumer motivation: Spatial control and arousal in the retail context." Environment and Behavior 44.6 (2012): 800–820.
[38] Wang, Lei, et al. "Measuring residents' perceptions of city streets to inform better street planning through deep learning and space syntax." ISPRS Journal of Photogrammetry and Remote Sensing 190 (2022): 215–230.

[39] Wang, Wei, et al. "Occupancy prediction through Markov-based feedback recurrent neural network (M-FRNN) algorithm with WiFi probe technology." Building and Environment 138 (2018): 160–170.
[40] Wang, Y. B. "Research on the influence of visual communication design based on color psychology on consumers' psychological needs." Psychiatria Danubina 34.Suppl 4 (2022): 1158–1163.
[41] Wang, Yang, et al. "Differences in urban daytime and night block vitality based on mobile phone signaling data: A case study of Kunming's urban district." Open Geosciences 16.1 (2024): 20220596.
[42] Wang, Zhanzhu, Maoting Shen, and Yongming Huang. "Combining eye-tracking technology and subjective evaluation to determine building facade color combinations and visual quality." Applied Sciences 14.18 (2024): 8227.
[43] Wei, J. The Patterning Study Of Commercial Walking District. Master's Thesis, Xi'an University of Architecture and Technology, 2010.
[44] Wu, Dong, et al. "Analyzing the influence of urban street greening and street buildings on summertime air pollution based on street view image data." ISPRS International Journal of Geo-Information 9.9 (2020): 500.
[45] Wu, Hangbin, et al. "Road pothole extraction and safety evaluation by integrating point cloud and images derived from mobile mapping sensors." Advanced Engineering Informatics 42 (2019): 100936.
[46] Wu, Hao. Research on the Application of Color Science in Landscape Design of Pedestrian Commercial Streets. Dalian Polytechnic University, MA thesis, 2013.
[47] Xia, Yuguo, Sheng Ding, and Li Zhao. "A Method for Deep Transfer Recognition of Electronic Components Based on Multi-Scale Attention Mechanism." Radio Engineering 53.09 (2023): 2174–2181.
[48] Yu, Chung-En, Selina Yuqing Xie, and Jun Wen. "Coloring the destination: The role of color psychology on Instagram." Tourism Management 80 (2020): 104110.
[49] Yu, Mingyang, et al. "Urban Color Perception and Sentiment Analysis Based on Deep Learning and Street View Big Data." Applied Sciences 14.20 (2024): 9521.
[50] Zhai, Yujia, et al. "Building facade color distribution, color harmony and diversity about street functions: using street view images and deep learning." ISPRS International Journal of Geo-Information 12.6 (2023): 224.
[51] Zhang, Jinjiang, et al. "Day–Night Synergy Between Built Environment and Thermal Comfort and Its Impact on Pedestrian Street Vitality: Beijing-Chengdu Comparison." Buildings 15.12 (2025): 2118.
[52] Zhang, Ke, Yuansi Hou, and Gang Li. "Color and naturalness: How color saturation shapes tourists' perception and purchase intention." International Journal of Tourism Research 26.4 (2024): e2717.
[53] Zhang, Le, Xueyan Li, and Yanlong Guo. "Research on the Influencing Factors of Spatial Vitality of Night Parks Based on AHP–Entropy Weights." Sustainability 16.12 (2024): 5165.
[54] Zhang, Longhao, et al. "Mechanisms influencing the factors of urban built environments and coronavirus disease 2019 at macroscopic and microscopic scales: The role of cities." Frontiers in Public Health 11 (2023): 1137489.
[55] Zhang, Wenwu, et al. "A Method for Building Segmentation of SAR Images Based on Improved DeepLabV3+." Radio Engineering 55.03 (2025): 475–483.
[56] Zhuang, Yi, and Chenyi Guo. "City architectural color recognition based on deep learning and pattern recognition." Applied Sciences 13.20 (2023): 11575.

[57] 인샤오옌, and 정태열. "관광객 공유한 사진 및 머신 러닝을 활용한 도시 색채 특성 분석 연구-중국 대리시를 대상으로." Journal of the Korean Institute of Landscape Architecture 52.2 (2024): 39–50.

Vulnerable Grid: Strategies and Challenges in Training "Disaster-Resilient Engineers"

Haoyu Liu and Ying Cheng

Hebei University of Engineering, Hebei, China
haoyuliu702@gmail.com

Abstract. The increasing frequency of extreme weather events has exacerbated the vulnerability of power systems, creating an urgent need to cultivate "disaster-resilient engineers" equipped with AI capabilities. However, a profound structural disconnect exists between the rapid iteration of industrial artificial intelligence technologies and the inherent resource constraints and static curricula of engineering education. This paper constructs an "Industrial AI Pedagogical Transformation" framework based on local resource calibration, which transforms complex industrial AI systems (such as eGridGPT dispatch models and GridFM prediction models for extreme weather response) into teachable curriculum components through three systematic pathways: model structure analysis, functional scenario simulation, and engineering capability mapping. Through preliminary practices including attention mechanism visualization, open-source simulation scenario construction, and the "AI-Capability Matrix" assessment tool, the feasibility of this framework has been validated. This paper further proposes a mixed-method validation scheme combining quasi-experimental research, qualitative analysis, and industry expert review, providing clear theoretical foundations and testable practical pathways for effectively integrating industrial AI with resilience engineering education under resource constraints.

Keywords: Industrial AI Pedagogization,Disaster-Resilient Engineers,Climate Resilience,Engineering Education Transformation ,Vulnerable Power Grids,Resource Constraints.

1 Introduction

The era of superintelligence and the climate crisis are jointly restructuring the technological paradigms of critical infrastructure. The increasing frequency and severity of extreme weather events have highlighted the vulnerability of global power systems, placing unprecedented demands on grid disaster resilience and rapid recovery capabilities [1,2]. Against this backdrop, artificial intelligence has rapidly

evolved from a research concept to a core technological driver for enhancing grid resilience [3]. Extensive research confirms that AI excels in tasks such as load forecasting, fault diagnosis, and real-time optimization scheduling under extreme weather conditions [4,13]. Industry-leading practices, such as Transformer-based dispatch strategy generators (eGridGPT) and graph neural network-based equipment health prediction systems (GridFM), clearly delineate the AI-driven landscape of future resilient grids [5,6].

However, a profound paradox emerges: these industrial-grade AI systems that determine industry directions and address climate risks largely exist as "black boxes" in university engineering education classrooms, or are confined to a few graduate-led research projects [7]. The direct consequence is that while graduates master traditional power grid analysis knowledge, they lack the necessary deconstruction capabilities and critical understanding of the core AI tools driving modern power systems for disaster prediction and response, creating a severe "capability gap" [10]. This structural contradiction is particularly acute in smart grids — a domain characterized by high complexity, social criticality, and climate vulnerability.

Current global engineering education reform initiatives, such as China's "New Engineering" construction [8] and MIT's "New Engineering Education Transformation" program [9], clearly point toward the intelligent transformation and industry-education integration of engineering disciplines. However, most existing discourse remains at the strategic level of advocating to "integrate AI" or "cultivate computational thinking," generally lacking specific methodologies for transforming complex, resource-intensive industrial-grade AI models into implementable and assessable teaching modules within typical university resource constraints [10,11]. A "last mile" gap urgently needs to be bridged between the real constraints in computing power, data, faculty, and curriculum systems in education and the industry's urgent demand for "disaster-resilient engineers."

To systematically address this challenge, this paper constructs a structured "Industrial AI Pedagogical Transformation" framework. Premised on local resource calibration, this framework aims to effectively "dimension-reduce" cutting-edge industrial AI systems for enhancing climate resilience into "transparent lesson plans" that cultivate students' higher-order engineering thinking through three core pathways: model structure analysis, functional scenario simulation, and engineering capability mapping. This paper not only preliminarily validates the framework's feasibility through typical cases but also designs a complete mixed-method validation scheme, providing clear theoretical foundations and practical pathways for effectively cultivating "disaster-resilient engineers" oriented toward future climate risks under resource constraints.

2 Requirements and Constraint Analysis

The core challenge of integrating industrial-grade artificial intelligence into engineering education stems from the profound tension between industry demands and educational realities. The global power industry is undergoing a profound

transformation driven by data and intelligence. The integration of high proportions of renewable energy has dramatically increased system complexity, making traditional knowledge systems inadequate. This trend has created an urgent industry demand for a new type of engineer—one who must not only be familiar with traditional principles but also possess the ability to work collaboratively with AI systems and engage in data-driven modeling and decision-making. The World Economic Forum's "Future of Jobs Report (2023)" lists AI and machine learning specialists as one of the fastest-growing positions, confirming this fundamental shift in capability paradigms [12]. At the practical level, from Transformer-based dispatch systems to graph neural network-driven predictive maintenance platforms, AI has been deeply embedded in core industrial processes, requiring engineers to master corresponding interaction, interpretation, and optimization skills. More fundamentally, the quantitative analysis, multimodal inference, and system resilience design required to address power grid uncertainties are evolving from specialized skills into fundamental thinking modes for solving complex engineering problems [5,6].

This transformation has received strong support at the national strategic level. Global initiatives such as China's "New Engineering" construction and MIT's "New Engineering Education Transformation" program clearly point toward the intelligent transformation of engineering disciplines and deep industry-education integration, providing clear policy orientation and practical references for educational reform. However, significant structural barriers exist between vision and reality. At the technical level, there is an order-of-magnitude gap between the computing power required for industrial-grade AI model training and typical university configurations, while core power grid data remains difficult to access for teaching due to security and commercial considerations, and the "black box" nature of models themselves hinders the principle tracing and step-by-step deconstruction necessary for teaching [7]. At the resource and capability level, existing faculty generally face knowledge gaps in transitioning from traditional power system analysis to AI-integrated teaching, while saturated and rigid curriculum systems present enormous resistance to adding new courses or deeply "AI-izing" existing courses through immersive transformation. Finally, traditional written examinations cannot effectively measure students' higher-order thinking abilities in using AI to solve complex engineering problems. How to scientifically assess their problem definition, iterative debugging, and critical validation processes in human-machine collaboration remains a key unresolved challenge [11].

In summary, the strong demands from industry and policy, together with the practical constraints in technology, resources, and assessment faced by education, constitute the core contradiction that this research must directly address and resolve. This contradiction also clearly reveals a critical research gap: there is an urgent need for a systematic "transformation methodology" that can guide educational practitioners to effectively "dimension-reduce" highly complex industrial AI technologies into teachable, learnable, and assessable teaching components within realistic resource boundaries [10].

3 Implementation Pathways of the Industrial AI Pedagogical Transformation Framework

This framework is premised on local resource calibration. Before introducing industrial AI models, it first assesses the implementing institution's computing power, data, faculty, and curriculum status. Based on criteria of interpretability, disaster scenario applicability, and interdisciplinary practicality, we selected eGridGPT and GridFM as representative models for pedagogical transformation (see Figure 1 and Figure 2) [5,6]. The framework achieves the pedagogical dimension reduction of industrial AI through three pathways: model structure analysis, functional scenario simulation, and engineering capability mapping.

Figure 1. Operating principle of eGridGPT

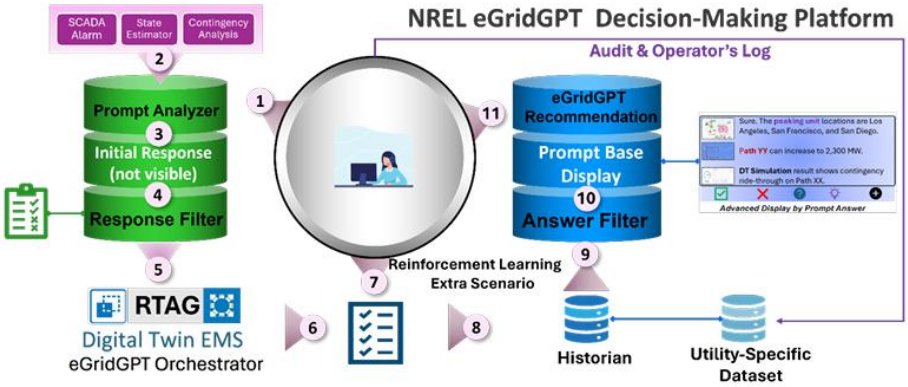
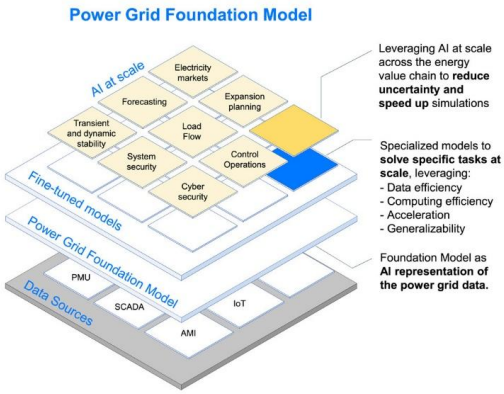


Figure 2. Operating principle of GridFM



3.1 Model Structure Analysis Pathway

This pathway aims to deconstruct the internal logic of complex models, making their core mechanisms traceable and understandable. Taking eGridGPT as an example, its parameter-heavy Transformer architecture is a "black box" in industry, but in pedagogical transformation, we convert it into understandable decision logic by stripping and visualizing its core attention mechanism [5,7].
Teaching Practice: In the "Power System Analysis" course, students input commands simulating extreme weather such as "typhoon passage, surge in load, sudden drop in wind and solar," and can observe the attention heatmaps output by the model. Students can intuitively see which key grid nodes, lines, and historical data at corresponding moments the model's "attention" focuses on when generating dispatch strategies. This transforms an abstract computational task into a visualized discussion about system operational patterns under disaster conditions, fostering deep understanding of AI's decision-making basis.

3.2 Functional Scenario Simulation Pathway

To overcome the limitations of sensitive real grid data and commercial software restrictions, this pathway constructs virtual simulation environments through open-source tools, simulating extreme weather or fault scenarios, allowing students to develop engineering judgment in highly realistic practice.
Teaching Practice: For the GridFM model, a "post-typhoon network reconfiguration" virtual simulation project can be constructed based on open-source tools such as Python, PyPSA, and Google Colab. In this scenario, students use the teaching version of the GNN model to predict equipment health states in the power grid, diagnose faults, and develop recovery strategies [6,14]. This not only trains their skills in using AI tools but, more crucially, cultivates their risk assessment and decision-making capabilities in disaster environments with data uncertainty and adverse conditions.

3.3 Engineering Capability Mapping Pathway

To overcome the limitations of traditional written examinations in assessing higher-order thinking abilities, this pathway develops specialized assessment tools that transform the vague concept of "AI empowerment" into observable and measurable capability indicators.
Teaching Practice: In the "Smart Grid Course Design," an "AI Model-Capability Matrix" assessment rubric was developed. After completing their "Microgrid AI Dispatch Scheme," students must conduct self-assessment and peer assessment based on this matrix across multiple dimensions including problem definition, data preprocessing, model selection and tuning, critical validation of results, and ethical and social considerations. This has preliminarily validated the tool's effectiveness in

transforming the vague concept of "AI empowerment" into observable and assessable capability indicators [10,11].

4 Validation Methodology and Discussion

It should be noted that due to project timeline constraints, this research focuses on framework construction and pathway validation, with systematic effectiveness evaluation planned as the core work for the next phase. To ensure the framework's testability, we have designed a complete validation methodology. Future work will employ quasi-experimental research, using experimental and control groups to quantitatively assess students' gains in core capabilities such as AI modeling and system analysis. This will be combined with in-depth interviews and analysis of teaching artifacts to qualitatively explore practical challenges in framework implementation. Finally, student outcomes will be submitted for blind review by industry experts to complete effectiveness validation from an industrial relevance perspective. This mixed research methodology aims to form an evidence loop from three dimensions—effectiveness, feasibility, and value—not only answering whether the framework "is effective" but also deeply revealing "why it is effective" and "under what conditions it is effective," laying the foundation for theoretical refinement and subsequent dissemination [11].

5 Conclusion and Outlook

Facing the urgent need for power grid resilience construction under the climate crisis and the practical contradiction between the high complexity of industrial-grade artificial intelligence systems and limited engineering education resources, this research has constructed and articulated a systematic "Industrial AI Pedagogical Transformation" framework. Premised on local resource calibration and aimed at cultivating disaster-resilient engineers, this framework completes the operationalized decomposition of typical industrial AI models such as eGridGPT and GridFM into standardized, implementable teaching components through three core pathways: model structure analysis, functional scenario simulation, and engineering capability mapping.

The main contributions of this research are threefold. First, it provides a clear structured methodology that bridges the "last mile" from macro educational reform slogans to micro teaching practices. Second, it profoundly insights into and responds to the practical constraints in technology, resources, and assessment faced by education, making the framework feasible for implementation under typical university conditions. Third, it transcends general assumptions by proposing a complete, multi-dimensional mixed-method validation scheme, providing a clear empirical roadmap for continued research deepening.

Of course, this research has its limitations. The universality and effectiveness of the framework still require more extensive quantitative validation through the

aforementioned quasi-experimental research, long-term case tracking, and industry feedback. Additionally, the framework's adaptability in institutions with different disciplinary backgrounds and resource levels requires further exploration.

Future work will strictly follow the validation pathway described in Chapter 4 to empirically test the framework's teaching effectiveness and continuously iterate and optimize the framework based on feedback results. Simultaneously, we will explore the framework's transfer application potential in other critical infrastructure domains such as transportation and construction, aiming to provide a replicable blueprint for achieving more resilient engineering education transformation oriented toward

References

1. Teng, F., Su, X., Wang, Z., Zhang, L.: Impact of Climate Change on China's Future Electricity Supply. In: Proc. 2023 IEEE 7th Conference on Energy Internet and Energy System Integration (EI2). pp. 4147–4153. IEEE (2023)
2. Panteli, M., Mancarella, P.: The Grid: Stronger, Bigger, Smarter? Presenting a Framework for Power System Resilience. IEEE Power & Energy Magazine. 13(3), 58–66 (2015)
3. Cowls, J., Tsamados, A., Taddeo, M., Floridi, L.: The AI Gambit: Leveraging Artificial Intelligence to Combat Climate Change — Opportunities, Challenges, and Recommendations. AI & Soc. 38, 283–307 (2023)
4. Huntingford, C., Jeffers, E.S., Bonsall, M.B., et al.: Machine Learning and Artificial Intelligence to Aid Climate Change Research and Preparedness. Environ. Res. Lett. 14(12), 124007 (2019)
5. Choi, S., Jain, R., Emami, P., et al.: eGridGPT: Trustworthy AI in the Control Room. Technical Report, NREL/TP-5D00-87740, National Renewable Energy Laboratory (2024)
6. Hamann, H.F., Gjorgiev, B., Brunschwiler, T., et al.: Foundation Models for the Electric Power Grid. Joule. 8(12), 3245-3258 (2024)
7. Wang, Z., et al.: A Trusted Measurement Model of AI Algorithm for Power Grid Scheduling in Predictive Scenarios. In: Proc. 2025 International Conference on Electrical Automation and Artificial Intelligence (ICEAAI). pp. 119–123. IEEE (2025)
8. Graham, R.: The Global State of the Art in Engineering Education. MIT Press, Cambridge (2018)
9. Li, X., Wang, Z., Chu, Z., et al.: A Multi-dimensional Integration Training Model for Outstanding Engineering Talents under the Background of Emerging Engineering Education. Higher Education Development and Evaluation. 41(3), 1-10 (2025)
10. Chakraborty, S., Galatro, D.: Artificial Intelligence (AI) Equality in Engineering Education: Strategies to Unite the AI Gap between the Global North & South. In: Proc. 2025 IEEE Global Engineering Education Conference (EDUCON). pp. 1-10. IEEE (2025)
11. Ortiz, E.A., et al.: AI and Education: Building the Future Through Digital Transformation. OECD Publishing, Paris (2025)
12. World Economic Forum: The Future of Jobs Report 2023. World Economic Forum, Geneva (2023)
13. Fan, S., Li, L., Wang, S., et al.: Research on the Application of Artificial Intelligence Technology in Power Grid Dispatch. Power System Technology. 44(2), 401–411 (2020)

14. Ju, P., Zhou, X., Chen, W., et al.: Review on ‘Smart Grid Plus’ Research. Power System Technology. 38(5), 2–11 (2018)

Teaching Civil Engineering Materials in the Era of AI and Carbon Neutrality

Ying Cheng,Yuyao Li, Haoyu Liu, Ruijie Shi

Hebei University of Engineering, China, University of Malaya, Malaysia, Hebei University of Engineering, China, Yanshan University, China
15231310477@163.com, liyuyao982@gmail.com, 2350277576@qq.com, 1365238660@qq.com

Abstract. This paper explores the challenges faced by the teaching of civil engineering materials in the context of the rapid development of artificial intelligence and the "carbon neutrality" strategy. At present, the teaching mode of the engineering materials course is "emphasizing theory while neglecting practice", which fails to cultivate students' ability to apply knowledge; at the same time, the content of the course is outdated, making it difficult for students to keep up with the cutting-edge developments in the field under the background of artificial intelligence and carbon neutrality. In response to the above issues, this article suggests a phased innovation of teaching content, integrating concepts related to artificial intelligence and the "carbon neutrality" strategy, enriching intelligent teaching methods (such as constructivist pedagogy), clarifying the transformation of university teachers' roles in the era of educational digitalization, aiming to cultivate talents with diverse capabilities, and implementing multidimensional teaching evaluation and assessment systems. Through these measures, it is expected to promote the reform of the civil engineering materials course, thereby cultivating comprehensive talents with a solid theoretical foundation, outstanding practical ability, innovative spirit, and green construction literacy.

Keywords: Artificial Intelligence,Carbon Neutrality,Civil Engineering Materials,Teaching Reform,Smart Education

1 Introduction

With the accelerated evolution of the new round of global technological revolution, the application of artificial intelligence (AI) in the teaching of civil engineering materials courses has become increasingly widespread, covering the application of traditional structural materials, the research and promotion of new engineering materials, as well as the detection and evaluation of engineering materials, thereby promoting the industry toward the direction of intelligence and digitalization [1,8]. Meanwhile, the proposal of China's "carbon neutrality" strategy has driven the transformation of civil engineering materials toward being green and low carbon (see AI-enabled low-carbon concrete design [5,6,9,10]).

For instance, Chinese Academy of Engineering academician Qingrui Yue delivered a report titled "Thoughts on the Development of Civil Engineering Materials in the Context of carbon peak and carbon neutrality" at the Shenzhen Academy of Experts Forum. He pointed out that the carbon emissions generated in the production of civil engineering materials have accounted for more than a quarter of China's total emissions and the task of carbon reduction in the civil engineering materials sector is arduous yet imperative.

The teaching of engineering materials courses should actively respond to the strategy for global climate change governance. In this field, universities have the responsibility to cultivate civil engineers with green environmental protection concepts. The civil engineering materials course, as a core curriculum for cultivating students' future practical skills, its teaching quality directly affects the adaptability of the talents.

However, there are significant shortcomings in the traditional teaching process of civil engineering materials courses. Therefore, based on artificial intelligence and the "carbon neutrality" background, combined with the current teaching situation of the Civil Engineering Materials course, this paper explores the specific paths of teaching reform.

2 The Characteristics of the Civil Engineering Materials Course

This course aims to enable students to master the composition structure, production processes, basic properties, and engineering application characteristics of various engineering materials, to fully understand the intrinsic relationship between material properties and microstructure, and to possess experimental testing capabilities for common civil engineering materials. It lays the necessary theoretical foundation for subsequent professional course learning and engineering practice. The course content covers major engineering materials such as building steel, inorganic cementitious materials, concrete and mortar, and asphalt materials, featuring a wide knowledge coverage, strong theoretical nature, and high degree of interdisciplinary integration. In terms of theoretical teaching, the current course content mainly focuses on the imparting of basic theoretical knowledge of traditional materials; in the practical teaching it mainly consolidates the theoretical teaching content through verification experiments such as cement performance and concrete mechanics.

3 Challenges in Curriculum Reform of Civil Engineering Materials

3.1 Delay in Curriculum Content Update, Failing to Reflect Green Low-Carbon and Intelligent Frontiers

Driven by the "carbon neutrality" strategic goals, green low carbon has become the core direction of development in materials science and engineering [1, 5, 6].

At the same time, emerging technologies such as artificial intelligence and big data are profoundly influencing material research and application. Currently, the content of civil engineering materials courses, including hydraulic cementitious materials (lime, gypsum, sodium silicate), cement, concrete, building mortar, fired bricks, building steel, building plastics, asphalt materials, and wood, all involve industrial production and application processes. Discussions and teaching on energy conservation, recycling, and environmental protection are scarce in these courses, which is one of the manifestations that the update speed of the curriculum content cannot keep up with the pace of development [1]. This lag is not only limited by the long revision cycle of text books and the solidification of knowledge, but also stems from the fact that some teachers lack sufficient understanding of industry frontiers and intelligent tools, making it difficult to integrate green low-carbon and intelligent concepts into the curriculum. As a result, students lack knowledge of green materials and intelligent application skills, making it difficult to meet the demand for interdisciplinary talents who can adapt to the green transformation and intelligent upgrading of the construction industry.

3.2 The Teaching Mode is Monotonous, Lacking Intelligent-driven Interaction and Exploration

The current teaching mode for the civil engineering materials course still mainly relies on the one-way transmission of "teacher's lecture-student's passive acceptance", relying on oral explanations, blackboard writing and slides presentation to impart theoretical knowledge. It lacks interactive, exploratory and personalized learning designs based on artificial intelligence and digital tools. Al though this mode has advantages in the completeness of the knowledge and the controllability of the progress, its inherent drawbacks have significantly restricted students' innovation ability and the cultivation of interdisciplinary comprehensive literacy. In the classroom, students' subjectivity is insufficient, the interaction is low, especially when understanding abstract concepts, they often remain at the stage of mechanical memorization without intuitive cognition and engineering thinking. The practical step lacks support from new means such as intelligent simulation and data-driven analysis, and cannot effectively enhance students' ability to solve engineering problems using AI tools. In addition, due to long-term exam-oriented orientation, insufficient training of teachers on new teaching methods and digital platforms, and high costs of resource reconfiguration, the reform of the teaching mode faces significant resistance.

3.3 The Assessment Model of the Course is Monotonous, Lacking Ability Orientation and Green Intelligent Literacy.

At present, the assessment system of the civil engineering materials course still remains at the simple combination of "final theoretical examination + laboratory report". The evaluation focus is mainly on the memorization and reproduction of theoretical knowledge, while ignoring the comprehensive ability assessment of

students in aspects such as green material design, AI-driven analysis, and engineering case resolution. This exam-oriented assessment leads students to tend to rely on short-term memory to cope with the exam, and teachers compress the practical and exploratory sections to ensure the theoretical scores, forming a vicious cycle of "emphasizing scores but neglecting abilities" and "high scores but low capabilities". In the experimental teaching, due to limited equipment resources and overly large group sizes, students often lack opportunities for practical operation and independent exploration, and the similarity of experimental reports is prominent, with serious deficiencies in the cultivation of innovation and problem-solving abilities. This model cannot reflect the comprehensive literacy that students should possess in the context of "carbon neutrality" and intelligence, and also restricts the pace of engineering education towards the transformation to new engineering disciplines.

4 Teaching Methods and Reform Paths

Artificial intelligence has brought about a revolutionary impact on traditional educational concepts. The structure of vocational education is constantly being adjusted. The contradiction between the quality of the workforce and market demand, as well as the contradiction between learning methods and the realization of educational achievement, have prompted vocational education to develop towards intelligence and automation. Integrating artificial intelligence into the teaching of civil engineering materials is an important measure to adapt to industry technological changes and improve the quality of talent cultivation. Its core value lies in three dimensions: aligning with the characteristics of the discipline, meeting the needs of the industry, and optimizing teaching effects. Through artificial intelligence teaching methods in universities, students can keep up with the development of technological trends and expand their horizons [11].

4.1 Revise Content and Integrate AI and Carbon Neutrality Concepts

Education should keep up with the times. Integrating artificial intelligence teaching content and "carbon neutrality" concepts into the curriculum reform of civil engineering materials is the first step. Firstly, it is necessary to enhance teachers' understanding of artificial intelligence teaching concepts and the basic theoretical knowledge of the latest disciplinary frontiers, so that they can incorporate the concept of using artificial intelligence to cultivate students in the teaching design of civil engineering materials courses and the development trends in the context of "carbon neutrality", thereby cultivating a group of interdisciplinary talents with a solid theoretical foundation, the ability to "apply knowledge flexibly", and the potential for "technological innovation". Secondly, the teaching points of the civil engineering materials course need to be integrated and optimized. The course content should be in line with the low-carbon development requirements of the industry, and should achieve inheritance and development

based on the traditional civil engineering materials course. For example, the physical and chemical properties of cement are key teaching contents in civil engineering materials, and its production process is one of the important reasons for the increase in carbon emissions in the industry.Therefore, when the instructor explains the content of this chapter, they can start by introducing the production process of cement to prompt students to think about the issue of carbon emissions. They can also use the current cutting-edge new type of green and low-carbon cement as an example to stimulate students' thinking on how to design, optimize, and produce low-carbon and green civil engineering materials. They can also utilize waste materials generated in daily life, such as sugarcane residue and fruit shells, to produce alternative products for concrete, thereby inspiring students' exploration spirit and the awareness of "a global community of shared destiny" [5, 6, 10].

4.2 Enrich Methods and Create Diversified Academic Atmosphere

Teaching methods are crucial tools for teachers to impart knowledge and cultivate abilities. By integrating artificial intelligence technology with the curriculum demands in the context of "carbon neutrality", just as instructors have made use of new green, low-carbon concrete alternatives, they could apply the same real-world use of theory in actual construction projects, guiding students to shift from passive reception to active construction. The constructivist teaching concept [7] can be adopted, integrating elements such as context, construction, concentration, ability, and community into the entire teaching process.

- **Contextualized Teaching:**The people, events, objects and their relationships in the learning environment are crucial for learning outcomes. Therefore, teaching design can contextualize learning environments based on real engineering tasks, providing students with opportunities for exploration, design and practice. Artificial intelligence technology offers a new way for contextualized learning. By using machine learning algorithms, it can analyze massive material performance data (such as strength, durability, carbon emission coefficient, etc.), quickly simulate changes in material performance under different ratios and environmental conditions, and help students intuitively understand the "variable-result" relationship, deepening their understanding of material properties [3].
- **Constructivist Learning:** Constructivism emphasizes the formation of mental models during the process of assimilating and adapting to new experiences. When facing unfamiliar knowledge, students need to restructure their existing cognition. The reconfiguration activities include "actual" construction, such as building with blocks, "virtual" construction like drawing on paper and computer screens, and imitation through SimCity simulations. Actual construction can refer to students integrating the learned knowledge into production, while virtual simulation means students using artificial intelligence to simulate the production and application of materials, thereby better understanding the nature of materials. Virtual simulation and AI

visualization technology provide effective support for this process. For example, through dynamic simulation of the microscopic structure evolution of concrete, the corrosion process of steel bars, or the stress distribution of asphalt, abstract concepts are transformed into visual and operational learning objects, thereby deepening understanding and memory [7].
- **Focus and Autonomy:** In the information age, learners need to possess the ability to be self-reliant, autonomous and creative in problem-solving. Teachers can use AI to generate material force short animations to attract attention before class, intersperse situational case tests during class, and assign AI virtual experiments after class, forming a closed-loop learning model of "pre-class guidance-in-class deepening-post-class consolidation". In a respectful and equal classroom atmosphere, students are more likely to remain focused and actively engaged.
- **Skill Enhancement and Diverse Thinking:** The cultivation of students' abilities should be based on diversified development. Artificial intelligence can assist students in efficient information retrieval and integration, expanding the thinking scope for problem-solving, and breaking through a single cognitive mode. By introducing AI data analysis, prediction and optimization functions in the course, students can not only master the core knowledge in the field of civil engineering materials, but also enhance cross-domain thinking and comprehensive abilities.
- **Community and Collaboration:** Group interaction, teacher-student relationships, and the cultural atmosphere of the learning community will all affect learning outcomes. In teaching, teachers can actively guide students to carry out group collaboration. Through discussions, debates and project cooperation, students' thinking horizons are broadened, communication and team collaboration skills are cultivated.

Implementation strategies: Integrate AI into constructivist models using virtual engineering scenarios; utilize digital interactive platforms (e.g. flipped classrooms, which is about students acquiring knowledge through self-study or online learning before class, and then they discuss, solve problems and apply the knowledge with the teacher and other students in the classroom.) to balance teacher guidance and student autonomy; combine classroom discussions with competitions; invite experts and rely on discipline platforms to track frontiers and popularize AI applications [2, 4, 11].

4.3 Implement a Multi-Dimensional Assessment Model

The traditional teaching evaluation model often relies solely on final exams and laboratory reports to assess students' performance. This single evaluation method can easily lead to a one-sided result and is difficult to comprehensively reflect students' daily learning attitude, effort level, and professional ability improvement. Therefore, process-based assessment and outcome-based assessment can be combined. In the classroom participation, project research, and experimental operation sections, artificial intelligence-assisted analysis and carbon

footprint assessment tools can be introduced to comprehensively record and analyze students' learning trajectories and ability growth. At the same time, implement interdisciplinary comprehensive project assessment, integrating the "carbon neutrality" concept into material selection, design optimization, and sustainability evaluation, etc. Encourage students to apply their knowledge to solve low-carbon innovation problems in realistic engineering scenarios. Teachers can appropriately increase the weight of practical courses and focus on cultivating students' ability to analyze and solve practical cases, and actively guide students to participate in professional-related competitions (such as BIM competitions) to enhance their intelligent construction and intelligent material application skills. In addition, include regular grades in the overall assessment. This multi-dimensional, process oriented evaluation system helps promote students' all-round development and lays a solid foundation for cultivating innovative civil engineering talents in line with the times [2, 4].

5 Conclusion

In the context of the rapid development of artificial intelligence and the in-depth implementation of the "carbon neutrality" strategy, the teaching reform of civil engineering materials courses has ushered in new opportunities and impetus. By reconfiguring the course content, integrating the green and low-carbon concepts with intelligent technologies, innovating teaching models, and improving the multi-dimensional evaluation system, students' engineering practical abilities, interdisciplinary comprehensive qualities, and sustainable development awareness can be significantly enhanced. In the future, it is necessary to strengthen the in-depth collaboration between industry, academia, and research, continuously optimize the course, so that it can be highly consistent with the industry's frontiers and social demands, and cultivate high-quality, multi-skilled talents who can lead the green transformation and intelligent upgrade of the civil engineering industry.

References

1. Awolusi, T.F., Finbarrs-Ezema, B.C., Chukwudulue, I.M., Azab, M. (2024). Application of artificial intelligence (AI) in civil engineering. In: *Studies in Systems, Decision and Control*, pp. 15–46. https://doi.org/10.1007/978-3-031-65976-8_2
2. Azam, R., Farooq, M.U., Riaz, M.R. (2024). A Case Study of Problem-Based Learning from a Civil Engineering Structural Analysis Course. *Journal of Civil Engineering Education*, 150(3). https://doi.org/10.1061/jceecd.eieng-1861
3. Diao, P., Shih, N. (2019). Trends and Research Issues of Augmented Reality Studies in Architectural and Civil Engineering Education—A Review of Academic Journal Publications. *Applied Sciences*, 9(9), 1840. https://doi.org/10.3390/app9091840
4. El-Adaway, I., Pierrakos, O., Truax, D. (2014). Sustainable construction education using Problem-Based Learning and Service Learning pedagogies. *Journal of Professional Issues in Engineering Education and Practice*, 141(1). https://doi.org/10.1061/(ASCE)EI.1943-5541.0000208

5. Guo, P., Mahjoubi, S., Liu, K., Meng, W., Bao, Y. (2023). Self-updatable AI-assisted design of low-carbon cost-effective ultra-high-performance concrete (UHPC). *Case Studies in Construction Materials*, 19, e02625. https://doi.org/10.1016/j.cscm.2023.e02625
6. Mahjoubi, S., Barhemat, R., Meng, W., Bao, Y. (2025). Review of AI-assisted design of low-carbon cost-effective concrete toward carbon neutrality. *Artificial Intelligence Review*, 58(8). https://doi.org/10.1007/s10462-025-11182-1
7. Moreno, L., Gonzalez, C., Castilla, I., Gonzalez, E., Sigut, J. (2006). Applying a constructivist and collaborative methodological approach in engineering education. *Computers & Education*, 49(3), 891–915. https://doi.org/10.1016/j.compedu.2005.12.004
8. Nyokum, T., Tamut, Y. (2025). Artificial intelligence in civil engineering: emerging applications and opportunities. *Frontiers in Built Environment*, 11. https://doi.org/10.3389/fbuil.2025.1622873
9. Taffese, W.Z., Hilloulin, B., Zaccardi, Y.V., Marani, A., Nehdi, M.L., Hanif, M.U., Kamath, M., Nunes, S., Von Greve-Dierfeld, S., Kanellopoulos, A. (2025). Machine learning in concrete durability: challenges and pathways identified by RILEM TC 315-DCS towards enhanced predictive models. *Materials and Structures*, 58(4). https://doi.org/10.1617/s11527-025-02664-3
10. Tipu, R.K., Rathi, P., Pandya, K.S., Panchal, V.R. (2025). Optimizing sustainable blended concrete mixes using deep learning and multi-objective optimization. *Scientific Reports*, 15(1). https://doi.org/10.1038/s41598-025-00943-1
11. Zawacki-Richter, O., Marín, V.I., Bond, M., Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education—where are the educators? *International Journal of Educational Technology in Higher Education*, 16(1). https://doi.org/10.1186/s41239-019-0171-0

Psychotherapy as A Near-Deterministic Event: Estimating Reward Frequency (N) Across AI and Human Platforms via Motion-in-Mind

Lulu Gao[1 [0009−0001−6113−0015]⋆], Shize Pan[1], Muhammad Numan[1], Mohd Nor Akmal Khalid[2], and Hiroyuki Iida[1]

[1] Japan Advanced Institute of Science and Technology (JAIST), Nomi 923-1292, Japan
s2420003@jaist.ac.jp

[2] Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM), Bangi 43600, Malaysia
akmal@ukm.edu.my

Abstract. We investigate whether psychotherapy behaves as a near-deterministic process at the session scale when read through Motion in Mind (MiM) and the Variable Ratio identity $N = 1/v$. We harmonize platform-level summaries from six AI therapy apps and four human-provided references, normalizing public ratings to per-session effectiveness $v \in (0, 1]$ and combining them with reported mean sessions to end point T to derive $G = vT$ and reward frequency $N = 1/v$. The results show a tight band $N \in [1.020, 1.316]$ across all entries (AI average $\bar{N} = 1.161$, human average $\bar{N} = 1.093$), indicating dense payoffs in almost every step of the MiM framework. Prior theory in Game Refinement (GR) and Gravity in Mind (GIM) situates therapy on the deterministic side of an informational continuum; convergent findings from digital mental health studies further support session-level benefits. While our dataset is compact and platforms differ in function (therapy oriented vs. companionship) and session duration, the cross-platform clustering around $N \approx 1$ is robust. We outline how extending MiM/GR/GIM with microinteraction metrics may help explain the residual AI–human gap (context sensitivity/empathy) without altering the main descriptive pattern.

Keywords: Motion in Mind (MiM), Game Refinement (GR), Gravity in Mind (GIM), Variable-Ratio reinforcement, Therapy platforms, Reward density, Determinism, Common sense, Empathy

1 Introduction

Many everyday human activities can be interpreted as informational processes that unfold through time, characterized by a certain rhythm of uncertainty reduction and closure. When we play a game, compose a sentence, or conduct

⋆ Corresponding author.

a conversation, each step transforms indeterminate possibilities into structured meaning, producing a sense of progress. Within this view, the dynamics of experience can be formalized through *Game Refinement* (GR) and its cognitive extensions such as *Gravity in Mind* (GIM), which describe how uncertainty resolves with an intrinsic "velocity" and an acceleration toward a sensible stopping point—often referred to as minimal objectivity or the moment of natural completion [1, 2]. Building upon this foundation, *Motion in Mind* (MiM) introduces a quantitative link between information gain and reward expectation, operationalizing per-step effectiveness as $v \in [0, 1]$ and its corresponding reward density through the Variable Ratio identity $N = 1/v$ [3, 4, 6]. Across domains such as board games, sports, and arcade or gacha systems, the N parameter delineates a continuous spectrum—from deterministic environments with dense, predictable payoffs to stochastic ones where rewards emerge sparsely and irregularly.

In this paper, we extend this unified framework from rule-based domains to the realm of psychotherapy, conceptualizing the therapeutic process as an informational event: a structured flow of uncertainty resolution through dialogue. While psychotherapy has often been examined through psychological or clinical lenses, it has rarely been modeled as a near-deterministic system of information transformation. By applying MiM's velocity formalism to session-level data, we propose that psychotherapy exhibits an exceptionally high reward density ($N \approx 1$), meaning that nearly every interaction yields informational or emotional gain—a property that sets it apart from most other human activities.

Question. Where does psychotherapy sit on this spectrum? Theory across behavioral, cognitive, humanistic, and common factor traditions predicts per-session benefits (e.g., activation, empathic receipt, alliance) that feel immediate in the moment. Digital mental health evaluations likewise report short horizon gains and perceived support for structured apps and companion chatbots [9–11, 15, 16, 19]. Therefore, we test whether platform-level data imply a *near-deterministic* session rhythm without presupposing it, i.e., N clustering close to 1 across both AI and human services.

Approach and contributions. (i) We harmonize heterogeneous sources by instantiating $v =$ rating/scale and reading endpoints on a common session unit [17, 18]; (ii) we derive $G = vT$ and $N = 1/v$ and report the empirical distribution per platform; (iii) we interpret the pattern through MiM/GR/GIM as a descriptive placement on the continuum of deterministic stochastic, while noting context-sensitivity alerts for AI systems [34, 35, 13]. Our findings show N tightly clustered around 1–1.3, visually supporting a near-deterministic session process.

2 Previous Research

2.1 Games as informational processes

Game Refinement (GR) and its extension Gravity in Mind (GIM) model interactive activities as informational processes in which uncertainty is resolved over time with a characteristic speed, and an acceleration toward a closure point

termed minimal objectivity [1]. Within this frame, activities are positioned on a continuum from stochastic (rewards are sparse and timing is volatile) to near-deterministic (each step tends to yield an immediate micro–reward). Subsequent work uses the same informational dynamics to account for engagement and addiction mechanisms, emphasizing that experienced momentum derives from structural properties of the activity rather than surface content [2].

2.2 From structured play to general information processes

The GR/GIM framework has been applied across structured human activities to describe how uncertainty, feedback, and closure interact to sustain engagement. In board games and ball sports, characteristic progress profiles and "flow zones" have been identified, revealing how skill-based systems maintain a balanced rhythm of uncertainty resolution [3, 4, 6]. At the opposite end, gacha and other variable-ratio systems deliberately introduce sparse and volatile reward schedules, providing an informational contrast to deterministic play. Together, these comparative benchmarks have helped position diverse activities along a stochastic–deterministic spectrum of experiential rhythm.

Beyond game contexts, the structural triad of speed–acceleration–closure captured in GR and its cognitive expansion, Gravity in Mind (GIM), offers a transferable model for human behavior beyond play [1, 2]. Processes that yield perceptible informational progress, such as artistic creation, dialogue, or learning are often perceived as coherent because they exhibit a similar rhythm of uncertainty resolution and a natural timing of closure. This generality motivates extending the same analytic logic to other human-centered contexts where interaction unfolds as a sequence of informational updates.

Building on these precedents, the present study examines psychotherapy as one such domain. If games provide a closed microcosm of uncertainty resolution, psychotherapy may represent its open-ended counterpart, inviting analysis through the same information framework introduced next.

2.3 Why therapy tends to reward action: theoretical foundations

Across multiple disciplines, therapy exhibits structural properties that make each session highly likely to yield a perceived reward or informational gain.

(1) Psychological micro–reward mechanisms. Traditional psychological models already predict that therapeutic engagement generates immediate reinforcement. Operant conditioning frameworks describe expressive behaviors as self-reinforcing within an empathic, accepting context [20]. Cognitive-behavioral approaches such as Behavioral Activation posit that "acting first" brings near-term emotional improvement [21]. Psychodynamic and humanistic traditions, through mirror empathy and unconditional positive regard, treat emotional attunement as intrinsically rewarding [22, 23]. Self-Determination Theory further shows that the fulfillment of autonomy, relatedness, and competence yields immediate intrinsic-motivation gains [24, 25], while affective-science research demonstrates that verbalization and affect labeling reduce arousal and restore coherence

[26–28]. Together, these mechanisms explain why therapeutic micro-interactions tend to be positively valenced at the session scale.

(2) Behavioral-economic / digital feedback perspective. From a behavioral-economics or digital feedback vantage, psychotherapy sessions form an unusually efficient reward system. Digital mental health interventions (DMHIs) often struggle with retention, but some studies show cases with high engagement and perceived immediate benefits, indicating that frequent feedback can sustain user behavior [29]. Unlike domains where reward is delayed (such as investments or education), therapy delivers rapid, salient micro-feedback (insight, validation, emotional shift) each session, creating conditions analogous to a low-variance variable-ratio schedule that encourages continued participation.

(3) Communication, narrative, and meaning reconstruction. From a narrative and communication perspective, psychotherapy can be viewed as a co-constructive meaning-making process. Each dialogic turn offers informational closure: ambiguity is reduced, self-narratives are revised, and emotional "noise" is translated into structured narrative coherence. This process corresponds to narrative reconstruction observed in psychotherapy research, where clients retrospectively craft coherent stories of change and agency [30, 31]. Systems of meaning-making align with communication theory's emphasis on stories as organizing frames of human experience (narrative paradigm) [32]. Consequently, therapy sessions tend to deliver not just emotional but informational reward in the form of increased coherence and predictability.

Synthesis. Taken together, these converging traditions suggest that psychotherapy embodies a multi-layered reward architecture. It couples intrinsic psychological reinforcement with efficient behavioral feedback and narrative-level uncertainty reduction. Viewed through an information-theoretic lens, this structure supports the hypothesis that psychotherapy approximates a near-deterministic informational event ($N \approx 1$), where nearly every bounded session reliably produces measurable informational gain.

3 Methodological Basis: GR, MiM/GIM, and the Relation $v = 1/N$

3.1 Unifying velocity v across game types

Following the Motion in Mind (MiM) framework, we model the pace at which uncertainty is resolved by a unified *velocity* $v \in [0, 1]$. For scoring games/sports, v is the success rate per attempt; for board games, v is the slope of information progress mapped from the game tree:

$$v = \begin{cases} \dfrac{G}{T}, & \text{(scoring games / sports)} \\ \dfrac{1}{2}\dfrac{B}{D}, & \text{(board games)} \end{cases} \tag{1}$$

where G is the average number of successful events (e.g., goals), T the average number of attempts, B the (effective) branching factor, and D the average game

length in plies.[3] MiM also introduces the *mass* (difficulty) m as the complement of v:

$$m \;=\; 1 - v. \tag{2}$$

3.2 Variable-ratio reinforcement and reward frequency N

Therapy/sports/board play can be viewed under a Variable Ratio (VR) reinforcement schedule, where a reward is obtained after a variable number of responses on average. Let N denote the expected number of steps per reward (reward frequency). Under $VR(N)$, the success probability per step equals the inverse of the expected steps to reward, giving the MiM identity:

$$N \;=\; \frac{1}{v} \quad\Longleftrightarrow\quad v \;=\; \frac{1}{N}, \tag{3}$$

with $N \geq 1$ and $v \in [0, 1]$.[4]

3.3 GIM as a consistency check

Gravity in Mind (GIM) conceptualizes information progress as the intersection of a linear baseline and an accelerated curve. Let $y = vt$ denote baseline progress and $y = \frac{1}{2}at^2$ denote the curvature toward closure (with $a > 0$ as an abstract acceleration parameter). Their crossover time—the *minimal objectivity*—is

$$T^\star \;=\; \frac{2v}{a}. \tag{4}$$

Eq. (4) gives the crossover time at which the accelerated GIM curve $y = \frac{1}{2}at^2$ meets the linear baseline $y = vt$ (see Fig. 1). We use this intersection only as a geometric sanity check. Holding a fixed, $T^\star = 2v/a$ increases with v; thus the intersection itself is not our operational marker of "earlier closure." In this paper, determinism is evaluated from reward density via the MiM identity $N = 1/v$: when $v \to 1$ (hence $N \approx 1$), rewards occur at almost every step and the process behaves as a near-deterministic event. The parameter a is not fitted to domain variables; it merely shapes the red curve in Fig. 1 to visualize how linear per-step effectiveness and accelerated resolution conceptually meet.

4 Data Collection and Processing

To ground our MiM/VR reading of therapy as a near-deterministic process, we assembled platform level records for six AI therapy applications (Woebot, Therabot, Wysa, Youper, Replika, Tolan) and four human delivered references

[3] Eq. (1) is standard in MiM/GR: $v = \frac{G}{T}$ for sports and $v = \frac{1}{2}\frac{B}{D}$ for board games. See [1] for derivations.

[4] In MiM, N is defined as the average trials per reward and is used alongside v and m to quantify reward pacing and difficulty; see [7].

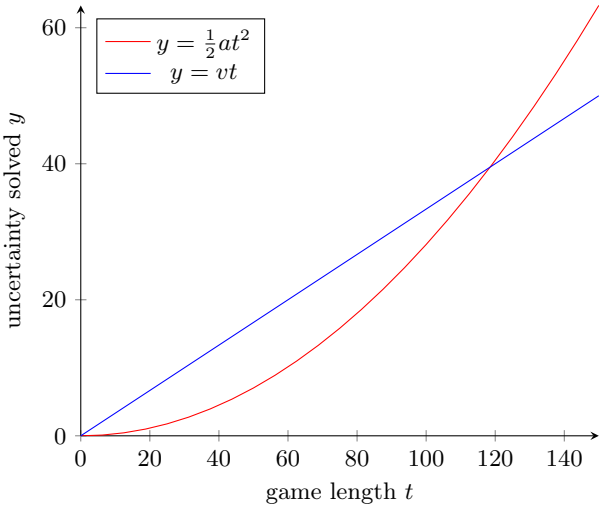


Fig. 1. Gravity in Mind curve (red, $y = \frac{1}{2}at^2$) vs. linear per-step progress (blue, $y = vt$). Their intersection $T^\star = 2v/a$ marks the *minimal objectivity*, i.e., the time when accelerated uncertainty resolution meets the baseline. When v is dense ($N = 1/v \approx 1$), $T^\star$ occurs earlier, indicating a near-deterministic process.

(NHS Talking Therapies, BetterHelp, conventional outpatient therapy, and a Chinese platform Yixinli). The rationale is twofold: (i) recent reports highlight safety and contextual–sensitivity issues in AI counseling [34, 35, 13], which makes it valuable to examine reward density at the platform level; (ii) session based progress is a natural analysis unit across heterogeneous services, consistent with industry usage of "sessions" [18] and with experience oriented evaluations of companion chatbots [16, 19].

Sources and coverage. Peer–reviewed studies were prioritized where available for AI platforms: Wysa and Youper have published evaluations on acceptability, usage, and outcomes [15, 10, 11, 9]. For Woebot we used widely cited public summaries when scholarly estimates were unavailable [14]. For Replika we consulted consolidated statistics to cross–check user base and platform scope [12], mindful of product change effects on user relationships [13]. For human references, we used practitioner/consumer summaries (e.g., BetterHelp) to obtain typical pathways and session counts [17]. When multiple reports existed, we preferred convergent ranges and documented the chosen value in our data sheet.

Endpoint operationalization. Across platforms, the notion of an "endpoint" varies according to service design but always marks a perceived resolution of uncertainty. In clinical trials and regulated interventions, endpoints correspond to symptom reduction or remission on validated scales (e.g., PHQ–9, GAD–7) [9, 11, 15]. Commercial platforms often define closure by completion of a structured program or module pathway [17], whereas open-domain chatbot deploy-

ments typically rely on user-declared goal attainment or disengagement after perceived benefit [14, 12]. To maintain conceptual consistency, all such endpoints are mapped onto a common unit of interaction, the session, which is defined as a bounded exchange window during which identifiable informational progress occurs, irrespective of duration or modality (e.g., 50-minute appointment or short asynchronous chat) [18]. This allows both human and AI services to be compared as discrete iterative events resolving uncertainty step by step.

From v to G and N. To compare heterogeneous platforms on a single MiM/VR scale, we require a common proxy for per–session informational effectiveness, the probability that a bounded interaction yields a subjectively rewarding or meaning–advancing outcome. However, each platform measures satisfaction differently: some employ 5–star or 10–point ratings, others rely on binary feedback ("helpful/not helpful") or structured post–session scales. Directly aggregating such values would conflate scale design with user experience. We therefore normalize all scores by their respective rating scales to obtain a dimensionless quantity $v \in [0, 1]$, representing the proportion of maximal reported satisfaction:

$$v \;=\; \frac{\text{rating}}{\text{scale}} \in [0, 1]. \tag{5}$$

This transformation does not assume linear psychometrics; rather, it interprets user ratings as bounded signals of perceived uncertainty reduction per session. A high v indicates that each session delivers a measurable gain toward closure on average, while a low v implies weaker informational convergence. In this sense, v behaves analogously to success probability in stochastic reinforcement learning, mapping diverse feedback systems into a unified information–processing metric.

Given this normalized effectiveness, we derive the two complementary quantities in the MiM/VR framework:

$$G \;=\; v\,T, \qquad N \;=\; \frac{1}{v}, \tag{6}$$

where T denotes the average sessions–to–endpoint reported by each source. G captures the expected number of sessions experienced as beneficial along a typical pathway, and N expresses the density of reward, which is the expected steps per subjective gain. No determinism is assumed a priori: high or low v values induce corresponding low or high N values mechanically. We report the empirical v and N across all platforms in Table 1 and analyze their distribution in the next section.

Rationalizing heterogeneous sources. Human and AI therapeutic interactions differ in rhythm, pacing, and closure conventions, yet both can be understood as session-bounded information processes. A human session involves a synchronous dialogue with temporal continuity and empathic feedback loops; an AI session, while shorter and often asynchronous, compresses multiple micro-interactions into a single cognitive episode of perceived support or insight. Following [18], we treat the session window instead of its duration as the analytical

Table 1. Platform summary focused on v, $N = 1/v$, and $G = v \cdot T$.

Platform	Rating	Scale	v	T	Lang.	G	N	User number
AI therapy platforms								
Woebot	4.3	5	0.860	12.0	EN	10.320	1.163	1.5 million
Therabot	5.3	7	0.760	8.0	EN	6.080	1.316	—
Wysa	4.7	5	0.940	10.8	EN	10.152	1.064	6 million
Youper	4.36	5	0.872	14.0	EN	12.208	1.147	3 million
Replika	4.3	5	0.860	18.2	Multi	15.652	1.163	10 million
Tolan	4.5	5	0.900	8.3	EN	7.470	1.111	3 million
Human therapist platforms								
NHS Talking	95.00	100	0.950	8.2	EN	7.790	1.053	1.2 million
BetterHelp	4.5	5	0.900	15.0	EN	13.500	1.111	5 million
Outpatient	4.2	5	0.840	8.0	EN	6.720	1.190	—
Yixinli	98.00	100	0.980	6.0	CH	5.880	1.020	3.5 million

Notes: v is the unified per-session speed; G is effective sessions ($G = vT$); N is expected steps-to-reward ($N = 1/v$). Language codes: EN = English, CH = Chinese; "—" indicates data not available. Sources compiled from peer-reviewed and public reports (details in Sec.4).

unit. This normalization abstracts away surface heterogeneity and retains the underlying cognitive function: each session represents an attempt to reduce uncertainty through guided reflection or feedback. Ratings are used only as cross-platform indicators of per-session informational gain, not as claims of clinical efficacy. Temporal variation in platform policies or user demographics is modeled as random noise on v, rather than as a redefinition of the construct itself [34, 35]. This yields a unified empirical framework in which $N = 1/v$ expresses reward density across modalities, supporting the cross-platform comparison in Sec. 5.

5 Results

Per–platform metrics. Using the harmonized definitions in Eqs. (5)–(6), we compute for each platform v = rating/scale, $N = 1/v$, and $G = vT$. The per–platform results are summarized in Table 1. For AI platforms, N values are: 1.163 (Woebot), 1.316 (Therabot), 1.064 (Wysa), 1.147 (Youper), 1.163 (Replika), 1.111 (Tolan). For human references, N values are: 1.053 (NHS Talking), 1.111 (BetterHelp), 1.190 (Outpatient), 1.020 (Yixinli). Across all entries, N lies in a narrow band [1.020, 1.316], i.e., rewards are obtained almost every session on average under the MiM reading of $N = 1/v$.

Group comparison. Figure 2 visualizes N by group (AI vs. human) with individual points and group means. The mean N for AI is $\bar{N}_{\text{AI}} = 1.161$ and for human platforms $\bar{N}_{\text{Human}} = 1.093$ (rounded to 3 d.p.). All platforms sit well inside a near-deterministic band $N \leq 1.5$ (shaded region), clustering tightly

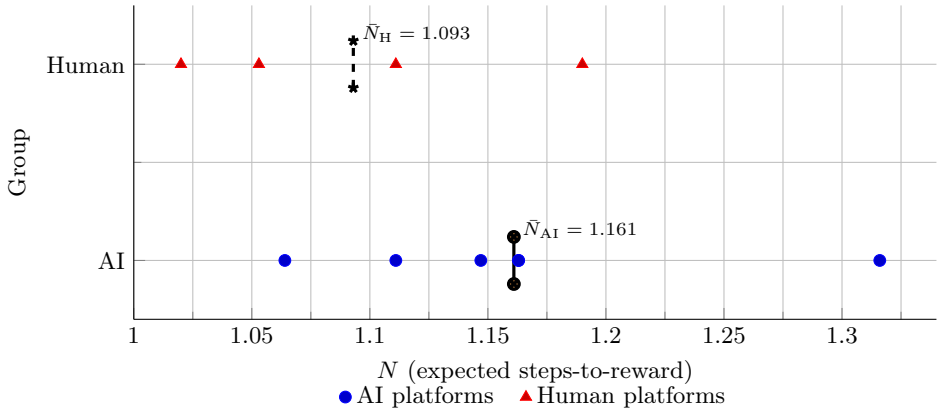


Fig. 2. Per–platform N values ($N = 1/v$) by group with means. Both clusters lie in $N \in [1.02, 1.32]$, indicating dense, near-deterministic reward pacing for AI and human platforms.

around $N \approx 1 \sim 1.3$. While human references show a slightly lower mean N (denser rewards), both groups exhibit high reward density, visually supporting that therapy behaves as a dense, near-deterministic process in the MiM sense.

Effective sessions. The effective–sessions measure $G = vT$ (Table 1) ranges from 6.08 to 15.65 for AI platforms and from 5.88 to 13.50 for human references. Operationally, this indicates that along a typical pathway to the declared endpoint, roughly 6–16 sessions are experienced as satisfactory/beneficial, with human references trending slightly lower N (and thus slightly denser reward pacing).

Ordering by N clarifies the deterministic→stochastic spectrum. Figure 3 arranges domains by increasing N (recall $N = 1/v$ in MiM). Therapy sits at the deterministic extreme: human and AI platforms yield $N \approx 1.09$ and 1.16, respectively, implying rewards (perceived benefits) arrive essentially every step. Borderline–dense activities follow (action games $N \approx 1.52$, Go $N \approx 1.73$, table tennis $N = 2.00$), where feedback remains frequent but not guaranteed each step. Skill + uncertainty mixes (Shogi $N \approx 2.87$, basketball $N \approx 3.70$, chess $N \approx 4.57$) show sparser reward pacing that requires several steps before payoff becomes likely. Finally, high–uncertainty environments (Mahjong $N \approx 9$, Soccer $N \approx 9.09$) and variable–ratio gambling (Gacha $N \approx 45$) occupy the stochastic end where rewards are rare. Overall, therapy's placement near $N \approx 1$ visually anchors the claim that its reward process is *near–deterministic*, while still allowing a nuanced comparison against games and sports on a single axis of reward density.

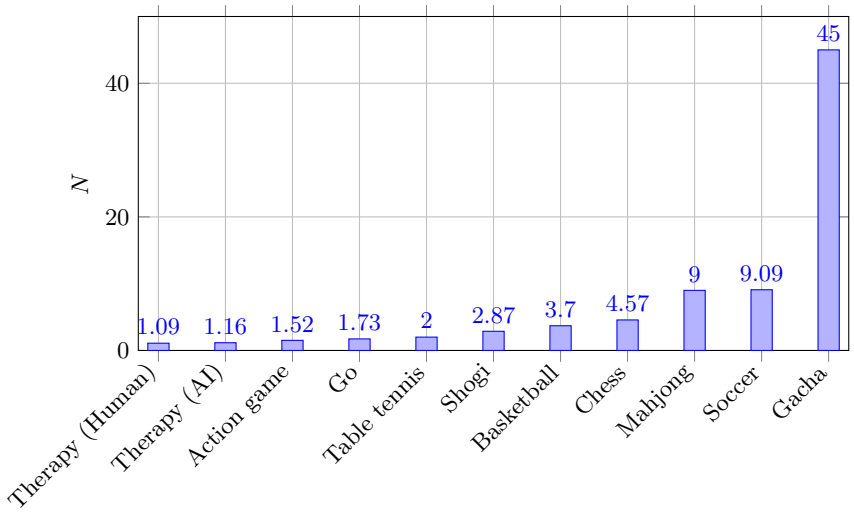


Fig. 3. Cross–domain comparison of N (lower N = denser rewards, more deterministic). Categories are ordered left–to–right by increasing N.

6 Limitations and discussion

Our dataset remains compact and relies on public or secondary sources, constrained by the need for both a normalized satisfaction score (to instantiate v) and a reported mean sessions-to-endpoint T. The notion of a "session" was standardized as a bounded interaction window, consistent with industry usage, but its concrete duration inevitably varies across modalities. In human-delivered therapy, sessions are typically fixed-length appointments (e.g., 45–60 minutes), whereas AI-based interactions unfold through shorter, asynchronous or message-based exchanges that may occur several times per day. This temporal heterogeneity limits the precision of cross-platform comparisons but does not affect the conceptual validity of using per-window v as a normalized measure of effectiveness. Additionally, AI platforms differ in orientation—from structured CBT(Cognitive Behavioral Therapy)/BA(Behavioral Activation) flows to companionship or wellness chatbots—while the human references focus on formal psychotherapy programs. These factors constrain generality, yet the overall clustering of $N = 1/v$ around 1–1.3 across all entries remains robust, supporting the interpretation of therapy as a near-deterministic process at the session scale.

7 Future work

We plan to expand this study by integrating first-hand, session-level data from our ongoing collaboration with an psychotherapy platform. This dataset

includes user-reported satisfaction ratings, session timestamps, and basic contextual descriptors, enabling fine-grained estimation of within-client v (reward velocity) dynamics across successive sessions. By quantifying how perceived reward changes over time, we aim to clarify the temporal structure of psychotherapy as an information process that gradually resolves uncertainty.

Building on the present MiM formulation, we will examine whether variations in v across therapeutic trajectories can predict outcomes such as continued engagement or symptom reduction. Empirical fitting will also allow testing whether $N = 1/v$ remains a stable indicator of reward density when applied to authentic, heterogeneous human–AI interaction data. Cross-validation against available short-form symptom scales (e.g., PHQ–9, GAD–7) will provide convergent evidence for the construct validity of v and N. This real-world expansion will transform the current conceptual framework into an empirically grounded tool for understanding psychotherapy not only as a healing practice but also as a dynamic information flow that embodies the resolution of human uncertainty.

References

1. Iida, H., Khalid, M.N.A.: Using games to study law of motions in mind. IEEE Access **8** (2020).
2. Khalid, M.N.A., Iida, H.: Objectivity and subjectivity in games: Understanding engagement and addiction mechanism. IEEE Access **9** (2021).
3. Xiong, S., Zuo, L., Chiewvanichakorn, R., Iida, H.: Quantifying Engagement of Various Games. In: *The 19th Game Programming Workshop (GPW 2014)*, pp. 101–106 (2014).
4. Iida, H.: What one likes, one will do well. In: *INCITEST 2018 Proceedings* (2018).
5. Numan, M.: *Velocity Dynamics: When Does the Game Get Engaging and Exciting? A Game Refinement Theory Perspective*. Doctoral dissertation, Japan Advanced Institute of Science and Technology (JAIST) (2025).
6. Thavamuni, Sagguneswaraan. *Entertainment Enhancements with Focus on the Difficulty in Stochastic and Skill Games*. Doctoral dissertation, Japan Advanced Institute of Science and Technology (JAIST), Nomi, Japan, [2024].
7. Kang, Xiaohan. *Using Games to Study Psychological Aspects based on Variable Ratio Reinforcement Schedule*. Doctoral dissertation, Japan Advanced Institute of Science and Technology (JAIST), Nomi, Japan, [2023].
8. Gao, Naying. *Computational Measures of Game Entertainment and Reward*. Doctoral dissertation, Japan Advanced Institute of Science and Technology (JAIST), Nomi, Japan, [2024].
9. Mehta, A., Niles, A.N., Vargas, J.H., Marafon, T., Couto, D.D., Gross, J.J.: Acceptability and effectiveness of artificial intelligence therapy for anxiety and depression (Youper): longitudinal observational study. Journal of Medical Internet Research **23**(6), e26771 (2021).
10. Malik, T., Ambrose, A.J., Sinha, C.: Evaluating user feedback for an artificial intelligence–enabled, cognitive behavioral therapy–based mental health app (Wysa): qualitative thematic analysis. JMIR Human Factors **9**(2), e35668 (2022).
11. Sinha, C., Dinesh, D., Heaukulani, C., Phang, Y.S.: Examining a brief web and longitudinal app-based intervention [Wysa] for mental health support in Singapore during the COVID-19 pandemic: mixed-methods retrospective observational study. Frontiers in Digital Health **6**, 1443598 (2024).

12. Roza, N.: Replika AI: Statistics, Facts and Trends Guide for 2025. Online; nikolaroza.com (Apr 2025). Accessed: 23 July 2025.
13. De Freitas, J., Castelo, N., Uğuralp, A.K., Oğuz-Uğuralp, Z.: Lessons from an app update at Replika AI: identity discontinuity in human-AI relationships. arXiv preprint arXiv:2412.14190 (2024)
14. Chatbots Magazine: A therapist bot actually works? Stanford study finds Woebot helps reduce anxiety and depression. Online; Chatbots Magazine (2017), Accessed: 2025-07-23.
15. Inkster, B., Vithlani, R., Narayanan, S., Venkateswaran, R., Aggarwal, J., Masurel, P.-E., Phiri, P., Bhattacharyya, S., Patnaik, S., Prakash, V.: AI-led mental health support (Wysa) for health care workers during COVID-19: service evaluation. JMIR Formative Research, **6**(7), e33500 (2022). https://doi.org/10.2196/33500.
16. Vidal, G., Sim, B., Brereton, M., Roe, P., Leech, M., Ploderer, B.: User experiences of social support from companion chatbots in everyday contexts: thematic analysis. JMIR Formative Research, **6**(9), e35904 (2022). https://doi.org/10.2196/35904.
17. Church, M., Fuller, K.: BetterHelp review 2025: cost, pros & cons, & my experience. Forbes Health (2025, March 12). In-depth practical review of BetterHelp's sign-up, pricing, platform features.
18. Adjust: Insights into what makes a good mobile app session. Adjust Blog (2023), Published on Adjust's official blog.
19. American Psychological Association. Therapists React to Study: AI Perceived as More Compassionate Than Humans. Psychology.org (2025). Available at: https://www.psychology.org/resources/ai-and-empathy/. Accessed 4 Aug 2025.
20. Skinner, B.F.: *Science and Human Behavior*. Macmillan, New York (1953).
21. Jacobson, N.S., Martell, C.R., Dimidjian, S.: Behavioral activation treatment for depression: Returning to contextual roots. Clinical Psychology: Science and Practice **8**(3), 255–270 (2001).
22. Rogers, C.R.: The necessary and sufficient conditions of therapeutic personality change. Journal of Consulting Psychology **21**(2), 95–103 (1957).
23. Kohut, H.: *The Restoration of the Self*. University of Chicago Press (1977).
24. Deci, E.L., Ryan, R.M.: The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. Psychological Inquiry **11**(4), 227–268 (2000).
25. Ryan, R.M., Deci, E.L.: Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. American Psychologist **55**(1), 68–78 (2000).
26. Gross, J.J.: The emerging field of emotion regulation: An integrative review. Review of General Psychology **2**(3), 271–299 (1998).
27. Lieberman, M.D., et al.: Putting feelings into words: affect labeling disrupts amygdala activity. Psychological Science **18**(5), 421–428 (2007).
28. Pennebaker, J.W., Smyth, J.M.: *Opening Up by Writing It Down* (3rd ed.). Guilford Press, New York (2016).
29. Boucher, A.-M., Raiker, J.S.: *Engagement and retention in digital mental health interventions: A scoping review.* BMC Digital Health, **3**(1), 1–15 (2024). https://doi.org/10.1186/s44247-024-00105-9
30. Adler, J.M., McAdams, D.P.: *The narrative reconstruction of psychotherapy.* Journal of Personality, **76**(6), 1231–1260 (2008). https://doi.org/10.1111/j.1467-6494.2008.00517.x
31. Adler, J.M.: *The narrative reconstruction of psychotherapy and psychological well-being.* Psychotherapy Research, **17**(6), 719–731 (2007). https://doi.org/10.1080/10503300701320685

32. Angus, L.E., McLeod, J.: *Narrative in psychotherapy theory, practice, and research: A critical review.* Psychotherapy Research, **24**(3), 293–308 (2014). https://doi.org/10.1080/10503307.2013.845420
33. Wampold, B.E., Imel, Z.E.: *The Great Psychotherapy Debate* (2nd ed.). Routledge, New York (2015).
34. Vincent, J.: AI therapy bots fuel delusions and give dangerous advice, Stanford study finds. Ars Technica (2025). Available at: https://arstechnica.com/ai/2025/07/ai-therapy-bots-fuel-delusions-and-give-dangerous-advice-stanford-study-finds/. Accessed 14 July 2025.
35. Moore, J., Grabb, D., Agnew, W., Klyman, K., Chancellor, S., Ong, D.C., Haber, N.: Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers. In: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, pp. 599–627 (2025).

A Controlled Framework for Generating Synthetic, Multi-Topic Thai Conversations for Healthcare Contact Centers

Kasidis Manasurangkul[1,*], Charnon Pattiyanon[1,*[0000−0003−3660−2962]], Niracha Janavatana[1], Supakorn Etitum[1], Pon Yimcharoen[1], Shine Min Kha[1]

[1]Department of Artificial Intelligence and Computer Engineering
CMKL University, Bangkok, Thailand
kasidism3010@gmail.com, charnon@cmkl.ac.th,
{niracha.mail, supakorn.etitum, pon.yi2025, fraserdoherthy}@gmail.com

Abstract. Developing conversational AI for healthcare contact centers requires large, high-quality datasets that accurately represent domain-specific knowledge while complying with strict privacy regulations. However, collecting real-world healthcare conversations is challenging due to the presence of personally identifiable information (PII) and the limited availability of curated Thai-language resources. To address these challenges, we introduce a controlled, modular framework for generating synthetic, multi-topic Thai conversations grounded in an expert-defined knowledge base. The pipeline incorporates multi-modal preprocessing using a Vision-Language Model (VLM) to convert images and diagrams into descriptive text, BM25-based topic grouping with balancing mechanisms, and a two-stage generation process involving scenario construction and dialogue synthesis using large language models (LLMs). The resulting dataset comprises over 58,000 multi-turn conversations annotated with 500 predefined healthcare topics, ensuring balanced label coverage and realistic interaction patterns. Evaluation with LLM-based scoring confirms strong topical accuracy and linguistic naturalness, with less than 0.08% of dialogues filtered for low quality. Diversity analysis using ROUGE and BLEU indicates structural repetition relative to an open-domain baseline, primarily due to the constrained nature of healthcare knowledge. This work provides a scalable solution for generating privacy-preserving, domain-specific conversational data in low-resource languages and supports the development of robust topic-ranking models for Thai healthcare applications.

Keywords: Conversational AI · Synthetic Data Generation · Healthcare · Thai Language · Large Language Models (LLMs).

1 Introduction

Healthcare systems worldwide provide essential services to patients, encompassing preventive care, diagnostics, treatment, and ongoing support. As these systems grow increasingly complex, non-clinical services (e.g., administrative coordination and customer support) have assumed a more critical role. A key

component of this support infrastructure is the healthcare contact center, where human agents manage calls and perform tasks including appointment scheduling, symptom triage, insurance verification, and other patient-related inquiries. These contact centers frequently serve as the initial point of contact between the public and the healthcare system.

In Thailand, contact centers are important for both public and private healthcare providers, handling high volumes of daily calls. During each interaction, human agents are required to manually complete Customer Relationship Management (CRM) forms by entering patient information, summarizing the conversation, and selecting relevant topics from the knowledge base. These form responses serve as valuable resources for driving continuous improvements in healthcare-related support infrastructure. However, the process is time-consuming, susceptible to errors, and often inconsistent across agents. This results in unreliable information within the CRM system, which may fail to accurately reflect the true quality of service provided by these contact centers.

With recent advancements in AI technologies, such as Automatic Speech Recognition (ASR), Large Language Models (LLMs), and Machine Learning-Based (ML) topic classification, these tools have emerged as promising solutions to address existing inefficiencies. ASR can be employed to transcribe recorded conversations into text, facilitating further analysis. LLMs are particularly effective for summarizing call content and evaluating the quality of agent interactions. ML-based topic classification enables the retrieval of relevant information to support agents in real time. When properly designed and supported by sufficient training datasets, these technologies can significantly enhance the efficiency and reliability of healthcare contact centers.However, developing such systems at scale necessitates access to large, high-quality conversational datasets. In the healthcare domain, collecting real-world conversations poses significant challenges due to concerns over personally identifiable information (PII), such as national identification numbers, addresses, and sensitive medical details, which are safeguarded by strict privacy regulations. In Thailand, this issue is further exacerbated by the limited availability of curated healthcare and conversational datasets, thereby hindering the progress of AI development in this field.

Synthetic data generation using Large Language Models (LLMs) has emerged as a promising alternative, allowing for experimentation without compromising privacy. A realistic synthetic dataset offers researchers and developers greater flexibility to train and implement tools more efficiently. However, current approaches predominantly target open-domain tasks, such as instruction-response generation, general dialogues, or classification with limited label sets [1–3]. These methods typically leverage broad, open-domain sources, most often text-based platforms like Wikipedia or online forums, to maximize linguistic diversity. Such pipelines, however, are poorly suited for domain-specific applications, such as healthcare services, which demand multi-label classification and precise label integration. Existing methods do not support complex label structures and often fail to ensure strong alignment between the generated content and the target labels, which are an essential requirement for supervised learning.

Focusing on conversational AI in the healthcare domain, we aim to train and evaluate a topic classification and ranking model that incorporates multiple topics. This is necessary because actual call patterns tend to be unpredictable, with a high likelihood that callers will inquire about multiple issues within a single conversation, driven by curiosity or the need for comprehensive information. Based on an investigation of call history in the call center of a nationwide healthcare service provider in Thailand, we observed that the number of topic categories can reach approximately 500 and continues to grow over time. Each topic is linked to a reference known as Knowledge Base (KB) documents, which may include text, images, or both. All generated conversations must be grounded in this constrained knowledge base to ensure relevance and factual consistency.

The combination of a narrow domain and a large label space presents several challenges for training conversational AI systems for Thai contact centers. The limited scope of reference materials further constrains topical diversity, in contrast to traditional synthetic data pipelines. Additionally, many domain-specific documents contain non-textual elements—such as diagrams or scanned forms—which reduce the amount of usable information for language model inputs. Ensuring balanced topic representation is also critical, as label imbalance can bias model training outcomes.

Synthesizing high-quality data in Thai adds another layer of complexity, as Thai remains underrepresented in most multilingual LLMs, resulting in reduced fluency, lower generation quality, and less stable outputs [4]. This underrepresentation makes it particularly challenging to generate coherent conversations that accurately integrate multiple constrained topics.

To address this range of challenges, this paper proposes a controlled, modular framework for synthesizing multi-topic Thai healthcare-related conversational dialogues grounded in a domain-specific KB. The framework is organized into two main phases within a system operation pipeline. In the first phase, the existing KB documents are preprocessed to generate textual descriptions of non-textual elements using a visual language model. In the second phase, a dataset is created through a combination of scenario pre-construction from KB topics and dialogue generation using LLMs. The resulting dataset comprises Thai conversational dialogues annotated with predefined healthcare-related topics and is evaluated in terms of diversity (using ROUGE [5] and BLEU [6] metrics), relevance, and naturalness. The main contributions of this paper are summarized as follows:

- This paper proposes an AI-based data preparation method for non-textual elements by converting them into descriptive alternative (alt, for short) text, enabling structured processing of multi-modal information.
- This paper introduces a novel two-phase prompting workflow using LLMs to generate and synthesize realistic Thai multi-topic conversational dialogues in the healthcare domain, creating opportunities for research and practical applications requiring domain-specific data for further analysis.
- This paper releases a synthesized dataset of Thai multi-topic healthcare conversational dialogues that mitigates the risk of exposing personal or sensitive healthcare-related information during the development of AI models.

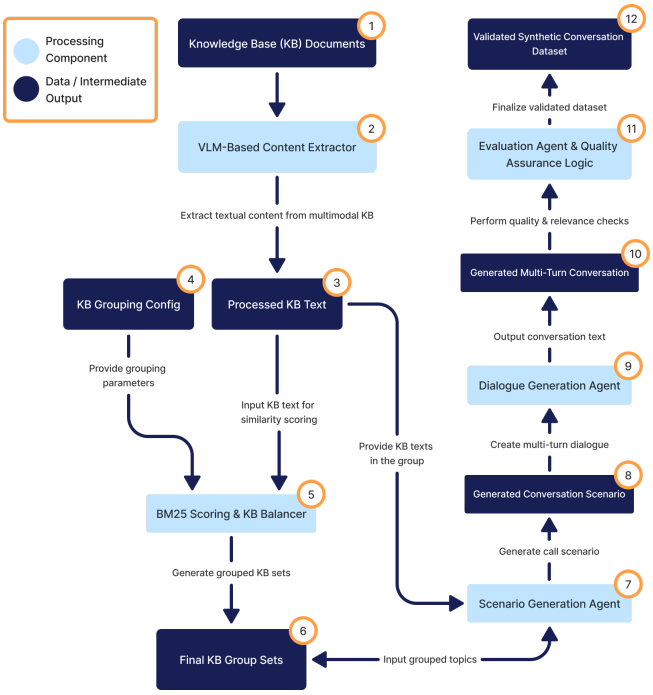


Fig. 1. System overview pipeline for synthesizing Thai healthcare conversations.

The remainder of this paper is organized as follows: Section 2 presents a detailed description of the proposed framework; Section 3 reports the experiments and results; Section 4 discusses the advantages and limitations of the framework; and Section 5 concludes the paper.

2 Methodology

2.1 Framework Overview

The proposed framework is structured as a two-phase system pipeline, as illustrated in fig. 1. The left-hand side represents the first phase (#1-#6), while the right-hand side corresponds to the second phase (#7-#12).

The first phase begins with the preprocessing of KB documents (#1), which may contain both textual and non-textual elements. In Step 2, a Visual-Language Model (VLM) [7] is employed to analyze the non-textual elements in the KB documents and generate descriptive alt text for each element. The details of this step are provided in Section 2.2. The textual content and generated alternative text are then compiled into a preprocessed KB document (#3). After introducing the grouping configuration (#4), Step 5 applies the BM25 Okapi scoring function [8], combined with a balancing mechanism, to calculate document-topic similarities

and organize KBs into well-structured topic clusters, resulting in grouped KB sets (#6). The details of the topic selection and grouping mechanisms are discussed in Section 2.3.

We assume that data synthesis is more effective when dialogue generation is guided by a pre-constructed scenario. In the second phase, the grouped KB set (#6) and the processed KB set (#3) are used to pre-construct scenarios that define the caller's context, intents, and interaction flow, using a small LLM in Step 7. The resulting scenarios (#8) resemble dialogues in structure but do not yet resemble realistic conversations. In Step 9, a larger LLM is employed to synthesize realistic multi-turn dialogues (#10) based on these pre-constructed scenarios. The details of the scenario and dialogue generation processes are provided in Sections 2.4 and 2.5, respectively. Finally, as the framework is designed to ensure the generation of a high-quality Thai conversational dialogue dataset, Step 11 is dedicated to evaluating and validating the synthesized dataset using various metrics. In this study, the proposed framework was applied to the specified healthcare-related KB documents, and the resulting synthesized dataset was evaluated using the metrics described in Section 3.

2.2 Knowledge Base Preparation

The knowledge base (KB) documents referenced in this paper refer to materials that consolidate the organizational and operational knowledge of an organization, ensuring continuity during staff turnover. KB documents can exist in various formats, such as Word documents, spreadsheets, PDF files, or relational database tables. They may also contain non-textual elements, such as diagrams, charts, images, or tables.

In this study, 468 KB documents in DOCX format were collected from a nationwide healthcare organization in Thailand that operates a contact center serving patients across the country. Each KB document corresponds to a predefined healthcare topic and contains both textual and visual representations, such as treatment flows, dosage charts, and procedural instructions, that contact center agents reference and communicate to callers.

The primary purpose of this preparation step is to convert the non-textual elements of KB documents into textual representations. Manually describing these elements is impractical given the large size of the KB. In this study, the KB documents are first preprocessed from DOCX files into a machine-readable format. Images within the files are then extracted and processed using the Gemini 2.0 API [9], guided by a structured system prompt specifically designed to produce descriptive alternative text for each image. The generated alternative text provides detailed interpretations of page layouts, capturing instructions from diagrams, embedded text, and labels in a consistent, context-preserving format. Following this image-to-text conversion, the generated alternative text is reintegrated into the corresponding KB document, replacing the non-textual element at its original location. This ensures that the KB document is normalized and consolidated into a clean text file. An example of KB document preparation and alternative text generation is illustrated in fig. 2, showing a figure described in

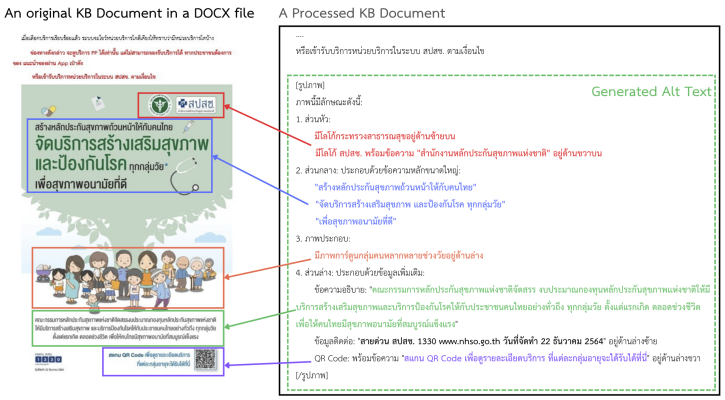


Fig. 2. An example of the interpretation of non-textual elements into alt text.

text. Notably, even when represented in Thai, the text contained in the image is accurately transcribed in the alternative text.

2.3 Topic Selection and Grouping

In a typical KB, it is common for multiple documents to be of similar topic. For example, the 14th KB document may address the universal coverage of healthcare services policy in Thailand, while the 15th KB document may discuss the use of the Line application for inquiries related to the same policy. Considering real-world application, it is more beneficial if scenarios and dialogues are generated across different but related topic groups, ensuring that the conversation content remains natural and the transition from one topic to another flows smoothly. Therefore, the objective of this step is to create combinations of similar topics that naturally complement each other and can be used to generate conversations covering a more diverse range of subjects, rather than focusing on a single topic. Since all KB documents are now in text format, a text similarity calculation approach is applied to identify which documents are most similar to others. The BM25 Okapi scoring algorithm [8] is used to measure the textual similarity of each document against all others. The most similar KB documents are then selected as candidate pairs for group formation. The number of top candidates is dynamically determined based on a predefined configuration. For instance, specifying "`topic_per_group: 2`" and "`expected_records_per_group: 5000`" means that each group will be associated with two similar topics. In this experiment, we generate combinations containing between one and four topics to better represent real-world datasets, where conversations may not always revolve around a single topic.

This approach allows candidate selection to scale appropriately with the requirements of conversational dialogue generation. To ensure coverage across the entire KB, the algorithm iterates over all KB documents, setting each document as the primary KB in turn when forming candidate groups. The process

then loops repeatedly until the number of generated combinations meets the "`expected_records_per_group`" configuration requirement. However, it is important to be aware of and prevent the over-representation of certain KBs, ensuring that no single topic from one KB document dominates a cluster. To address this, this paper introduces a penalty mechanism that reduces the similarity score of KB documents that have already been used multiple times. The adjusted score can be calculated as follows:

$$\text{AdjustedScore}_{p,c} = \text{BM25}_{p,c} - (\alpha \times U_c) \tag{1}$$

where $\text{AdjustedScore}_{p,c}$ denotes the adjusted similarity score between the primary KB_p and the candidate KB_c, $\text{BM25}_{p,c}$ represents the original similarity score computed between KB_p and KB_c, α is a constant that controls the penalty applied to overused KBs, and U_c indicates the total number of times candidate KB_c has been assigned to any group.

The penalty mechanism reduces the likelihood of frequently similar KBs dominating the dataset by deducting their similarity scores when they have been selected multiple times. This approach is particularly useful for mitigating bias in downstream tasks. At the end of this step, the process outputs a list of topic groups along with the KB documents contained within each group.

2.4 Scenario Construction

Since large language models in the Thai language generally exhibit lower fluency and limited diversity in generated content when synthesizing conversational data, this study separates the process into two distinct tasks: scenario construction and dialogue generation. The objective of this separation is to allow each model to focus more effectively on a specific aspect of generation, thereby yielding higher-quality outputs.

In this section, scenarios refer to a structured sequence of stories or topics that a synthesized call dialogue should address. While they may resemble conversations, scenarios primarily outline the overall flow and context without including detailed exchanges. The scenario generation task focuses on producing diverse conversation content and ensuring natural transitions between topics, rather than emphasizing linguistic fluency. An example is shown on the left-hand side of fig. 3, where a scenario specifies that a caller inquires about services at a drug store covered under their healthcare plan and seeks information on changing their registered healthcare service center.

For constructing these scenarios, this study found that Gemma-3 14B [11] performed best after experimentation with other LLMs capable of processing Thai text, including Qwen2.5-14B-Instruct [10] and Google's Gemma-3-27B-it [11]. The selected model consistently produced contextually accurate scenarios (based on Thai native speakers' manual evaluation) while remaining feasible within reasonable computational resources. The processed KB documents and KB group list are provided to the model with a structured prompt, and the model then constructs a collection of scenarios.

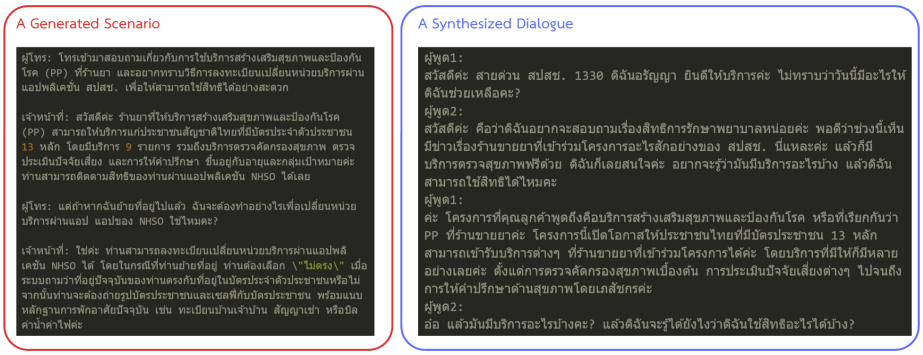


Fig. 3. An example of a generated scenario and a synthesized call dialogue in Thai.

2.5 Dialogue Generation

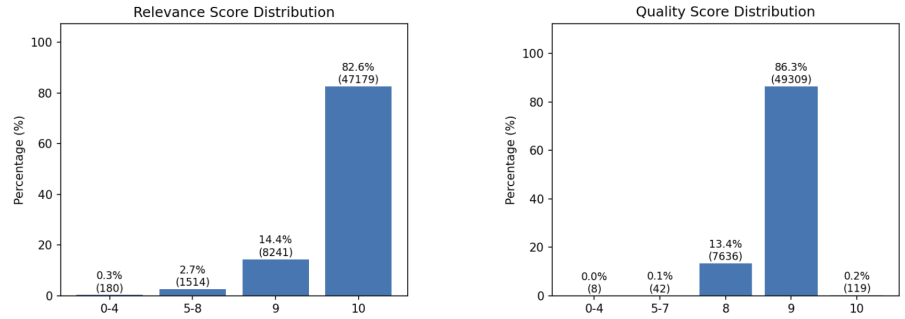
The constructed scenarios can be directly used to synthesize call dialogues. The ultimate goal of the proposed framework is for the synthesized dataset to contain realistic, multi-topic call exchanges.

This study identifies Gemma-3 27B [11] as the most effective model for dialogue generation in terms of both accuracy and efficiency. This conclusion follows an extensive evaluation of several LLMs, including Llama-3.1-70B [13], Qwen2.5-72B-Instruct [10], Qwen2.5-14B-Instruct [10], Gemma-3-14B-it [11], and Gemini 2.0 API [9]. In contrast to scenario construction where the model must process a large set of KB documents, the dialogue generation works with short-text scenarios, allowing the use of a larger LLM without excessive computational cost.

At this stage, each scenario is fed into the selected LLM using a structured prompt that instructs the model to generate a call dialogue between two speakers in accordance with the provided scenario. The synthesized output is a complete text-based conversation, as illustrated on the right-hand side of fig. 3. This example demonstrates that the model is capable of producing realistic, natural-sounding dialogues, as confirmed by evaluations from native Thai speakers. At the end of this step, the proposed framework has achieved its goals to provide a way to generate a realistic chat dialogue for Thai contact centers.

3 Experiments and Results

After obtaining the synthesized dataset from the system pipeline, it is essential to ensure that its quality meets the requirements for practical adoption. This paper evaluates the dataset within the context of Thai healthcare-related contact centers across three dimensions: (1) relevance of the dialogues to the assigned topics and their linguistic fluency, (2) diversity of the generated dialogues, and (3) bias in topic distribution.



(a) Relevance Score Distribution of the Synthesized Call Dialogues

(b) Quality Score Distribution of the Synthesized Call Dialogues

Fig. 4. Score Distribution of the Generated Call Dialogues.

3.1 Relevance of the Topics and the Fluency of the Generation

Relevance of the topics is defined as the extent to which the synthesized dialogues accurately reflect their assigned KB topics. Fluency, in contrast, measures the coherence of the generated dialogue and the natural integration of multiple topics within a single conversation. The fluency score is sometimes referred to as a quality score. Both the relevance score and the fluency score range from 0 to 10, where a score of 10 indicates that the dialogue is fully aligned with the assigned topics (for relevance) or that the topics are seamlessly and naturally integrated (for fluency). In this study, a Generative Adversarial Network (GAN)-inspired approach is employed, in which another LLM acts as a discriminator to evaluate the correspondence between each synthesized dialogue and its assigned topics. Specifically, a Gemma-3 27B [11] model is provided with the dialogue and instructed to assign both a relevance score and a fluency/quality score.

The resulting relevance scores are summarized in the bar chart shown in fig. 4(a). Analysis of the distribution reveals that only 1,694 out of 57,114 dialogues (approximately 3%) received a relevance score lower than 9, indicating that the vast majority of generated dialogues are highly relevant to their assigned topics. On the other side, the resulting fluency/quality score are summarized in the bar chart shown in fig. 4(b). It shows the similar result compared to the relevance score where only 50 out of 57,114 dialogues (approximately 0.08%) received the fluency/quality score lower than 8.

3.2 Dataset Diversity

Another key quality aspect of a synthesized dataset is diversity, defined as the extent to which the dataset avoids containing redundant or near-duplicate dialogues. To quantify diversity, this study adopts ROUGE [5] and BLEU [6] metrics, which measure the similarity between pairs of dialogues. In this context, a higher similarity score indicates greater redundancy and thus lower diversity.

Table 1. Resulting Dataset Diversity Scores.

Dataset	Total Comparisons	ROUGE-1	ROUGE-2	ROUGE-L	BLEU
Ours	3,695,258	**0.2596**	**0.1237**	**0.2324**	**0.1357**
Baseline	3,515,340	0.0215	0.0022	0.0201	0.0234

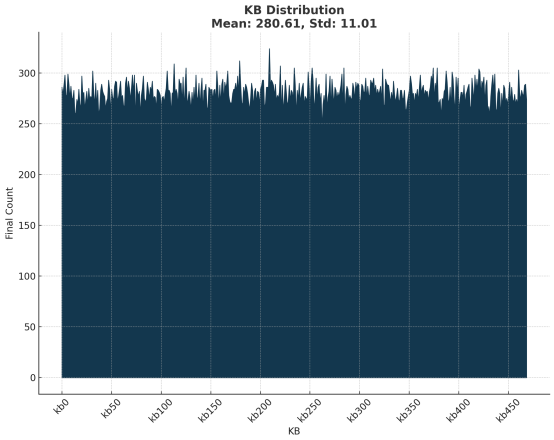


Fig. 5. Distribution of Synthesized Dialogues Per KB Documents.

To provide a meaningful reference point, the *WangchanThaiInstruct Multi-turn Conversation Dataset* [14], a widely recognized Thai synthetic conversation dataset, is used as a benchmark. The goal of this evaluation is to compare the ROUGE and BLEU scores of the proposed synthesized dataset against those of the baseline dataset, thereby assessing whether the proposed framework produces more diverse conversations.

This paper sampled 10% of dialogue pairs that has the same assigned KB documents to minimize computational requirements, which have high possibility to be redundant. Moreover, to ensure a fair comparison, speaker labels are dropped and the first three lines of each dialogue are removed since it typically contains repetitive greetings according to the controlled instruction of the contact center. table 1 shows a collection of ROUGE and BLEU scores for both the synthesized dataset and the benchmark dataset. It depicts that the synthesized dataset in this paper has lower diversity than the baseline, as reflected by higher similarity scores across all metrics. This outcome is likely due to structured KB-based grouping in our generation pipeline, which enforces consistent topic coverage but leads to more uniform phrasing and conversational patterns.

3.3 Bias Measurement

To ensure that the synthesized dataset is balanced across KB documents, and to prevent potential bias during model training and evaluation, the number of gen-

erated conversation dialogues associated with each KB document was counted and their distribution plotted. The results in fig. 5 show a nearly uniform distribution, with the gap between the most and least represented KBs being approximately 40 dialogues. This balance ensures that no single KB disproportionately dominates the dataset, maintaining broad and topic coverage.

4 Discussion

The evaluation results show that the proposed framework can be implemented efficiently as a system pipeline, which performs strongly in relevance and quality, with most generated conversation dialogues meeting or exceeding the defined thresholds, and without introducing noticeable bias across KB topics. This confirms that the synthesis process, i.e., scenario construction followed by dialogue generation, produces contextually accurate conversations with natural flow.

However, diversity scores were lower than expected compared to the benchmark dataset [14], which was drawn from general-purpose sources with fewer constraints. The narrower domain and limited KB document pool inherently increase lexical and structural similarity. Additionally, the dialogue generation process followed a consistent interaction pattern (greeting $\rightarrow$ problem description $\rightarrow$ agent response), which likely contributed to higher similarity scores.

Another factor was the decision not to apply chunking to large KB documents. Without segmentation, the LLM may repeatedly focus on similar content, producing repetitive structures. The use of a single prompt template and identical one-shot examples for each number of topics also reinforced uniformity. Since the one-shot example was manually created without real-world dialogue samples, it may not fully capture the diversity found in authentic interactions.

While quality scores remained high, qualitative review revealed areas for improvement. The LLM struggled with natural transitions when handling more than four topics per conversation, despite real-world dialogues sometimes spanning up to eight topics. Generated conversations were also shorter, typically equivalent to three to five minutes of text, compared to real-world dialogues that can extend up to twenty minutes of text, reflecting both coherence challenges and computational constraints. These findings highlight trade-offs between resource efficiency and realism, suggesting potential gains from refined prompt design, document chunking, and diversified generation strategies.

5 Conclusion and Future Direction

This study introduces a modular, controlled framework for synthesizing Thai healthcare contact center conversation dialogues grounded in a domain-specific knowledge base. By combining topic grouping, scenario planning, and dialogue generation, the system produced a realistic conversation dialogue dataset aligned with expert-defined topics while maintaining label balance and strong relevance

and quality scores. Although diversity remains a challenge due to domain constraints and uniform prompt design, the dataset provides a practical resource for training and evaluating topic-ranking models in Thai healthcare contexts.

Several directions can further enhance the quality and diversity of generated conversations. First, we plan to apply chunking to KB documents before feeding them to the LLM. This approach will mitigate context length limitations and increase topic diversity, as the model will only have access to a portion of the document at a time, reducing the tendency to repeatedly select the same sections. Second, we aim to create multiple system prompts with variations in tone, structure, and length, tailored for different conversation types. This will introduce stylistic diversity and make the outputs more representative of real-world interactions. Additionally, replacing synthetic one-shot examples with real conversational examples can improve realism and enhance topic transitions, resulting in more natural and contextually rich dialogues.

References

1. Ren, J., Du, Z., Wen, Z., Jia, Q., Dai, S., Wu, C., Dong, Z.: Few-shot LLM Synthetic Data with Distribution Matching. In: Companion Proceedings of the ACM on Web Conference 2025 (WWW '25). Association for Computing Machinery, New York, NY, USA, 432–441. (2025). https://doi.org/10.1145/3701716.3715245
2. Moe, L., Nguyen, U.T., Luu, B.T.: Mitigating Class Imbalance in Fact-Checking Datasets Through LLM-Based Synthetic Data Generation. In: Proceedings of the 4th ACM International Workshop on Multimedia AI against Disinformation (MAD' 25). Association for Computing Machinery, New York, NY, USA, 73–80. (2025). https://doi.org/10.1145/3733567.3735571
3. Jagatap, A., Merugu, S., Comar, P.M.: Improving Search for New Product Categories via Synthetic Query Generation Strategies. In: Companion Proceedings of the ACM Web Conference 2024 (WWW '24). Association for Computing Machinery, New York, NY, USA, 29–37. (2024). https://doi.org/10.1145/3589335.3648299
4. Pipatanakul, K., et al.: Typhoon 2: A Family of Open Text and Multimodal Thai Large Language Models. ArXiv Preprints. (2024). https://doi.org/10.48550/arXiv.2412.13702
5. Lin, C.-Y.: ROUGE: A Package for Automatic Evaluation of Summaries. In: Proceedings of the Workshop on Text Summarization Branches Out. Barcelona, Spain (2004). https://aclanthology.org/W04-1013/
6. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: BLEU: a Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 311–318. (2002). https://doi.org/10.3115/1073083.1073135
7. Li, J., Li, D., Xiong, C., Hoi, S.,: BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. ArXiv Preprints. (2022). https://doi.org/10.48550/arXiv.2201.12086
8. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M.M., Gatford, M.: Okapi at TREC-3. In: Proceedings of the Third Text REtrieval Conference (TREC-3). (1995).
9. Comanici, G., et al.: Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. ArXiv Preprints. (2025). https://doi.org/10.48550/arXiv.2507.06261

10. Yang, A., et al.: Qwen2.5 Technical Report. ArXiv Preprints. (2024). https://doi.org/10.48550/arXiv.2412.15115
11. Kamath, A., et al.: Gemma 3 Technical Report. ArXiv Preprints. (2025). https://doi.org/10.48550/arXiv.2503.19786
12. Kwon, W., et al.: Efficient Memory Management for Large Language Model Serving with PagedAttention. In: Proceedings of the 29th ACM Symposium on Operating Systems Principles. (2023). https://doi.org/10.1145/3600006.3613165
13. Aaron Grattafiori, A., et al.: The Llama 3 Herd of Models. ArXiv Preprints. (2024). https://doi.org/10.48550/arXiv.2407.21783
14. Thammaleelakul, S., Wannaphong, P.: WangchanThaiInstruct Multi-turn Conversation Dataset. Zenodo. (2024). `https://zenodo.org/records/13132633`

Breaking Free from the GPA Assembly Line

Rebuilding Individual Growth Paths in the Era of Superintelligence

ZHANG Yuxuan[1][0009-0005-2369-2382], NIU Jiaxing[2][0009-0002-5791-4201] and SHANG Jinghan[1][0009-0008-1011-725X]

[1] School of Mathematical Sciences and Physics, Hebei University of Engineering, No. 19 Taiji Road, Handan 056107, Hebei, China
[2] School of Architecture and Art, Hebei University of Engineering, No. 19 Taiji Road, Handan 056107, Hebei, China
yuxuanhue@qq.com

Abstract. In the era of rapidly advancing superintelligence, traditional higher education faces fundamental challenges in balancing general and specialized education while fostering creativity and transferable skills. Drawing an analogy between human learning and large language model (LLM) training, this study highlights parallels between Supervised Fine-Tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF) with standardized versus interest- and goal-driven learning. While SFT mirrors rote memorization, limiting generalization and creativity, RLHF promotes autonomous problem-solving and knowledge transfer. Building on these insights, we propose AI-driven personalized education systems that integrate adaptive evaluation, dynamic learning paths, and teacher–AI collaboration. Such systems enable multidimensional competency assessment, flexible specialization, and holistic student development. Finally, we envision a future university model where AI empowers lifelong, individualized learning, supporting creativity, critical thinking, and the cultivation of unique talents.

Keywords: Superintelligence; Higher Education Reform; Large Language Models; Personalized Learning; AI-Driven Assessment; SFT; RLHF; Teacher–AI Collaboration.

1 Introduction

In the current era, where the advent of "superintelligence" is rapidly approaching, the challenges faced by higher education extend beyond the iterative updating of curricula; they lie more fundamentally in the reshaping of its core mission and evaluation logic. Over the past century, university education has upheld the ideal blueprint of "a balance between general and specialized education," striving to cultivate "well-rounded talents" with broad vision and cross-disciplinary competence through comprehensive curricula and a unified Grade Point Average (GPA) standard. However, technological innovation, the increasing specialization of industrial division of labor, and employers'

preference for readily deployable skills have led to growing doubts about the effectiveness and legitimacy of this traditional model.

Standardized GPA systems compel students to spread efforts evenly across subjects, often sacrificing deep engagement in areas of interest to offset weaknesses in others. This suppresses intrinsic learning, stifles creativity, and limits cross-domain knowledge transfer.

In recent years, research on the training of large language models (LLMs) shows that excessive SFT boosts targeted skills but weakens generalization, mirroring the trade-offs of standardized education. RL, emphasizing exploration and feedback, can enhance core abilities while maintaining transfer capacity[1].

This study asks whether, in the superintelligence era, universities should retain the GPA-centered model or use AI to build flexible, fair systems that support deep specialization. When personal goals clash with standardized models, institutions must balance being "enablers" and "intelligent governors."

To address this question, the main contributions of this paper are as follows:

- **Proposal of an analogy framework:** Establishing a mapping between LLM training mechanisms and human learning patterns, revealing the potential erosion of transferability and creativity caused by standardized requirements.
- **Integration of research evidence:** Systematically reviewing the differences between SFT and RLHF in terms of training effectiveness, transfer performance, and catastrophic forgetting, and translating these findings into reference signals for educational reform.
- **Combination of institutional and technological solutions:** Proposing a dynamic GPA and personalized evaluation framework based on AI-generated learning profiles, preserving essential general education while granting students greater freedom to delve deeply into areas of interest.
- **Forward-looking risks and governance pathways:** Examining potential challenges such as algorithmic bias, educational equity, and the empowerment of teachers and students, along with governance recommendations and pilot strategies.

This paper proceeds as follows: Section 2 examines LLM–brain analogies; Section 3 links training strategies to educational reflections; Section 4 proposes university reform strategies; and Section 5 discusses universities' role with superintelligence.

2 The Human Brain and Large Language Models: Unexpected Kindred

2.1 Analogies in Learning Processes

LLMs undergo massive unsupervised pretraining on broad text corpora (so-called "foundation models") [2], akin to how human infants and children absorb language patterns through passive exposure. Both rely on extracting statistical regularities from rich input. Indeed, LLMs have been described as "resurrecting associationist principles" of learning[3], mirroring theories that humans learn language via repeated

exposure and pattern induction. While it is true that LLM training methods were partly inspired by human learning processes, our analogy aims to reflect on potential implications for human education, rather than suggest a direct causal relationship.

After pretraining, LLMs are supervised-fine-tuned on specific tasks, analogous to formal schooling or targeted training in humans. In both cases, corrective feedback on errors refines the system's representations: for humans this may be teacher correction, while for LLMs it is gradient descent on labeled examples.

LLMs can be further adjusted by reinforcement learning from human feedback (RLHF), which shapes output towards preferred behavior. This parallels operant conditioning in humans and animals, where positive/negative feedback (rewards or punishments) modify future responses.

2.2 Knowledge Formation and Generalization

Cognitive neuroscience shows that human concept formation engages multiple brain systems (hippocampus, prefrontal cortex, etc.) and flexibly builds abstractions from examples[4]. For instance, medial prefrontal cortex encodes the abstract structure of a task across different instances, while the hippocampus maps that schema to specific details[5]. Humans thus form hierarchical, generalized representations (schemas, prototypes) that support transfer to new situations. LLMs likewise build distributed concept representations through exposure, enabling broad generalization in many tasks[3]. However, LLMs may lack explicit analogical mechanisms; their "generalization" arises from interpolation in vector space rather than symbolic schema.

Sensorimotor Grounding: Neuroscience emphasizes that human knowledge is grounded in multimodal experience (sensory, motor, social contexts). Recent work finds that text-only LLMs capture non-sensorimotor aspects of concepts reasonably well, but fail to recover the sensory/motor features that humans learn through embodied experience. For example, without visual or motor input, an LLM has little basis for understanding how to smell a flower or move in space, whereas human concepts integrate these dimensions. Models trained with multi-modal data reduce this gap (Xu et al., 2025)[6].

3 Research Echoes: The Mirror Between AI Transferability and Educational Models

3.1 SFT and RL—From "Giving a Man a Fish" to "Teaching a Man to Fish"

The training of large language models (LLMs), typically through Supervised Fine-Tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF), offers a useful analogy for higher education: SFT mirrors traditional, standardized knowledge transmission, whereas RLHF reflects personalized skill development, highlighting contrasting educational philosophies.

The SFT stage resembles the traditional process of knowledge impartation, in which the model passively learns from a dataset composed of high-quality "question–standard answer" pairs[7]. Its goal is to replicate examples with precision, striving to imitate specific knowledge or stylistic patterns. This parallels the "standard answer" pedagogy in education—akin to "giving a man a fish"—which focuses on ensuring that students master predefined knowledge points[8].

RLHF introduces a value-driven mechanism in which models, guided by a reward model trained on human preferences, explore and optimize behaviors to produce high-quality content. This approach, akin to "teaching a person to fish," cultivates autonomous judgment and problem-solving skills.

This section compares the two training paradigms and highlights how personalized, goal-oriented education is essential for developing comprehensive student competencies in the superintelligence era.

3.2 Paradigm Comparison: The Limitations of SFT and the Generalization Advantages of RL

SFT builds foundational instruction-following skills, while RL refines behavior to align with complex human preferences, differing in mechanisms, objectives, generalization, and risks.

Table 1. Comparative analysis of Supervised Fine-Tuning (SFT) and Reinforcement Learning (RL) in large language model training, highlighting differences in mechanisms, goals, data requirements, generalization capabilities, risks, and costs.

Evaluation Dimension	Supervised Fine-Tuning, SFT	Reinforcement Learning, RL
Core Mechanism	Supervised learning is conducted on high-quality instruction–response pairs, aiming to maximize the log-likelihood of the target responses.	Learns a reward model and applies reinforcement learning algorithms to optimize the language model's policy in order to maximize expected rewards.
Optimization Goal	Local optimization: Imitates specific expressions and knowledge contained in the dataset.	Global optimization: Acquires general strategies for producing outputs aligned with abstract values rather than imitating specific samples.
Transfer & Generalization	Has limited generalization capability, tending to "memorize" patterns from the training data, and may perform poorly in new	Exhibits stronger generalization capability; by learning the principles underlying the reward signals, the model can better handle unseen tasks and

Evaluation Dimension	Supervised Fine-Tuning, SFT	Reinforcement Learning, RL
	domains or on unseen instruction types[7].	demonstrate cross-domain adaptability.
Capability Degradation Risk	Carries the risk of catastrophic forgetting or the so-called alignment tax, where excessive fine-tuning may degrade foundational abilities acquired during pre-training.	Also carries the risk of capability degradation, but this can be mitigated through carefully designed reward structures to preserve foundational abilities[9].

SFT's main limitation is its narrow focus: overfitting to training exemplars can cause fragility and reduce generalization[7]. Task-specific fine-tuning may also impair other abilities, a phenomenon termed capability degradation or alignment tax[9]. For example, conversational fine-tuning can diminish analytical skills like math or coding.

RLHF offers a flexible, holistic optimization by evaluating response quality against human preferences. This promotes deeper understanding, enhancing robustness and generalization for complex or novel tasks. Despite risks like reward model bias[10], RLHF fosters transferability, making models versatile problem-solvers.

3.3 Empirical Evidence: The Evolution of Model Capabilities under Different Training Paradigms

Recent research shows that while SFT is necessary for skill acquisition, RLHF uniquely enhances generalization, safety, and alignment with human values.

1. Ouyang et al. (2022) — The InstructGPT Study[11]: As a pioneering work in the RLHF domain, OpenAI's study clearly demonstrated the superiority of RLHF. They first trained a base model using SFT and then optimized it via RLHF. The results revealed that, despite having far fewer parameters than contemporary SFT models (such as GPT-3), the RLHF-optimized model achieved significantly higher "helpfulness" and "truthfulness" scores in human evaluations. A key finding was that simply enlarging the SFT dataset could not effectively address the issues of "hallucination" (i.e., generating false information) and harmful content, whereas RLHF substantially suppressed such undesirable behaviors. This indicates that RLHF instilled in the model a set of "behavioral principles" rather than merely imparting knowledge.
2. Touvron et al. (2023) — The LLaMA 2 Study[12]: In the technical report accompanying Meta's release of the LLaMA 2 series, a detailed comparison was provided between SFT and RLHF (which incorporated rejection sampling and the PPO algorithm). Experimental results validated RLHF's value across multiple dimensions. In human evaluations of "helpfulness" and "safety," the LLaMA-2-Chat model fine-tuned with RLHF scored far higher than its SFT-only counterpart. Notably, the report also highlighted the phenomenon of alignment tax: in certain academic

benchmarks (e.g., MMLU), the zero-shot performance of RLHF models was occasionally slightly lower than that of SFT models. This suggests that while RLHF makes a model more "cooperative" and "safe," it may do so at the expense of part of its original, unconstrained knowledge-retrieval capability—underscoring the trade-offs inherent in the optimization objectives of the two paradigms. Figure 1 conceptually illustrates this performance difference based on the trends observed in the study.

3. Rafailov et al. (2023) — Direct Preference Optimization (DPO)[13]: This work proposed a simpler and more stable alternative to RL—DPO—which directly leverages preference data to fine-tune an LLM using a straightforward classification loss, thereby bypassing the need to train a separate reward model and perform complex reinforcement learning. The key contribution lies in showing that preference-based optimization (whether RLHF or DPO) is far superior to SFT in controlling model behavior. Experiments demonstrated that DPO could achieve more precise and reliable control than SFT over the sentiment, style, or level of detail in generated summaries, while avoiding the common pattern-collapse issues seen in SFT. This further confirms that value- and goal-driven optimization (RL/DPO) cultivates more generalizable and controllable capabilities than example-based imitation (SFT).
4. Wang et al. (2024) — Reward Function Complexity in RLHF[10]: A study published at ICLR 2024 delved into the design of reward models in RLHF. The findings revealed that a carefully crafted, multi-objective reward function—e.g., simultaneously rewarding helpfulness, truthfulness, and penalizing verbosity—can guide models to acquire more nuanced capabilities. This stands in sharp contrast to the single imitation objective of SFT, which is incapable of handling such multi-objective, and potentially conflicting, optimization tasks. These results underscore the intrinsic potential of the RL paradigm in navigating the complex value systems encountered in real-world applications.

SFT builds a model's knowledge base, while RL instills generalized behavioral and value-judgment abilities; relying solely on SFT may yield knowledgeable but rigid models.

3.4 Cross-Domain Analogy: From Model Training to the Cultivation of the Human Mind

The contrast between SFT and RLHF parallels Rote versus Meaningful Learning: Rote learning, like SFT, emphasizes passive memorization and reproducing standard answers to maximize exam scores. In the short term, this method is efficient for tackling standardized tests and rapidly acquiring specific factual knowledge[8]. However, its drawbacks align closely with the limitations of SFT:

- Knowledge acquired through rote memorization is isolated and unstructured, lacking deep connections to the learner's existing cognitive framework. As a result, such knowledge is easily forgotten and difficult to retrieve in novel, non-standardized contexts—demonstrating poor *knowledge transfer* capability.

- An excessive emphasis on memorization and reproduction crowds out the time and cognitive resources needed to develop critical thinking, creative thinking, and complex problem-solving skills. This is analogous to the "capability degradation" observed with SFT, where students may achieve high scores in a specific subject but show stagnation in comprehensive analytical and innovative abilities.
- When learning degenerates into an externally imposed task, with the sole "reward" being grades or the avoidance of punishment, students' intrinsic interest in learning and their natural curiosity are gradually eroded[14].

Interest- and goal-driven learning, like RLHF, is fueled by curiosity and long-term goals. Educators create motivating environments and offer timely feedback, helping students align strategies with personal objectives[15].

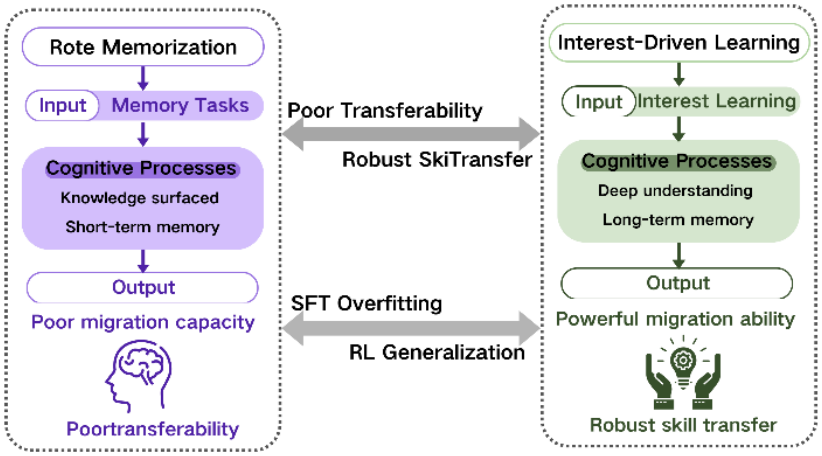


Fig. 1. presents a conceptual model that visually contrasts these two learning pathways and their differential impacts on cognitive development.

- Knowledge Construction through Interest: Learning based on genuine interest encourages students to actively construct knowledge, integrating new information with prior understanding to form robust cognitive schemas. Such deeply processed knowledge is more durable and flexible, facilitating transfer and application in novel contexts[16].
- Metacognitive Development through Goal-Driven Learning: In goal-oriented learning, students continuously plan, monitor, evaluate, and adjust their learning processes, thereby exercising metacognitive skills. They acquire the ability to learn how to learn—teaching them to fish—a higher-order skill that can accompany them throughout life.
- Enhanced Motivation and Well-Being: According to Self-Determination Theory, when individuals' needs for autonomy, competence, and relatedness are satisfied, their learning motivation and psychological well-being are significantly enhanced[17]. This intrinsically motivated learning experience itself constitutes a central aim of education.

4 Reconstructing University Education: Specialization, Autonomy, and AI Empowerment

4.1 AI-Driven Personalized Education Systems

Artificial intelligence can enable personalized instruction through the use of learner models. A learner model is employed to estimate a student's current knowledge state and proficiency. For instance, Bayesian Knowledge Tracing (BKT) treats the mastery of skills as a latent Markov process, allowing the probability of skill acquisition to be updated according to the following formula:

$$P(K_t) = P(K_{t-1})(1 - p_s) + \big(1 - P(K_{t-1})\big)p_\ell$$

Here, p_s and p_l denote the probabilities of slip and learning, respectively. Similarly[18]. Item Response Theory (IRT) estimates a student's ability θ using a logistic function:

$$P(\text{Right}|\theta) = \frac{1}{1 + e^{-a(\theta - b)}},$$

Here,a and b represent the item's discrimination and difficulty parameters, respectively. In practice, parameters a (discrimination) and b (difficulty) can be estimated from historical student performance data or calibrated with expert judgments, allowing adaptive learning systems to adjust content delivery accordingly.Based on these cognitive models, adaptive learning systems can dynamically tailor instructional content for individual students.

Learning paths can be dynamically tailored using knowledge graphs and reinforcement learning, while recommendation systems adaptively suggest materials based on student profiles, enhancing engagement, retention, and efficiency[19]. A scoring function can be defined, for example:

$$\text{score}(s, c) = \sigma(w_s \cdot x_c),$$

where w_s denotes the student feature vector, x_c represents the content features, and $\sigma(\cdot)$ is an activation function. AI-driven ITS use scoring functions to recommend content and dynamically adapt to students' learning states, offering timely personalized guidance[20].

4.2 Reconstructing the GPA System

Traditional GPA systems overlook the learning process and individuality [21], prompting universities to adopt "de-quantified" assessments that emphasize holistic competencies over single-score evaluations [22].

Competency-based multidimensional assessment, often implemented through competency maps and portfolios, offers a comprehensive view of students' strengths and gaps[23]. In fields such as medical education, portfolios document practice data and reflections, though flexible design and guidance are needed to ensure representative evidence of abilities.

At a mathematical modeling level, a multidimensional composite evaluation function can be designed. Suppose student i has mastery levels $C_{i1}, C_{i2}, \ldots, C_{iK}$ across K competency dimensions. A simple scoring model can be expressed as:

$$E_i = \sum_{k=1}^{K} w_k \ C_{ik}$$

Where w_k denotes the weight of the k-th competency. By dynamically adjusting these weights and incorporating nonlinear transformations, the model can be adapted to individual student profiles. For instance, to balance depth and breadth, a weighted average of the competency scores can be computed as follows:

$$E_i = \frac{\sum_{k=1}^{K} w_k C_{ik}}{\sum_{k=1}^{K} w_k}.$$

With AI assistance, these weights can be dynamically optimized in real time based on a student's background and career objectives, rendering the evaluation results more personalized. Furthermore, learning trajectory optimization objectives—such as maximizing overall competency gain—can be incorporated to guide the design of the assessment framework.

4.3 Teacher - AI Collaboration

In a smart education ecosystem, teachers and AI should operate in a complementary and collaborative manner. At the micro level, AI systems are responsible for real-time monitoring and support: they can analyze student responses, emotional indicators, and other data to provide timely personalized hints, learning recommendations, and feedback[20, 21]. For example, AI can generate targeted prompts or supplementary exercises based on students' current error patterns, making the learning process more adaptive. Additionally, AI can employ predictive models to identify at-risk or disengaged students early, issuing alerts to teachers to enable timely intervention.

At the macro level, teachers assume an irreplaceable role: they provide emotional support, value guidance, and cultivate learning motivation. As research has highlighted, AI reduces administrative burdens on teachers, allowing them to focus more on fostering students' creativity and higher-order thinking skills. Teachers guide students' critical thinking and humanistic development through encouraging discussions, collaborative learning, and shaping classroom culture. This human–AI complementarity implies that AI handles routine instructional adjustments automatically, freeing teachers to attend to students' psychological well-being and the meaningfulness of learning.

- Micro-level intervention (AI): AI systems perform real-time diagnostics based on learner models, offering hints or pushing relevant content when learning bottlenecks are detected. They also provide immediate feedback on correctness after task completion, analyze error patterns, and guide students in timely self-correction.

- Macro-level intervention (Teachers): Teachers focus on students' overall growth and value development, providing personalized emotional support and learning guidance. For instance, they can organize subject discussions or workshops on study strategies to help students cultivate a positive learning mindset and long-term goals[21].

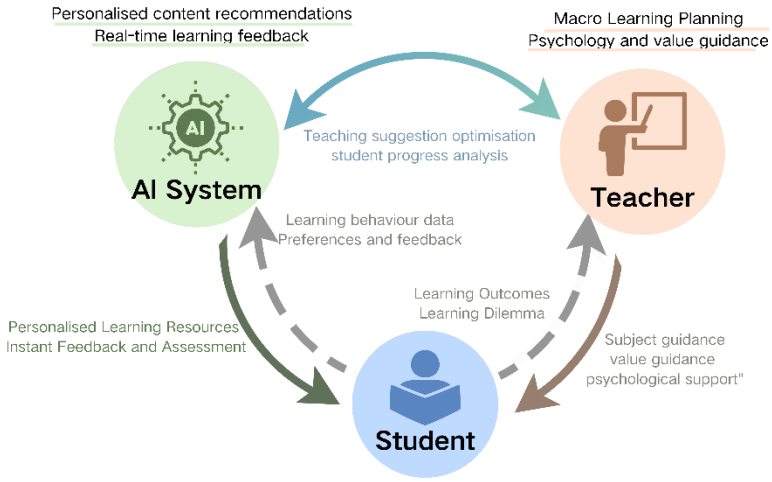


Fig. 2. Conceptual Diagram of Teacher–AI Collaborative Learning Support Structure

Through these mechanisms, AI and teachers form a complementary "educational team," enhancing classroom efficiency while addressing students' individualized needs and humanistic care. Ultimately, the integration of AI-driven dynamic evaluation with teacher guidance establishes a more flexible, equitable, and student-centered university education model.

5 Future Outlook: Becoming Unique Individuals in the Era of Superintelligence

As discussed, traditional education has long adopted a top-down approach—teachers functioned like gardeners "watering and nurturing" students, and course content was pre-defined. This resembles supervised learning in the field of artificial intelligence. Building on this analogy, we can interpret supervised fine-tuning (SFT) and reinforcement learning (RL) in educational terms: SFT resembles providing students with abundant exercises and answers, enabling them to accumulate knowledge through imitation, whereas RL encourages students to explore, experiment, and discover patterns independently, thereby enhancing their generalization capabilities.

To alleviate GPA pressures, assessment systems should leverage AI and big data for longitudinal and multidimensional evaluation, emphasizing innovation and holistic competence over mere grades, as demonstrated by initiatives at Peking University and Fudan University.

Personalized education and human–AI collaboration can enhance learning, with AI providing individualized guidance and teachers nurturing humanistic qualities. Ethically, AI should assist rather than dominate, ensuring technology supports curiosity, creativity, and human development.

The future of education envisions lifelong, flexible, and personalized learning, where AI supports holistic growth and celebrates creativity, critical thinking, and individual uniqueness, enabling each learner to thrive in a diverse society.

6 References

1. Huan, M., Li, Y., Zheng, T., Xu, X., Kim, S., Du, M., Poovendran, R., Neubig, G., Yue, X.: Does Math Reasoning Improve General LLM Capabilities? Understanding Transferability of LLM Reasoning, http://arxiv.org/abs/2507.00432, (2025). https://doi.org/10.48550/arXiv.2507.00432.
2. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., Arx, S. von, Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J.Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D.E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P.W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X.L., Li, X., Ma, T., Malik, A., Manning, C.D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J.C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J.S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A.W., Tramèr, F., Wang, R.E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S.M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., Liang, P.: On the Opportunities and Risks of Foundation Models, http://arxiv.org/abs/2108.07258, (2022). https://doi.org/10.48550/arXiv.2108.07258.
3. Language models and psychological sciences - PubMed, https://pubmed.ncbi.nlm.nih.gov/37941751/, last accessed 2025/08/15.
4. Zeithamova, D., Mack, M.L., Braunlich, K., Davis, T., Seger, C.A., Kesteren, M.T. van, Wutz, A.: Brain Mechanisms of Concept Learning. The Journal of Neuroscience. 39, 8259 (2019). https://doi.org/10.1523/JNEUROSCI.1166-19.2019.
5. Complementary task representations in hippocampus and prefrontal cortex for generalizing the structure of problems | Nature Neuroscience, https://www.nature.com/articles/s41593-022-01149-8, last accessed 2025/08/15.
6. Xu, Q., Peng, Y., Nastase, S.A., Chodorow, M., Wu, M., Li, P.: Large language models without grounding recover non-sensorimotor but not sensorimotor features of human concepts. Nat Hum Behav. 1–16 (2025). https://doi.org/10.1038/s41562-025-02203-8.

7. Wei, J., Bosma, M., Zhao, V.Y., others: Finetuned Language Models Are Zero-Shot Learners. In: Proceedings of the 10th International Conference on Learning Representations, ICLR 2022 (2022).
8. Karpicke, J.D., Butler: A.C.: Rote rehearsal, elaboration, and retrieval practice in verbal learning. Journal of Experimental Psychology: Learning, Memory, and Cognition. 38, 812–818 (2012).
9. Gao, L., Schulman, J., Hilton, J.: Scaling Laws for Reward Model Overoptimization. In: Proceedings of the 40th International Conference on Machine Learning (2023).
10. Wang, Z., Clarkson, M.R., Ganti, T.: The surprising effect of reward composition on RLHF. In: The Twelfth International Conference on Learning Representations, ICLR 2024 (2024).
11. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback, http://arxiv.org/abs/2203.02155, (2022). https://doi.org/10.48550/arXiv.2203.02155.
12. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: LLaMA: Open and Efficient Foundation Language Models, http://arxiv.org/abs/2302.13971, (2023). https://doi.org/10.48550/arXiv.2302.13971.
13. Rafailov, R., Sharma, A., Mitchell, E.: Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In: Advances in Neural Information Processing Systems 36 (NeurIPS 2023 (2023).
14. Hulleman, C.S., Harackiewicz, J.M.: Promoting interest and performance in high school science classes. Science. 326, 1410–1412 (2009).
15. Bai, Y., Jones, A., Ndousse, K.: Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback, (2022).
16. Bransford, J.D., Brown, A.L., Cocking, R.R.: eds.): How People Learn: Brain, Mind, Experience, and School. National Academy Press (2000).
17. Ryan, R.M., Deci, E.L.: Self-determination theory: Basic psychological needs in motivation, development, and wellness. Guilford Press (2017).
18. Pelánek, R.: Bayesian knowledge tracing, logistic models, and beyond: an overview of learner modeling techniques. User Model User-Adap Inter. 27, 313–350 (2017). https://doi.org/10.1007/s11257-017-9193-2.
19. Zhao, Z.: Optimization of Personalized Learning Paths in Educational AI Driven by Student Behavior Data. Journal of Industrial Engineering and Applied Science. 3, 16–25 (2025). https://doi.org/10.70393/6a69656173.323738.
20. Lin, C.-C., Huang, A.Y.Q., Lu, O.H.T.: Artificial intelligence in intelligent tutoring systems toward sustainable education: a systematic review. Smart Learning Environments. 10, 41 (2023). https://doi.org/10.1186/s40561-023-00260-y.
21. Joseph, S.: Rethinking assessment: how AI is changing the way we measure student success? AI & Soc. (2025). https://doi.org/10.1007/s00146-025-02255-4.
22. 大学生不考试啦，那未来怎么评价？_凤凰网, https://news.ifeng.com/c/8lbTlrEiElJ, last accessed 2025/08/14.

23. Oudkerk Pool, A., Jaarsma, A.D.C., Driessen, E.W., Govaerts, M.J.B.: Student perspectives on competency-based portfolios: Does a portfolio reflect their competence development? Perspect Med Educ. 9, 166–172 (2020). https://doi.org/10.1007/s40037-020-00571-7.

Quantitative Optimization of Gamified Museum Experiences: A Motion in Mind Approach

Jiayu LIU, Jiahao ZHANG, and Yuexian GAO

Hebei University of Engineering, Handan, China
ljy10070026@163.com, zjh975640888@gmail.com, gaoyuexian@hebeu.edu.cn

Abstract. Museums are increasingly adopting immersive, interactive, and gamified strategies to enhance visitor engagement and learning. This paper examines how *Jubensha* - a Chinese narrative-driven live-action role-playing (LARP) format - can inform the design and evaluation of gamified museum experiences. While such initiatives have proliferated, most rely on qualitative iteration rather than systematic metrics, resulting in inconsistent engagement and limited comparability across contexts. To address this gap, we introduce the *Motion in Mind* framework, which models cognitive-emotional dynamics through analogies to physical parameters such as mass, velocity, and potential energy. We integrate this theoretical model with empirical data from Chinese gamified museum cases, including the *Sea Monster Exhibit*, *Canal Mystery*, and *Jiangmen Script Tour*. Our results demonstrate how quantitative indicators can identify optimal engagement ranges and guide iterative improvements. The study contributes a reproducible, data-driven approach for analyzing and optimizing gamified cultural experiences.

Keywords: Gamified Museum Experience · Jubensha · Motion in Mind · Engagement Optimization · Cultural Interaction Design

1 Introduction

In recent years, museums have shifted from an *object-centered* to a *visitor-centered* paradigm. Immersive, interactive, and gamified experiences have become key strategies for attracting diverse audiences and achieving educational objectives [20,14,4]. This transition is supported by the adoption of digital technologies—such as location-based mobile games [17], virtual reality installations [14], and interactive storytelling platforms [4,17]—that enable new forms of participatory cultural engagement. Museum learning is increasingly recognized as a process that fosters curiosity, social interaction, and emotional connection, rather than the simple transfer of factual knowledge [3,5].

Within this context, *Jubensha* (also known as *script murder games*), a culturally distinctive form of live-action role-playing (LARP) rooted in Chinese narrative traditions, is being adopted through such activities by an increasing number of museums in China, which aim to attract more visitors and help them

better understand the culture embodied by their collections. It combines storytelling, character immersion, and social interaction within a structured, time-limited format, integrating Chinese cultural and historical elements to promote learning through role-play and collaborative problem solving [19,16,6]. However, despite the growing popularity of gamified museum practices, their design and evaluation often rely on qualitative intuition and ad hoc iteration, leading to inconsistent visitor experiences, limited cross-case comparability, and potential issues such as superficial engagement or cognitive overload.

Previous research emphasizes the importance of balancing accessibility and challenge [4,17], addressing diverse audience profiles [8,2], and embedding interactive content within coherent narrative or spatial frameworks [17,1]. However, the lack of clear and reproducible evaluation metrics continues to constrain systematic optimization of engagement, learning quality, and visitor well-being.

To address these challenges, this paper introduces the *Motion in Mind* framework, which models cognitive - emotional dynamics through analogies to physical concepts such as mass, velocity, momentum, and potential energy. This framework enables the derivation of interpretable parameters - including success rate, challenge magnitude, and potential emotional energy - for evaluating interactive cultural systems in a quantitative and reproducible manner.

We apply this integrated approach to a set of Chinese gamified museum cases, including the *Sea Monster Exhibit*, *Canal Mystery*, and *Jiangmen Script Tour*. By analyzing task-level performance data, we identify optimal engagement ranges and propose data-driven strategies for iterative design refinement. This study thus contributes a reproducible, theoretically grounded methodology for the design, comparison, and enhancement of gamified museum experiences.

2 Case Studies

To demonstrate the applicability of the Motion in Mind model in gamified museum contexts, we selected three representative projects: *Jiangmen Script Tour*, *Canal Mystery Adventure*, and *Sea Monster Exhibit*, as shown in Figure 1. Each case is briefly introduced below, with emphasis on its design features, current challenges, and task structure. This contextualization ensures that the subsequent quantitative analysis is aligned with the museums' goals and realities.

2.1 Jiangmen Script Tour (Board-like Narrative)

The Jiangmen Museum implemented a "script-tour" format, combining role-playing with historical storytelling. Visitors assume specific roles within a branching narrative, making decisions that unlock different storylines. **Challenges**: While the immersive narrative enhances emotional resonance, the high entry difficulty limits accessibility for casual visitors. Balancing depth of story with broad participation remains a central issue. **Tasks**: The volunteers were required to complete a series of branching story tasks. The criterion for task success was to be able to narrate the story's conclusion without quitting halfway and without any external assistance.

Fig. 1: Overview of the three study cases: (top-left) *Jiangmen Script Tour* (Jubensha-style role-play), (top-right) *Sea Monster Exhibit* (scoring-oriented digital interactives), (bottom) *Canal Mystery Adventure* (hybrid puzzle-education).

2.2 Canal Mystery Adventure (Hybrid Puzzle-Education)

This youth-oriented exhibition integrates puzzle-solving with historical education about canal engineering. Participants navigate through themed spaces, solving sequential riddles while collecting historical knowledge. **Challenges**: The exhibition succeeds in delivering knowledge, but some puzzles are overly complex, creating uneven progression across age groups. Ensuring progressive difficulty remains a design concern. **Tasks**: Volunteers completed a sequence of three hybrid tasks: (1) board-style navigation of a puzzle route, (2) historical riddle-solving with partial scoring, and (3) final synthesis of clues. Success was defined as completing the majority of checkpoints.

2.3 Sea Monster Exhibit (Scoring-oriented Interactive)

This digital exhibit engages participants in interactive mini-games related to marine reptile evolution. Visitors earn scores by playing AR interactive projection games, motion-sensing throwing games, and educational jigsaw puzzles. **Challenges**: The exhibit is highly attractive to young visitors but sometimes emphasizes entertainment over in-depth educational outcomes. The challenge is to preserve engagement while deepening knowledge retention. **Tasks**: Volunteers engaged in multiple short scoring tasks (e.g., interactive matching games). Success was measured as achieving percentage of the maximum possible score.

3 Methodology

3.1 Motion in Mind Model

Game Refinement Theory, which quantifies engagement through measures of information acceleration in game progress [11,12,13], usually denoted as $GR = \sqrt{B}/D$ in board-like games and $GR = \sqrt{G}/T$ in sports like games, which contributes to the research of The Motion in Mind model [9,7,10] (**MiM**). The MiM can be adapted to quantify audience engagement in museum gamification by analyzing task performance data. In a success-driven context, the model uses three parameters:

Success rate v: Let S be the number of successful task completions (e.g., solving a puzzle, completing an AR segment) and N the total number of attempts. Then:

$$v = \frac{S}{N}$$

An "attempt" refers to any discrete opportunity to achieve a defined objective within the museum's interactive design.

Average challenge magnitude m: Rather than representing negative outcomes, m captures the *pacing* of the experience and the *density of rewarding moments*. Then:

$$m = 1 - v$$

When m is close to zero ($v \to 1$), successes occur almost continuously, leading to low emotional fluctuation and a flat experience rhythm. When m approaches one ($v \to 0$), failures dominate, causing long intervals between rewards and potential disengagement. Moderate m values (typically 0.4–0.6) generate alternating sequences of challenges and rewards, creating an emotionally engaging rhythm that sustains attention over time.

Emotional potential E_p: In the success-driven formulation:

$$E_p = 2 \times m \times v^2$$

E_p captures the potential emotional intensity from the combination of challenge magnitude and success probability; higher values indicate stronger engagement potential.

Remarks:

- v reflects accessibility and the balance between difficulty and attainability.
- m captures difficulty and the density of rewarding moments.
- E_p links these core parameters to an interpretable measure of emotional engagement potential.

In summary, v measures accessibility, m reflects pacing, and E_p links both to an interpretable indicator of emotional engagement. These parameters will be revisited in the *Results and Discussion* to show how data-informed adjustments can enhance emotional immersion and sustained participation.

3.2 Measures in Gamified Museum Contexts

In the *Motion in Mind* model, the probability of success in a single attempt is denoted by v, and failure by $m = 1 - v$. To apply these measures in gamified museum contexts, tasks are categorized as *board-like*, *scoring-like*, or *hybrid*. Board-like tasks involve sequential levels or nodes cleared in order (e.g., puzzle routes). Scoring-like tasks consist of independent scoring attempts (e.g., mini-games, shooting challenges, quizzes). Hybrid tasks combine both structures, requiring separate computation of success rates and subsequent weighted integration.

The computation of v follows the *Motion in Mind* framework, representing the probability of success per attempt. In this study, task-specific adaptations were informed by prior applications in sports and game design research [11,18].

For board-like structures (e.g., scripted tours), v was approximated as the ratio of successfully completed bottleneck stages (B_{eff}) to total decision points (D). For scoring-like structures (e.g., digital interactive tasks), v was calculated as the mean participant score relative to the maximum score. In hybrid cases, v was derived through a weighted combination of the two measures.

For board-like tasks, the success rate v_{board} is calculated as

$$v_{\text{board}} = \frac{B_{\text{eff}}}{D}, \tag{1}$$

where D is the total number of levels, and B_{eff} represents the *effective number of bottleneck stages* encountered and solved by an average participant, rather than the literal count of mandatory decision points. A larger number of bottlenecks generally reduces the success rate. In most museum-based gamified experiences, designers deliberately limit the number of decisive checkpoints to a little proportion, which reflects an intentional design strategy: to provide moments of challenge without creating excessive barriers that could discourage casual visitors [15].

For scoring-like tasks, the success rate v_{score} is determined by the ratio of the average participant score to the maximum possible score:

$$v_{\text{score}} = \frac{\text{Average Score}}{\text{Max Score}}. \tag{2}$$

For hybrid tasks, the overall success rate is a weighted combination of the two components:

$$v = \alpha_{\text{board}} \cdot v_{\text{board}} + (1 - \alpha_{\text{board}}) \cdot v_{\text{score}}, \tag{3}$$

where α_{board} is a weighting factor between 0 and 1, determined by the relative importance of the board-like component in the game design. Once v is obtained, the failure rate is given by

$$m = 1 - v. \tag{4}$$

This framework can be applied to specific case studies. In the *Canal Mystery* (hybrid type), v_{board} is calculated from B_{eff} and D, while v_{score} is derived from average score proportions; applying a weighting factor (based on the amount of

time audiences spend on different types of puzzles or tasks) yields the overall v and m. In the *Jiangmen Script Tour* (board-like dominant), the total number of narrative branches and bottlenecks is identified to compute v_{board}, which serves directly as v. In the *Sea Monster Exhibit* (scoring-like dominant), participant score data from interactive activities is aggregated to determine v_{score}, which is then used to derive m.

3.3 Data Collection Procedures

We recorded volunteer interactions on site for the three cases (20 participants for *Canal Mystery*, 14 for *Jiangmen Script Tour*, and 17 for *Sea Monster Exhibit*). For each discrete task opportunity, we recorded task-level attempts (N) and successes (S); for scoring-oriented activities we logged raw scores and normalized them by the maximum possible score to obtain v_{score}. For board-like sequences we identified decision points (D) and estimated the effective bottlenecks B_{eff} as those checkpoints at which an average visitor's progress meaningfully stalled (based on time-to-clear and assistance requests). Hybrid cases combined both via Eq. (2), with weights derived from the proportion of time spent in each task type.

3.4 Analysis Pipeline

After computing case-level v, we derived $m = 1 - v$ and the success-driven potential $E_p = 2(1 - v)v^2$. We then visualized E_p across v to diagnose whether each experience lies near the theoretical optimum ($v \approx 0.67$) and formulated design adjustments (e.g., progressive hints, adaptive tiers) based on the direction of deviation.

4 Results and Discussion

This research recorded the participants' performance in the process, mainly focusing on the efficiency of solving the puzzles or tasks. At the end, their subjective satisfaction level was recorded.

The summarized results are presented in Table 1 and visualized in Figure 2. Figure 2 illustrates the relationship between the success probability v and potential emotional energy E_p. The theoretical optimum occurs at $v \approx 0.67$, which maximizes E_p and represents an ideal balance between challenge and accessibility, as indicated by the red marker. Case-specific markers reveal deviations from this optimum: the *Sea Monster Exhibit* ($v = 0.7154$, $E_p = 0.2913$) is near the peak, suggesting a well-calibrated challenge level; the *Canal Mystery* ($v = 0.4221$, $E_p = 0.2059$) underperforms due to a relatively low success rate; and the *Jiangmen Script Tour* ($v = 0.2653$, $E_p = 0.1034$) shows the largest gap, indicating a high difficulty that may limit sustained engagement.

Examining the three cases in detail, the *Sea Monster Exhibit* demonstrates a slightly higher-than-optimal success rate, which can lead to a gradual reduction

Table 1: Calculated Motion in Mind parameters for the three museum gamification cases.

Case	D	B_{eff}	v_{board}	α	Total Score	Average Score	v_{score}	v	E_p	Satisfaction
CM	8	1.4984	0.1873	0.6	10	7.743	0.7743	0.4221	0.2059	85%
JST	12	3.1836	0.2653	–	–	–	–	0.2653	0.1034	85.71%
SME	–	–	–	–	8	5.7232	0.7154	0.7154	0.2913	94.11%

CM, JST, SME are relatively refer to Canal Mystery, Jiangmen Script Tour and Sea Monster Exhibit

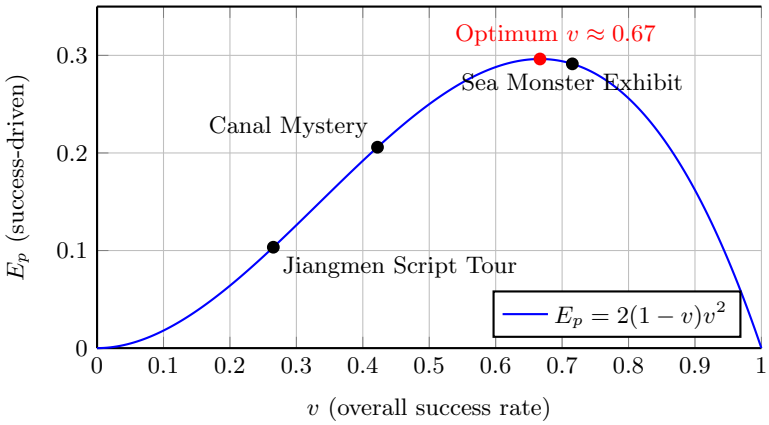


Fig. 2: E_p versus overall success rate v under the MiM formulation; the red marker denotes the theoretical optimum near $v \approx 0.67$.

in emotional tension p during extended interactions. Introducing adaptive challenge tiers, such as optional "hard mode" missions, could help maintain engagement. The *Canal Mystery* shows a moderately low v, suggesting that its challenge level may be discouraging for average participants. Simplifying task mechanics or providing progressive hints could increase v toward the optimal range of 0.6–0.7, enhancing potential emotional energy E_p. The *Jiangmen Script Tour* exhibits both low v and E_p, indicating limited sustained engagement. Adjusting the difficulty curve—beginning with easier tasks and offering more frequent positive reinforcement—could improve emotional immersion and encourage longer participation.

Overall, designs with v values between 0.6 and 0.7 tend to achieve higher E_p and richer experiential outcomes, providing a quantitative target for iterative adjustments in gamified museum design. The satisfaction data in Table 1 support these findings, confirming that aligning task difficulty with the theoretical optimum maximizes audience engagement while preserving the depth of cultural content delivery.

The Sea Monster Exhibit demonstrates a slightly higher-than-optimal success rate v, which can lead to a gradual reduction in emotional tension p over extended interaction. Introducing adaptive challenge tiers, such as optional "hard mode" missions, could preserve engagement dynamics. The Canal Mystery case shows a moderately low v, suggesting that the challenge level might be discouraging for average participants. Simplifying task mechanics or implementing progressive hints could raise v toward the optimal range of 0.6 - 0.7, thereby increasing potential emotional energy E_p. The Jiangmen Script Tour exhibits both low v and E_p, indicating insufficient sustained engagement. Adjusting the difficulty curve - starting with easier tasks and providing more frequent positive reinforcement - could enhance emotional immersion and encourage longer participation. Collectively, these adjustments align each case more closely with the theoretical optimum, maximizing audience engagement while preserving the depth of cultural content delivery.

5 Conclusion and Future Work

This study applied the Motion in Mind framework to quantitatively evaluate three gamified museum experiences—the *Sea Monster Exhibit*, *Canal Mystery*, and *Jiangmen Script Tour*. Using key parameters such as overall success rate v, emotional tension p, and potential emotional energy E_p, we assessed how each design balances challenge and accessibility. Results indicate that experiences with v between 0.6 and 0.7 consistently achieve higher E_p values, aligning with the theoretical optimum and fostering richer engagement. The *Sea Monster Exhibit* approaches this ideal, while the other two cases could benefit from calibrated adjustments to task difficulty and accessibility.

Beyond case-specific insights, the Motion in Mind framework provides a reproducible, data-driven approach for iterative optimization in cultural and educational contexts. By setting measurable targets for v and E_p, designers can fine-tune emotional pacing to sustain both cognitive and affective engagement.

This study has several limitations. First, the operationalization of v relies on case-specific assumptions, such as the definition of B_{eff} and the use of small, uncontrolled participant groups. Although uncertainty analysis mitigates some variability, reproducibility and cross-case comparability remain constrained. Future research should employ standardized data collection protocols—e.g., task logs, eye-tracking, and physiological measures—with larger, more representative samples. Second, the model emphasizes performance-based engagement measures. While this provides a quantifiable lens on emotional tension and sustained participation, museum experiences also pursue broader goals, including learning outcomes, exploration, and creative interaction. The Motion in Mind framework should therefore be considered one component of a multi-dimensional evaluation toolkit rather than a comprehensive measure of success. Third, the cases analyzed were drawn from relatively successful and well-documented projects, introducing a survivorship bias. Consequently, the identified optimal engagement range ($v \approx 0.6$ - 0.7) may reflect the characteristics of this sample rather than a

universal principle. Future studies should include less successful or experimental cases to verify the generalizability of these findings.

Future work will integrate real-world behavioral and physiological data to validate and refine the model's predictive capabilities, enabling more precise alignment between gamified museum experiences and evolving audience expectations.

References

1. Anderson, S.L.: The interactive museum: Video games as history lessons through lore and affective design. E-Learning and Digital Media (2019)
2. Antoniou, A.: Predicting cognitive profiles from a mini quiz: a facebook game for cultural heritage. In: Games and Learning Alliance Conference (2019)
3. Apostolellis, P., Bowman, D.A.: Small group learning with games in museums: Effects of interactivity as mediated by cultural differences. Proceedings of IDC (2015)
4. Bellotti, F., Berta, R., De Gloria, A., D'Ursi, A., Fiore, V.: A serious game model for cultural heritage. ACM Journal on Computing and Cultural Heritage (2012)
5. Bohlmeijer, M.: Systematic literature review on interaction design used for museum learning. TScIT Conference (2024)
6. Bowman, S.L., Diakolambrianou, E., Brind, S.: Transformative Role-playing Game Design. Uppsala University Publications (2025)
7. Gao, Y., Liu, C., Gao, N., Khalid, M.N.A., Iida, H.: Nature of arcade games. Entertainment Computing **41**, 100469 (2022)
8. Hsu, T.Y., Liang, H.Y.: Museum engagement visits with a universal game-based blended museum learning service for different age groups. Library Hi Tech (2021)
9. Iida, H., Takahashi, N.: Game Refinement Theory and Its Application. Springer Series on Cultural Computing, Springer (2016)
10. Iida, H., Khalid, M.N.A.: Using games to study law of motions in mind. IEEE Access **8**, 138701–138709 (2020)
11. Iida, H., Takeshita, N.: An application of game refinement theory to sports games. In: IEEE Conference on Computational Intelligence and Games (2008)
12. Iida, H., et al.: Game refinement theory for sports games. Entertainment Computing (2014)
13. Iida, H., et al.: Game refinement theory and its applications. In: Advances in Computer Games (2016)
14. Lepouras, G., Vassilakis, C.: Virtual museums for all: employing game technology for edutainment. In: Virtual Reality (2004)
15. Li, Q.: Implementation mechanism and prospects of the "museum + scripted game" model in museum education: A case study of jiangmen museum. Bowuguan Yanjiu (Museum Research) (4), 112–120 (2022)
16. Liu, Y.: Experiencing china's intangible cultural heritage in role-playing games: Comparative studies between mmorpgs and larps. International Journal of Role-Playing (14), 41–46 (2023)
17. Rubino, I., et al.: Integrating a location-based mobile game in the museum visit. ACM Journal on Computing and Cultural Heritage (2015)
18. Siqi, L., Khalid, M.N.A., Iida, H.: Enhancing auction experiences: Game dynamics and customer experience design. Entertainment Computing **54**, 100959 (2025)

19. Xiong, S., Wen, R., Zheng, H.: Player category research on murder mystery games (2023)
20. Ćosović, M., Ramić Brkić, B.: Game-based learning in museums—cultural heritage applications. Information **11**(1), 22 (2020)

Spoof Detection in Automatic Speaker Verification Using ResNet-34 and Early-Stage Cepstral Coefficient Fusion

Kosin Kalarat[1], Sasiporn Usanavasin[1], Thanaruk Theeramunkong[1], Kasorn Galajit[2], and Jessada Karnjana[2]

[1] Sirindhorn International Institute of Technology, Thammasat University, Pathum Thani, 12120, Thailand
6622300090@g.siit.tu.ac.th.
[2] NECTEC, National Science and Technology Development Agency, Pathum Thani, 12120, Thailand

Abstract. Spoofing attacks pose a significant challenge to the reliability of automatic speaker verification (ASV) systems, particularly in real-world applications where adversaries can generate highly convincing synthetic or replayed speech. To address this, a binary classification approach is investigated for spoof detection, focusing on feature extraction methods and deep learning architectures optimized for performance on the ASVspoof 2019 Physical Access (PA) dataset. In this study, mel-frequency cepstral coefficients (MFCC), linear-frequency cepstral coefficients (LFCC), and inverse MFCC (IMFCC) are employed to generate time–frequency representations, which are then processed by a ResNet-34 model. These three representations are mapped into a three-channel input, enabling the model to capture spectral variations relevant to spoof detection. The model is trained with binary cross-entropy loss, class weighting, and regularization techniques to mitigate data imbalance and overfitting. Our system achieves an equal error rate (EER) of 3.68% and an F1-score of 0.97 on the evaluation set, outperforming baseline systems by 33.3% in EER reduction. These results suggest that integrating multiple cepstral feature types can effectively enhance spoof detection performance in binary classification settings. This approach provides a scalable framework for real-world ASV security applications, particularly where computational efficiency and robustness are critical.

Keywords: spoof detection, binary classification, speaker verification, MFCC, LFCC, IMFCC, ResNet-34, ASVspoof.

1 Introduction

Automatic speaker verification (ASV) systems have become integral to applications such as banking authentication, virtual assistants, and secure access control. However, their increasing adoption has been accompanied by heightened vulnerability to spoofing attacks, including replay, voice conversion, and speech synthesis [1]. These attacks

can severely undermine system reliability, leading to potential security breaches and identity fraud.

To address these vulnerabilities, the Automatic Speaker Verification Spoofing and Countermeasures (ASVspoof) challenges have provided standardized datasets, including Logical Access (LA) and Physical Access (PA), to benchmark countermeasure systems. The LA dataset comprises voice conversion, speech synthesis, and deepfake attacks, while the PA dataset focuses on replay attacks [2]. Our study focuses on replay attacks from the PA dataset. These attacks are a growing concern because of their simplicity: anyone can use a smartphone to record and replay a target's voice without specialized knowledge, making them a significant threat to voice authentication.

ASVspoof 2019 challenge is the latest updated version of replay attack which is PA sub-dataset. The challenge provided the baseline method using constant-Q cepstral coefficients (CQCC) [3] and linear-frequency cepstral coefficients (LFCC) [4] with the same classifier which is Gaussian mixture model (GMM). While deep learning methods like convolutional neural networks (CNN) and residual networks (ResNet) have improved spoof detection by extracting patterns from speech's time-frequency representations, many existing studies still use only one type of acoustic feature (e.g., mel-spectrograms [5], MFCC [6], gammatone cepstral coefficients (GTCC) [7], or LFCC [8]). This single-feature approach may limit the model's ability to learn rich, discriminative representations [14].

This study focuses on leveraging multiple filterbank-based features MFCC, LFCC, and inverse mel-frequency cepstral coefficients (IMFCC) to enhance spoof detection performance in a binary classification setting. These features are independently extracted and mapped to different channels, allowing the ResNet-34 architecture to learn complementary spectral information. Unlike previous approaches, this method avoids overly complex fusion strategies while maintaining computational efficiency.

While the overall extraction process for MFCC, LFCC, and IMFCC is similar, they differ in frequency scaling strategies. MFCC applies to a nonlinear mel scale after computing the power spectrum from the Fourier transform, emphasizing lower frequencies in line with the human ear’s higher sensitivity in this range [9]. This makes MFCC particularly effective for speech and speaker recognition tasks [9]. LFCC, on the other hand, uses a linear frequency scale after computing the magnitude spectrum, maintaining constant spectral resolution across the frequency range-an advantage when fine-grained detail at higher frequencies is important [10]. IMFCC employs an inverse mel scale, placing greater emphasis on higher frequencies [11], thereby capturing information that MFCC and LFCC may overlook, especially in certain speech types

By fusing these three complementary feature representations, the resulting composite feature set captures the spectral envelope across low, mid, and high frequencies more comprehensively. Although this process slightly increases computational dimensionality, it significantly improves the descriptive power of the input representation for modeling the complex acoustic patterns of both bona fide and spoofed speech. Furthermore, the proposed fusion method is straightforward to implement, computationally efficient, and exhibits strong robustness in diverse spoofing detection scenarios.

The rest of this paper is organized as follows. Section 2 briefly provides background knowledge of this research. Section 3 presents our proposed method. Section 4

describes the experimental results. Section 5 presents the discussion of this research. Finally, the conclusion of this research will be presented in Section 6.

2 Background

This section provides an essential overview of the key components that form the foundation of our methodology. The fundamental concepts in speech signal processing will be introduced, including feature extraction techniques such as MFCC, LFCC, and IMFCC, as well as the ResNet-34 architecture used for our classification task.

2.1 Feature Extraction

The input speech is represented as a tensor comprising three spectral feature types: MFCC, LFCC, and IMFCC. These coefficients are widely used in speech processing for their ability to capture robust, discriminative characteristics across different frequency bands [9][10][11].

To exploit the complementary strengths of these feature types, MFCC, LFCC, and IMFCC are employed independently during model training shown in Fig. 1. This approach allows the model to benefit from the distinct frequency-domain perspectives offered by each scaling technique.

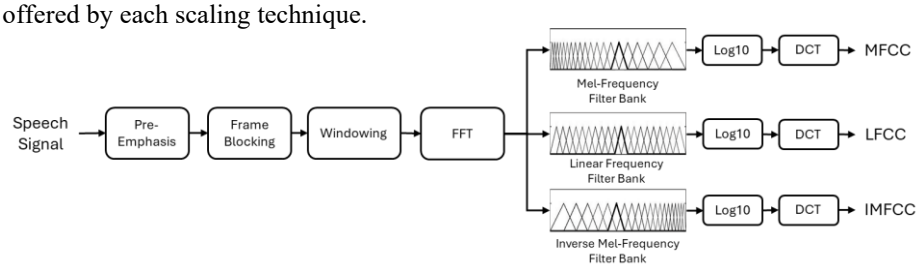


Fig. 1. Feature extraction of MFCC, LFCC and IMFCC block diagrams.

2.1.1 Mel-Frequency Cepstral Coefficients (MFCC)

The MFCC is widely recognized and has been effectively applied in numerous speech recognition tasks [9]. The extraction process begins by segmenting the speech signal into equal-sized frames. Each frame is then transformed using the fast Fourier transform (FFT) to obtain its spectrum, which is subsequently filtered through a mel-scaled filter bank composed of triangular bandpass filters. A logarithmic scale is applied to the filter bank outputs, and finally, the cepstral coefficients are derived using the discrete cosine transform (DCT).

$$\mathrm{MFCC(q)} = \sum_{m=1}^{M} \log[\mathrm{MF(m)}] \cos\left[\frac{q\left(m-\frac{1}{2}\right)\pi}{M}\right], \tag{1}$$

and

$$\mathrm{MF(m)} = \sum_{k=1}^{K} |X^{DFT}(k)|^2 H_m(k)\,, \tag{2}$$

where q denotes the cepstral coefficient index, while M refers to the total number of triangular filters in the mel filter bank. The term $\mathrm{MF}(m)$ corresponds to the logarithmic mel-scale filter bank energy for the m-th filter. Meanwhile, $X^{DFT}(k)$ represents DFT of the speech signal at frequency bin k, and $H_m(k)$ describes the frequency response of the m-th mel triangular band-pass filter. Finally, K specifies the total number of DFT frequency bins.

2.1.2 Linear Frequency Cepstral Coefficients (LFCC)

The LFCC are obtained through the same procedure as MFCC, except that a linear frequency filter bank replaces the mel filter bank [10]. In the linear frequency filter bank, the triangular filters maintain a constant bandwidth, unlike the mel filter bank where the band-pass varies across the frequency range. This characteristic allows LFCC to capture high-frequency artifacts effectively.

$$f(m) = \left(\frac{N}{F_s}\right)\left(f_1 + m\frac{f_h - f_1}{M + 1}\right), \tag{3}$$

where $f(m)$ represents the m-th boundary point of the linear filter bank, while N denotes the length of FFT corresponding to the number of points in the discrete Fourier transform. The parameter $F(s)$ is the sampling frequency of the input signal, f_1 is the lowest frequency of the filter bank, and $\mathrm{f_h}$ indicates the highest frequency considered. The total number of filters in the linear filter bank is represented by M, and m refers to the filter index.

2.1.3 Inverse Mel-Frequency Cepstral Coefficients (IMFCC)

The IMFCC are a variation of MFCC applied in voice spoofing detection [11]. They capture key characteristics of speech signals by focusing on high-frequency components through an inverse mel-scale filter bank composed of triangular bandpass filters. The extraction process follows the same steps as MFCC, but with the inverted filter bank.

$$\mathrm{IMFCC}(q) = \sum_{m=1}^{M} \log[IMF(m)] \cos\left[\frac{q\left(m - \frac{1}{2}\right)\pi}{M}\right], \tag{4}$$

and

$$\mathrm{IMF}(m) = \sum_{k=1}^{K} |X^{DFT}(k)|^2 H_m(k)\,, \tag{5}$$

where most variables are the same as MFCC except IMF(m) and $H_m(k)$. The term IMF(m) corresponds to the logarithmic inverse mel-scale filter bank energy for the m-th filter and $H_m(k)$ describes the frequency response of the m-th inverse mel triangular band-pass filter.

2.2 Residual Network with 34 Layers (ResNet-34)

ResNet-34 is a variant of the residual neural network (ResNet) architecture comprising 34 layers [12]. It overcomes the vanishing gradient problem by employing shortcut connections with identity mappings, enabling more effective training of deep networks [13]. Figure 2 illustrates the shortcut connection where *x* represents the input to the block, which could be the output of a previous layer or an earlier residual block. This input is passed through a series of transformations, denoted as *F(x)*, which typically consists of two weighted layers with a rectified linear unit (relu) activation function.

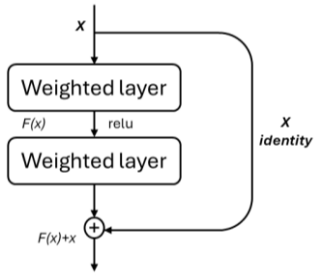


Fig. 2. Basic building blocks of ResNet.

3 Proposed Method

This research proposed the fusion of three features which are MFCC, LFCC and IMFCC using ResNet-34 as classifier. Frequency alignment on the extracted features is performed to ensure they correspond to the same frequency scale. Subsequently, standard normalization is applied to remove dimensional discrepancies between the different feature types. Finally, these normalized features are concatenated along the frequency axis to produce fused feature maps, which are then fed into the model.

Feature fusion refers to the integration of features derived from different sources or representations to enhance model performance. In deep learning, the objective of feature fusion is to exploit the complementary information inherent in diverse features, thereby increasing model expressivity and improving performance in classification, regression, or other downstream tasks. This approach has been widely recognized as effective for boosting the training efficiency of neural network models [14][15][16].

.

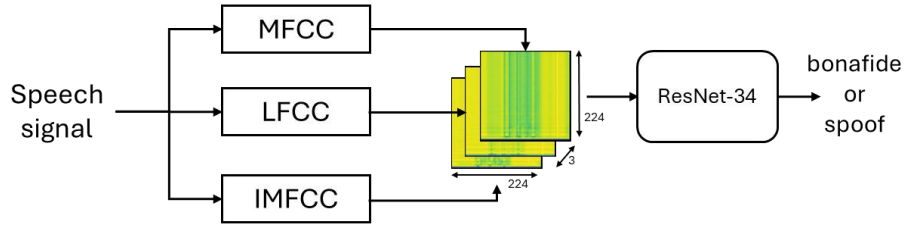


Fig. 3. Proposed framework.

While concatenating the three features along the frequency axis is straightforward, it disrupts the spatial structure of the spectrum within the feature. To preserve the local correlation in the time–frequency plane, the three feature spectrograms are stacked in a new dimension to form a tensor of size 3×224×224. Each audio file produces a 224×224 spectrogram after feature extraction, which is then stacked to form a three-channel format similar to RGB images, resulting in a size of 3×224×224, as illustrated in Fig. 3. This approach not only converts the traditional speech sequence into an image-like representation but also integrates multiple features into a format better suited for convolutional neural network processing.

4 Experimental Results

In this study, the widely used ASVspoof 2019 dataset was employed, which addresses three primary attack types: synthetic speech (SS), voice conversion (VC), and replay attacks [1]. It is divided into two subsets: logical access (LA), containing spoofed speech generated via text-to-speech (TTS) and voice conversion methods, and physical access (PA), comprising replayed speech recordings. Each subset is further split into training, development, and evaluation sets, as summarized in Table 1. Our work focuses exclusively on the replay attack as PA subset.

Table 1. Logical Access (LA) and Physical Access (PA) subsets in ASVspoof 2019 dataset.

Subset	ASVspoof 2019 LA		ASVspoof 2019 PA	
	Spoof	Bonafide	Spoof	Bonafide
Train	22,800	2,580	48,600	5,400
Dev	22,296	2,548	24,300	5,400
Eval	63,882	7,355	135,000	8,000

All three features MFCC, LFCC, and IMFCC were extracted from 4 seconds speech signal using the following configuration. A 30-ms Hamming window was applied with a 15-ms frame shift. The spectrum was computed using a 2048-point FFT, with the frequency range analyzed from 0 Hz to 8 kHz. A bank of 128 mel-scale filters was used to generate the filter bank energies, from which 60 cepstral coefficients were then derived. A pre-emphasis filter with a coefficient of 0.97 was applied to the signal before

feature extraction. The final feature set was normalized before being used for classification.

The classifier used in this study is based on ResNet-34 architecture. Training was conducted with a mix of genuine and spoofed speech samples from the ASVspoof 2019 training set. The approach described involves training our model entirely from scratch, without leveraging any pre-trained data. The model was optimized using the Adam optimizer with a learning rate of 0.0001 and trained for 75 epochs with batch size of 16, using sparse categorical cross-entropy as the loss function.

Performance was evaluated on the development and evaluation sets, which contain previously unseen genuine and spoofed samples. Metrics including equal error rate (EER), accuracy, precision, recall, and F1-score were used to assess the model's effectiveness in spoof detection. The EER is a standard metric in biometric security, defining the threshold where the probability of falsely accepting a non-matching sample equals the probability of falsely rejecting a matching one. Accuracy evaluates the system's ability to correctly classify genuine and spoofed speech by calculating the percentage of all samples that were classified correctly. The F1-score is the harmonic mean of precision and recall. Precision measures the fraction of correctly classified positive samples out of all samples predicted as positive, whereas recall measures the fraction of correctly classified positive samples out of all actual positive samples. This combined metric is especially useful for providing a balanced performance assessment in situations with imbalanced classes.

Table 2. EERs comparison.

Method		EER (%)
Feature Extraction	Classifiers	
LFCC	GMM	13.54 [3]
CQCC	GMM	11.04 [4]
Mel-Spectrogram	ResNet-34	5.74 [5]
GTCC	ResNet-34	11.03 [7]
MFCC	ResNet-34	4.35
LFCC	ResNet-34	5.02
IMFCC	ResNet-34	9.66
Proposed method	ResNet-34	**3.68**

Table 2 summarizes the EER performance of different feature extraction methods and classifiers for replay attack detection in automatic speaker verification shown in Table 2. The results include both the standalone feature approaches and the proposed fusion method.

5 Discussions

The experimental results demonstrate that the choice of acoustic features significantly influences the performance of replay attack detection systems in automatic speaker verification. As shown in Table 2, traditional GMM-based systems employing LFCC and CQCC features yield relatively high equal error rates (13.54% and 11.04%, respectively). In contrast, ResNet-34 based classifiers provide substantial performance improvements across different features. For instance, mel-spectrogram combined with ResNet-34 achieves an EER of 5.74%, while GTCC with ResNe-34 results in 11.03%.

Among the tested features, MFCC and LFCC demonstrated strong performance when employed individually with ResNet-34, achieving EERs of 4.35% and 5.02%, respectively. In contrast, IMFCC alone resulted in a notably higher EER of 9.66%, indicating that IMFCC is less effective in discriminating replay-specific artifacts compared to MFCC and LFCC. This outcome aligns with expectations, as IMFCC applies to a smoother filter bank that avoids capturing excessive detail in these noisy bands. If a system relies too heavily on high-frequency information, it might introduce environmental noise and replay artifacts, which increase ambiguity for discrimination. IMFCC mitigates this risk by not enforcing strict sensitivity to high frequencies, making it a valuable complementary feature when integrated with MFCC and LFCC. This explains why the fusion of all three features achieved superior performance compared to using them individually. Therefore, the integration of MFCC, LFCC, and IMFCC features yielded the best result, reducing the EER to 3.68%. This performance not only surpasses that of individual feature-based systems but also outperforms previously reported approaches, underscoring the complementary nature of different cepstral features and demonstrating the efficacy of feature-level fusion in replay spoofing detection.

Table 3. Classification performance evaluation of the classifier when taking a single feature and taking a combined feature.

Features	Accuracy	Precision	Recall	F1	EER
MFCC	0.95	0.98	0.96	0.97	4.35
LFCC	0.95	0.97	0.96	0.97	5.03
IMFCC	0.90	0.98	0.89	0.93	9.66
MFCC/LFCC/IMFCC	0.96	0.96	0.98	0.97	3.68

The classification metrics presented in Table 3 further reinforce these findings. The fusion of MFCC, LFCC, and IMFCC not only produces the best EER but also achieves the highest recall (0.98) and balanced F1-score (0.97), indicating a reliable ability to detect spoofed signals while maintaining strong precision. This suggests that the proposed multi-feature ResNet-34 system provides a more generalized representation of spoofing cues compared to single-feature systems.

6 Conclusions

This paper investigated the effectiveness of cepstral-based features for replay attack detection in automatic speaker verification systems. While traditional single-feature approaches such as LFCC and MFCC achieved competitive performance, the results indicate that IMFCC alone is less effective in this context. The fusion of MFCC, LFCC, and IMFCC, however, yielded the best overall performance with an EER of 3.68%, outperforming both individual feature models and several previously reported benchmarks.

The findings confirm that leveraging complementary information from multiple cepstral representations enhances the discriminative power of deep learning models such as ResNet-34 for spoofing detection. This supports the use of feature fusion strategies in future designs of automatic speaker verification countermeasure systems.

Acknowledgments. This work is supported by Royal Thailand Government Scholarship for Science and Technology. Additional support for this research was provided by IISI and SIIT COE, Thammasat University. The ASEAN IVO project "Spoof Detection for Automatic Speaker Verification" was involved in the production of the contents of this presentation and was financially supported by NICT.

References

1. Tan, C.B., Hijazi, M.H.A., Khamis, N., Nohuddin, P.N.E.B., Zainol, Z., Coenen, F., Gani, A.: A survey on presentation attack detection for automatic speaker verification systems: State-of-the-art, taxonomy, issues and future direction. Multimedia Tools and Applications 80(21-23), 32725–32762 (2021)
2. Villalba, J., Lleida, E.: Detecting replay attacks from far field recordings on speaker verification systems. In: Biometrics and ID Management: COST 2101 European Workshop, BioID 2011, pp. 274–285. Springer, Brandenburg (Havel), Germany (2011)
3. Tak, H., Patino, J., Nautsch, A., Evans, N., & Todisco, M.: Spoofing attack detection using the non-linear fusion of sub-band classifiers. arXiv preprint arXiv:2005.10393 (2020)
4. Todisco, M., Delgado, H., & Evans, N.: Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. Computer Speech & Language 45, 516-535 (2017)
5. Aravind, P. R., Nechiyil, U., & Paramparambath, N.: Audio spoofing verification using deep convolutional neural networks by transfer learning. arXiv preprint arXiv:2008.03464 (2020)
6. Duraibi, S., Alhamdani, W., & Sheldon, F.T.: Replay spoof attack detection using deep neural networks for classification. In: 2020 International Conference on Computational Science and Computational Intelligence (CSCI), pp. 170-174. IEEE (2020)
7. Chaiwongyen, A., Duangpummet, S., Karnjana, J., Kongprawechnon, W., & Unoki, M.: Replay attack detection in automatic speaker verification using gammatone cepstral coefficients and resnet-based model. Journal of Signal Processing 26(6), 171-175 (2022)
8. Mon, K. Z., Galajit, K., Mawalim, C. O., Karnjana, J., Isshiki, T., & Aimmanee, P.: Spoof detection using voice contribution on lfcc features and resnet-34. In: 2023 18th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), pp. 1-6. IEEE (2023)

9. Muda, L., Begam, M., & Elamvazuthi, I.: Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. arXiv preprint arXiv:1003.4083 (2010)
10. Chakroborty, S., & Saha, G.: Improved text-independent speaker identification using fused MFCC & IMFCC feature sets based on Gaussian filter. International Journal of Signal Processing 5(1), 11-19 (2009)
11. Zhou, X., Garcia-Romero, D., Duraiswami, R., Espy-Wilson, C., & Shamma, S.: Linear versus mel frequency cepstral coefficients for speaker recognition. In: 2011 IEEE workshop on automatic speech recognition & understanding, pp. 559-564. IEEE (2011)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)
13. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016, LNCS, vol. 9908, pp. 630–645. Springer, Cham (2016)
14. Zhang, J., Tu, G., Liu, S., & Cai, Z.: Audio anti-spoofing based on audio feature fusion. Algorithms 16(7), 317 (2023)
15. Xue, J., Zhou, H., Song, H., Wu, B., & Shi, L.: Cross-modal information fusion for voice spoofing detection. Speech Communication 147, 41-50 (2023)
16. Xue, J., & Zhou, H.: Physiological-physical feature fusion for automatic voice spoofing detection. Frontiers of Computer Science 17(2), 172318 (2023)

From Digital Transformation to Human–AI Symbiosis: The Evolving Role of Universities in Shaping Education and Society

Xiaokun Shi[1⋆] and Jizong Jia[1⋆⋆]

Hebei University of Engineering,
No. 19, Taiji Road, Congtai District, Handan City, Hebei Province, China

Abstract. The proliferation of artificial intelligence (AI) is compelling universities to evolve beyond basic Digital Transformation (DT). This paper proposes and empirically tests a three-stage evolutionary model (DT → HAIC → HAIS), framed by the Technology-Organization-People (TOP) model, to guide and assess this transition. Drawing on a mixed-methods study of fifteen universities across diverse global contexts, our findings reveal that progression towards Human-AI Symbiosis (HAIS) hinges on organizational readiness and human agency, not merely technological infrastructure. Two key patterns emerged: effective ethical governance is crucial for building stakeholder trust, and dynamic feedback loops between AI deployment and policy accelerate integration. This research provides practical guidelines for university leaders to navigate this complex transformation. It concludes by positioning universities as critical societal actors in shaping responsible AI paradigms and fostering inclusive innovation. **Keywords:** Digital Transformation; Human–AI Collaboration; Human–AI Symbiosis; Higher Education; AI Governance; Educational Technology

1 Introduction

The explosive growth of artificial intelligence (AI) technologies has fundamentally reshaped how we think about higher education. Universities—long regarded as knowledge creators and drivers of social change—now face a dual challenge that we find particularly compelling: they must modernize their own operations through AI adoption while simultaneously guiding society toward responsible technology use [8, 11]. Unlike previous waves of digital innovation, today's AI systems possess remarkable sophistication, independence, and learning capabilities that create unprecedented opportunities alongside significant risks—far exceeding what earlier technology rollouts could achieve.

We have observed this transformation happening in distinct phases that universities can identify and track. Early Digital Transformation (DT) efforts

⋆ adcsss7@qq.com
⋆⋆ 2394492900@qq.com

typically involved layering digital tools over existing teaching methods and administrative workflows—essentially digitizing what already existed rather than reimagining it. As AI capabilities matured, universities entered a Human–AI Collaboration (HAIC) stage, where AI systems began to act as partners in teaching, learning, and institutional decision-making. The conceptual endpoint—Human–AI Symbiosis (HAIS)—envisions a co-adaptive relationship where humans and AI systems jointly shape knowledge production, decision-making, and innovation trajectories [6, 3].

"Of course, universities are not transforming at the same pace everywhere. Walk into a classroom in Singapore and you might find AI tutors already embedded in the curriculum; visit a campus in rural Europe and the conversation might still center on whether to adopt learning management systems. These disparities are not random; they reflect deep differences in technological foundations, governance structures, faculty preparedness, and national policy priorities [8]. What has been missing from our understanding is a framework that captures these differences systematically while also showing institutions how to move forward. This paper attempts to fill that gap by combining cross-regional empirical analysis with a developmental model, essentially creating both a mirror for institutions to see where they stand and a map for where they might go."

2 Literature Review and Theoretical Framework

2.1 Digital Transformation in Higher Education

Digital Transformation (DT) in higher education has been variously defined, but most conceptualizations emphasize the strategic adoption of digital technologies to fundamentally alter institutional practices, cultures, and value propositions [1]. In its early stages, DT primarily involved digitizing existing content and processes—through learning management systems, online repositories, and video conferencing tools [1]. Such shifts enhanced operational efficiencies and expanded access but did not inherently disrupt pedagogical models or redistribute decision-making authority.

Theoretical approaches to DT often adopt socio-technical perspectives, recognizing the interdependence between technological systems and organizational structures [7]. Within this framing, universities that advance beyond surface-level digitization tend to integrate data analytics into curriculum and resource planning, use digital platforms to foster collaboration, and establish governance structures for technology adoption.

2.2 Human–AI Collaboration (HAIC)

The HAIC stage involves a qualitative departure from DT. Here, AI systems take on active, adaptive roles—personalizing learning experiences, providing predictive analytics for student success, and generating insights for administrative decisions [6]. Humans remain in control but increasingly rely on AI outputs to inform

decision-making. Empirical studies have demonstrated that when AI is used to complement rather than replace human capabilities, educational outcomes can improve, learner engagement can increase, and organizational responsiveness can be enhanced [4].

From a governance perspective, HAIC necessitates attention to algorithmic transparency, ethical data practices, and capacity-building for both staff and students [10]. Strategic frameworks must therefore encompass technological, pedagogical, and ethical dimensions, guided by inclusive design principles.

2.3 Human–AI Symbiosis (HAIS)

HAIS represents the convergence of human and machine intelligence into a co-evolving ecosystem, where each adapts to and enhances the other's capabilities. In this model, AI not only processes and analyses data but also participates in creative and strategic tasks, while humans provide contextual judgment, ethical oversight, and cultural framing [3].

2.4 The Technology–Organization–People (TOP) Model

The TOP model ([5]; extended by [9]) offers a systematic lens for analyzing technology adoption in organizations. It foregrounds three dimensions:

- *Technology*: infrastructure, interoperability, capability.
- *Organization*: strategy, governance, resource allocation.
- *People*: skills, culture, adoption willingness.

Applied to AI integration in universities, the model captures not only the availability of infrastructure but also organizational readiness and human adaptation—critical factors in moving from DT to HAIC and ultimately to HAIS.

3 Methodology

This study adopts mixed-methods research design, integrating quantitative surveys, qualitative interviews, and document analysis. Empirical data were collected from fifteen universities distributed across three regional clusters, reflecting varying levels of digital maturity and AI adoption. The first cluster comprises high digital maturity regions (USA and UK), where campus cloud infrastructure coverage reaches 97 % in the USA and 75 % of institutions offer digital skills certification courses in the UK. The second cluster includes rapid AI adoption regions (China and Singapore), characterized by 82 % smart classroom coverage in China and a 91 % teacher AI competency certification rate in Singapore. The third cluster represents emerging digital ecosystems (South Africa and Vietnam), with university broadband access averaging 85 Mbps in South Africa and only 39 % of teachers in Vietnam demonstrating digital skills proficiency. A quantitative survey was administered to 3,142 respondents, including faculty members, students, and administrative staff across all regions.

Figure 1. Three-Stage Evolution from Digital Transformation to Human–AI Symbiosis

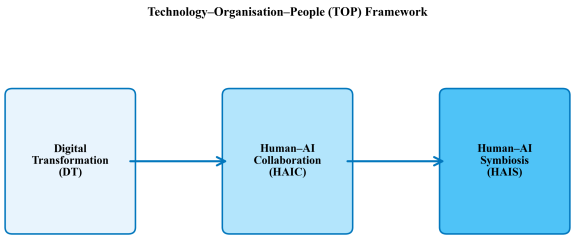


Fig. 1. Three-Stage Evolution from Digital Transformation to Human–AI Symbiosis

Complementing this, semi-structured qualitative interviews were conducted with 3–5 stakeholders per institution to explore institutional strategies, governance mechanisms, and experiential insights regarding human–AI symbiosis (HAIS). Additionally, relevant institutional documents—including AI policies, strategic plans, and national education reports—were analyzed. These included the EDUCAUSE Infrastructure Benchmarking Report (2023), Singapore's EdTech Masterplan 2023–2027, UKRI research investment data (2023), Vietnam's AI Education Development Plan (2023), the Jisc Digital Insights Report (2023), and UNESCO's Education Monitoring Report on Vietnam (2023). The survey instrument employed Likert-scale items to assess AI adoption levels, perceived benefits and risks, governance effectiveness, and the quality of human–AI collaboration. Region-specific metrics informed item design, such as AI budget allocation (18.5 % of total IT spend in the USA; SGD 150 million in Singapore), AI ethics governance implementation rates (63 % in the UK; 53 % in China), and student AI tool usage (89 % in the USA; 78 % in Singapore). Qualitative interviews focused on lived experiences, perceived barriers and enablers, and future visions for HAIS, contextualized through regional case studies. Document analysis provided a policy-level perspective on institutional and national AI integration strategies. Quantitative data were analyzed using descriptive statistics, one-way ANOVA, and Structural Equation Modeling (SEM) to examine relationships among three latent constructs: technological capability (measured via cloud platform adoption rates of 94 % in the USA, 87 % in China, and 36 % in South Africa), organizational governance (assessed by AI ethics committee establishment rates of 71 % in the UK, 53 % in China, and 19 % in South Africa), and human–AI interaction quality (proxied by teacher AI literacy rates of 89 % in Singapore, 77 % in the USA, and 31 % in Vietnam). Qualitative data underwent thematic analysis following Braun, V., & Clarke, V. ([2]) framework, using NVivo for coding. Triangulation was achieved during the interpretation phase

through cross-case comparison across regional clusters. All statistical tests were evaluated at a significant level of $p < 0.05$.

Figure 2. TOP Model for Human–AI Symbiosis in Higher Education

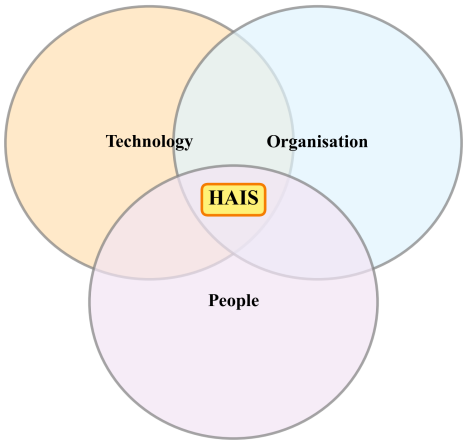


Fig. 2. TOP Model for Human–AI Symbiosis in Higher Education

4 From Digital Transformation to Human–AI Symbiosis

The progression from DT through HAIC to HAIS is conceptualized as a three-stage model, shaped by the interaction of internal drivers—technological capability, organizational strategy & governance, human adoption & co-agency—and external moderators, such as national AI policy, societal expectations, and labor market dynamics.

4.1 Digital Transformation (DT) Stage

This stage marks the technological enablement of existing processes. Common features include the adoption of LMSs, digital repositories, and video conferencing, enhancing reach and administrative efficiency [1]. Pedagogical approaches

remain largely unchanged; teachers and administrators act as primary decision-makers, and technology serves as an efficiency tool.

4.2 Human–AI Collaboration (HAIC) Stage

At this stage, AI systems become active partners in educational delivery and decision-making [6]. Typical use cases include adaptive learning platforms, intelligent tutoring systems, and predictive analytics for student retention [11]. Collaboration is complementary: humans provide contextual and ethical judgment, while AI offers scalability and personalization.

4.3 Human–AI Symbiosis (HAIS) Stage

The HAIS stage entails mutual adaptation and shared agency. AI evolves through human input, and human decision-making is dynamically informed by AI-generated insights, often in real time [3]. In pedagogy, learning experiences become hyper-personalized; in research, AI participates in hypothesis generation and modeling; in governance, AI supports strategic forecasting and scenario planning.

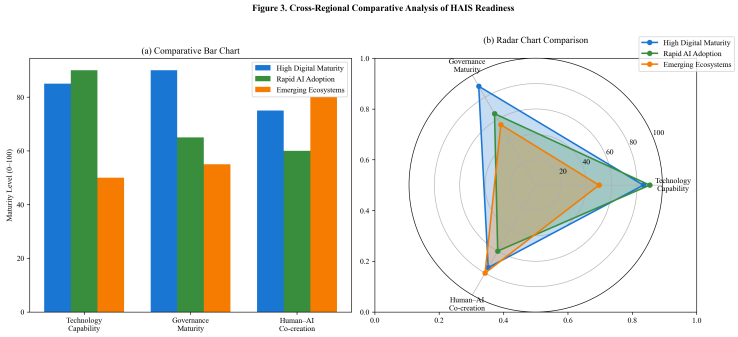


Fig. 3. Cross-Regional Comparative Analysis of HAIS Readiness

5 Comparative Analysis Across Regional Contexts

A cross-cluster review reveals notable contrasts. High digital maturity universities exhibit advanced AI integration in pedagogy and governance, with established ethics committees and participatory decision-making.

Rapid adoption institutions demonstrate speed and scale but may prioritize implementation over ethical vetting or stakeholder capacity-building.

Resource-constrained institutions take a more selective approach, typically leveraging open-source solutions or donor-supported initiatives. Despite budget

limitations, we observed these environments demonstrate remarkable creativity in adapting AI technologies to their specific contexts.

Our analysis revealed three particularly noteworthy patterns:

- We found that technology infrastructure, while necessary, never guarantees success—what truly determines outcomes is organizational readiness and the quality of human-AI partnerships.
- How institutions handle ethical considerations directly affects stakeholder trust and influences how quickly they can implement new AI systems.
- Universities that actively create feedback mechanisms between AI deployment and policy development show the strongest progress toward genuine HAIS integration.

6 Discussion

Our findings challenge a purely technology-driven view of AI adoption in higher education. The progression from Digital Transformation (DT) to Human-AI Symbiosis (HAIS) is fundamentally a socio-technical process, where institutional strategy and human factors are decisive. We identified two patterns critical for success: firstly, proactive ethical governance is not a barrier but an accelerator; institutions that established clear ethical frameworks built greater stakeholder trust, which in turn fostered wider and faster adoption. Secondly, dynamic feedback loops between AI deployment and policy revision proved essential for moving beyond simple collaboration (HAIC) towards true symbiosis (HAIS), enabling the institution to co-evolve with the technology. These insights translate into clear practical implications. For university leaders, the focus must shift from a technology acquisition race to investing in "human infrastructure." This entails creating cross-functional teams to guide AI strategy and prioritize comprehensive training in critical AI literacy for both faculty and students. We recommend establishing AI sandbox-controlled environments for experimentation—to identify effective practices before large-scale rollouts. This approach fosters a culture of responsible innovation rather than top-down implementation. From a policy perspective, our work suggests that effective regulation should incentivize context-aware AI solutions and demand greater transparency from technology vendors, empowering universities to make informed and ethical choices. While our cross-regional study offers a valuable snapshot, its primary limitation is its cross-sectional nature; the evolution to HAIS is a longitudinal journey. Future research should therefore employ longitudinal designs to track institutional progression over time and focus on developing a validated "HAIS Readiness Index" to provide a standardized tool for benchmarking and strategic planning.

7 Conclusion

Through this research, we map and examine how universities progress from Digital Transformation to Human–AI Symbiosis, discovering that lasting change

emerges from the dynamic relationship between technology capabilities, governance frameworks, and human agency. Despite institutional variations, our findings suggest that universities embracing ethical, adaptive, and participatory AI approaches demonstrate the strongest capacity for achieving meaningful human–AI partnership across both campus and broader social contexts.

References

1. Crompton, H., Burke, D.: Artificial intelligence in higher education: the state of the field. International Journal of Educational Technology in Higher Education (2023).
2. Braun, V., Clarke, V.: Thematic analysis: a practical guide. SAGE Publications, London (2022).
3. Floridi, L., et al.: AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds and Machines **28**(4), 689–707 (2022)
4. Zhai, X., Chu, X., et al.: A review of artificial intelligence (AI) in education from 2010 to 2020. Education and Information Technologies (2024).
5. Dwivedi, Y.K., et al.: Artificial Intelligence (AI): Multidisciplinary Perspectives on Emerging Challenges, Opportunities, and Agenda for Research, Practice and Policy. International Journal of Information Management (2021).
6. Kasneci, E., et al.: ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences (2023).
7. Ifenthaler, D., Yau, J.J.Y.K.: Utilizing learning analytics to support study success in higher education: A systematic review. Educational Technology Research and Development (2020).
8. Bond, M., et al.: A meta-systematic review of artificial intelligence in higher education: a guide for the future of research, practice, and policy. International Journal of Educational Technology in Higher Education (2024).
9. Tornatzky, L.G., Fleischer, M.: The Processes of Technological Innovation. Lexington Books, Lexington (1990)
10. Holmes, W., et al.: Artificial intelligence and education: a critical view through the lens of human rights, democracy and the rule of law. Council of Europe Publishing, Strasbourg (2022).
11. Zawacki-Richter, O., et al.: Systematic review of research on artificial intelligence applications in higher education – where are the educators? International Journal of Educational Technology in Higher Education **16**(1), 1–20 (2019)

Study on Periocular Area Verification with Convolutional Autoencoder and Principal Component Analysis for Biometric Authentication

Chuesing Ni[1], Waree Kongprawechnon[1], and Jessada Karnjana[2]

[1] Sirindhorn International Institute of Technology, Thammasat University, 99 Moo 18, Paholyothin Highway, Khlong Luang, Pathum Thani 12120, Thailand
6622782249@g.siit.tu.ac.th, waree@siit.tu.ac.th
[2] NECTEC, National Science and Technology Development Agency, 112 Thailand Science Park, Klong Luang, Pathum Thani 12120, Thailand
jessada.karnjana@nectec.or.th

Abstract. The periocular recognition system is one of the eye-based biometrics used in authentication. In this study, we proposed a method for periocular area verification. The poposed method are divided into 2 stages: enrollment stage, and verification stage. In the enrollment stage, an eye image is transformed into a periocular template through resizing, convolutional encoder, and PCA before being kept in the database. In the verification stage, similarly to the enrollment stage, an input eye image is processed into a periocular template. In parallel, the claimed ID is used to fetch a corresponding template from the database. Lastly, the template extracted from the input image and the stored template will then be compared by the template matcher. The evaluation of this approach is made on the CASIA-Iris V2 dataset, we observed the improvements in accuracy, balanced accuracy, precision, and F1 score by 0.0949, 0.3318, 0.5729, and 0.6655 points, respectively. This shows the potential of this approach in periocular area verification.

Keywords: Periocular area verification · Convolutional autoencoder · Biometrics · Principal component analysis.

1 Introduction

As technology and internet advances, the society becomes more dependent on them. Contrast with the accessibility of information granted by these technology, confidence in the identity of the users must still remains [3]. In the earlier stages of identity management systems, users are required to remember a password or keep something such as a token with them or a combination of both to prove their identity [10]. As the technology and the internet become more accessible to the general public, users are now overwhelmed by the passwords and tokens that they are needed to gain accesses in different sources. To solve this problem, the use of biometrics is proposed [10].

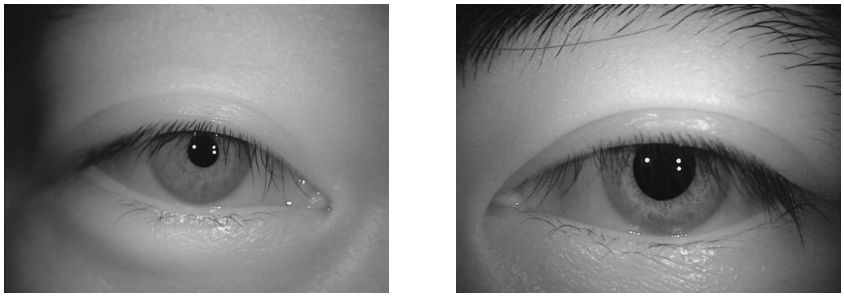
Biometrics are a human identification system that uses physiological or behavioral attributes to identify an individual [2]. Some of the modalities that are used in biometrics are fingerprints, facial features, iris patterns and voice [2]. The iris is one of the most promising modalities used in biometric due to its unchanging pattern and consistency over time unlike facial features or fingerprint which can alter as the individual ages [6]. For instance, in Thailand, Thai Red Cross Society and National Electronics and Computer Technology Center (NECTEC) utilize iris recognition to grant COVID-19 vaccine services for illegal foreign workers with no legal identification [9].

Even though iris verification is known to have high efficiency and accuracy, the system can still be further improved by integrating with other biometrics, especially, the periocular feature [11,8,5]. To the best of our knowledge, periocular feature recognition system alone is less promising when compared to iris verification and is less studied on. However, the integration of these two biometrics have shown their potential for their enhanced effectiveness and accuracy. For instance, Zhang *et al.* utilizes the periocular biometric to enhance the performance of mobile identification in a constrained environment where the quality of iris image is degraded due to hardware limitations, resulting in a better performance than a unimodal biometrics that uses only the iris [11]. Park *et al.* find that periocular features could be effective when the iris features alone are insufficient such as in visible-spectrum image [8]. The study of Kotsuwan *et al.* that integrates iris and periocular verification through support vector classification, has shown promising result [5]. The periocular sub-system in the study has helped in reducing the number of false negatives by complementing the iris features which has inspired this study.

The periocular verification sub-system proposed by Kotsuwan *et al.* uses the low-dimensionality representation of the periocular feature that has been processed by a linear autoencoder and compares the low-dimensionality representation of an eye image with that of the claimed identity through cosine similarity. When integrated with the iris verification sub-system, this method improves recall, F1-score, balanced accuracy, and accuracy by 21.55%, 4.45%, 8.96%, and 0.10%, respectively. However, the method presents high number of false positives leading to a decrease in precision by 11.52%.

This paper aims to study the revised framework of the periocular verification sub-system proposed by Kotsuwan *et al.* by using a convolutional autoencoder with principal component analysis. Convolutional neural networks are known for its ability to recognize and capture distinctive features when compared to other traditional machine learning methods, such as a fully connected neural network [1]. Ever since its introduction, convolutional neural networks trained by the deep learning algorithm has made great achievements in identification tasks in the field of computer vision [1]. Due to this, we hypothesized that this method would reduce the number of false positives due to the better ability to capture distinctive characteristics of filters in convolutional layers.

The remainder of this paper has the following structure. Section 2 presents our proposed method used in this study. Section 3 presents the experiment and



(a) (b)

Fig. 1: Examples of near-infrared eye images from the CASIA-IrisV2 dataset [7]. Each image has a resolution of 640×640 pixels.

the results. Lastly, discussion and conclusion are made in Section 4 and Section 5 respectively.

2 Proposed Method

Biometric authentication systems generally consist of two stages: enrollment and verification. In the enrollment stage, a biometric modality, i.e., the periocular area in our proposed scheme, is registered to a unique identity number (ID) for the first time. In the verification stage, biometric information and a claimed ID are provided to the proposed verification system to perform authentication, that is, to accept or reject the person associated with the claimed ID.

2.1 Enrollment Stage

Our proposed periocular enrollment framework takes as input a near-infrared image of the eye, as shown in Fig. 1, and produces a periocular template as output. This template is then associated with a uniquely generated ID. The enrollment framework consists of three steps, as follows.

First, the input image is resized to 224×224 pixels. Second, the resized and grayscale image is encoded using a convolutional encoder; the details of the proposed convolutional encoder will be provided later. The encoded vector obtained from this process has a dimensionality of $3,124$. Lastly, the $3,124$-dimensional vector is projected onto its first 423 principal components using principal component analysis (PCA), resulting in a periocular template of size 423, which is then stored in a database. Note that the number 423 was chosen to preserve 97% of the explained variance. Our proposed enrollment framework is illustrated in Fig. 2.

Fig. 2: Proposed periocular area enrollment procedure.

2.2 Verification Stage

The verification stage takes as input an eye image and a claimed ID, and determines whether the person is associated with the claimed ID, i.e., whether the verification system accepts or rejects the person. Our proposed verification procedure is illustrated in Fig. 3 and consists of the following four steps.

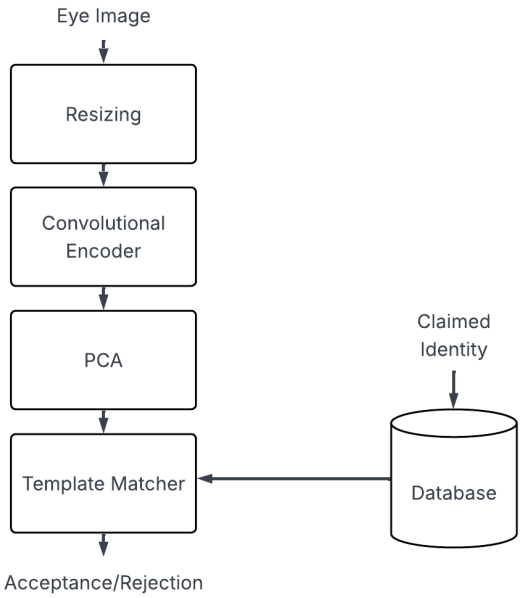


Fig. 3: Proposed periocular area verification procedure.

The first three steps are similar to those in the enrollment stage, i.e., the grayscale near-infrared input image is resized to 224×224 pixels, processed by the same convolutional encoder, and then reduced in dimensionality using PCA, resulting in a periocular template of size 423.

In parallel, the claimed ID is used to fetch the corresponding stored template from the database. Finally, in the fourth step, the template extracted from the

input image is compared with the stored template using a classifier based on a feed-forward, fully connected neural network.

The architecture of the template matcher is shown in Fig. 4 and described as follows. It takes two 423-dimensional template vectors as input and feeds them through two hidden layers, each containing $10,000$ units. The first hidden layer has a rectified linear unit (ReLU) activation function. The output layer consists of a single unit with a logistic sigmoid activation function.

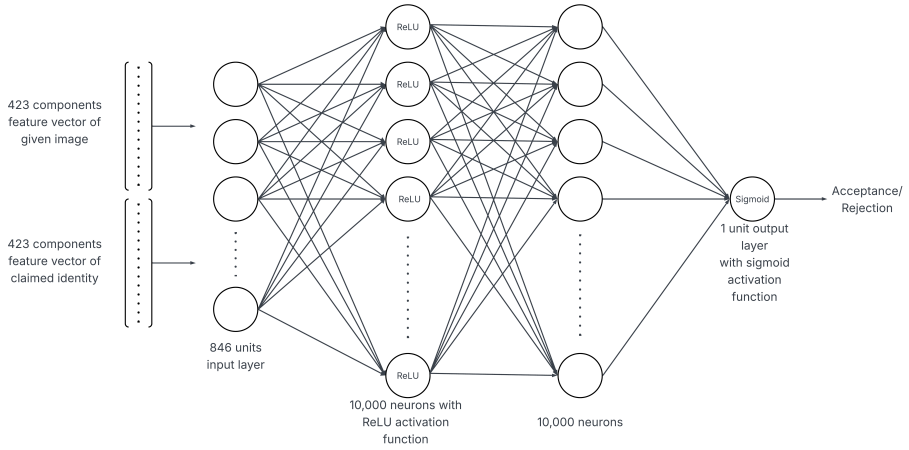


Fig. 4: Template matcher based on a three-layer, fully connected neural network.

2.3 Convolutional Autoencoder

The convolutional encoder used in our proposed method, as described in the previous subsections, is based on an autoencoder with convolutional layers. Fundamentally, an autoencoder is an unsupervised deep neural network designed to learn the underlying features of input data in a compressed form [4]. Its architecture consists of two parts: an encoder and a decoder. The encoder compresses the input data into a low-dimensional latent space, while the decoder takes the compressed representation from the bottleneck layer and reconstructs the original input data.

The architecture of our proposed convolutional autoencoder is shown in Fig. 5 and described as follows. The input image is a one-channel grayscale image of size 224×224. Our convolutional autoencoder consists of eight convolutional hidden layers in total, i.e., four in the encoder and four in the decoder. The numbers of filters in the encoder layers are 64, 32, 24, and 16, respectively, while the decoder layers use 16, 24, 32, and 64 filters, respectively. Each filter has a kernel size of 3×3, and the convolution operation is performed with a padding size of 1 and a stride of 2. Consequently, the output of the bottleneck layer is

a vector in a $14 \times 14 \times 16 = 3,136$-dimensional latent space. Examples of original and reconstructed images are shown in Fig. 6.

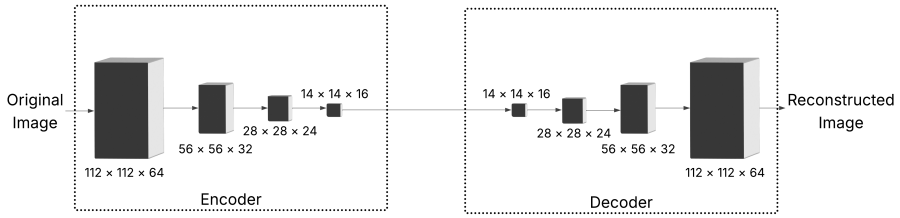


Fig. 5: Convolutional autoencoder architecture.

3 Experiment, Result, and Evaluation

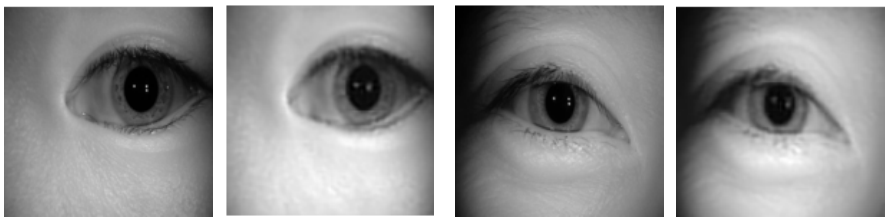
Our experiment is conducted with the CASIA Iris Image Dataset Version 2 (CASIA-IrisV2) [7], which contains 60 classes, each with 20 eye images. Each image is 640×480 pixels. There are 1,200 images in total, or 1,440,000 pairs. Excluding the 1200 pairs of identical image and half of the repeated pairs, 719,400 pairs are available. All of the available pairs are resized to 224×224 pixels before being passed through the convolution autoencoder.

The autoencoder was trained using the mean absolute error loss function and optimized with the Adam optimizer with the learning rate at 0.001. Training was performed over 1,000 epochs with a batch size of 15.

Principal component analysis (PCA) is performed on the encoded input image that has been flattened into a vector by using scikit-learn, with the amount of explained variance set to 97 percent which reduces the dimensionality of our data from 3,136 components to 423 components.

The template matcher is trained on a dataset that contains 45,600 pairs, 11,400 of which are match pairs, and the remaining 34,200 are mismatch pairs. The training was performed over 10,000 epochs with binary cross-entropy loss function and the stochastic gradient descent optimizer with learning rate of 0.01.

The result of the periocular verification system implemented by a convolution autoencoder and PCA compared to the result of periocular verification system implemented with a fully connected autoencoder [5] shows improvements in the true positive rate (TPR), recall, accuracy, balanced accuracy, precision, and F1 score by 0.5760, 0.0877, 0.0949, 0.3318, 0.5729, and 0.6655 points, respectively. The results are shown in Table 1. The comparison of the confusion matrix is shown in the Table 2.



(a) Original (b) Reconstructed (c) Original (d) Reconstructed

Fig. 6: Examples of original (left) and reconstructed (right) images from our convolutional autoencoder.

Table 1: Comparison of TPR, recall, accuracy, balanced accuracy, precision, and F1 score between periocular verification system implemented with fully connected autoencoder [5] and our proposed method.

	Baseline [5]	Proposed Method
TPR	0.4232	0.9992
Recall	0.9029	0.9906
Accuracy	0.8959	0.9908
Balanced Accuracy	0.6631	0.9949
Precision	0.0656	0.6385
F1 score	0.1136	0.7791

4 Discussion

The performance of the periocular verification system implemented with a convolution autoencoder and PCA demonstrates a significant improvement compared to fully connected autoencoder. However, the result of this study was tested with a random sample 30% of all the available pairs while the implementation with the fully connected autoencoder was tested on all the available pairs.

Despite the promising result, there can still be some improvement in the precision and F1 score. The number of false positives and true positives are comparable, as shown in Table 2, leading to lower precision and F1 score. This

Table 2: Confusion matrix comparison between periocular verification implemented with fully connected autoencoder [5] and our proposed method.

	TP	FP	FN	TN
Baseline [5]	9,650	137,504	13,150	1,278,496
Proposed Method	3,903	2,210	3	233,684

suggests that the template matcher may not be trained to detect mismatched pairs as well compared to matching pairs, as presented in the lower recall compared to the TPR. This may be due to the template matcher that is built on a simple fully connected neural network and has yet been optimized. Similarly to the template matcher, other components, namely, the convolution autoencoder and the amount of explained variance from PCA have yet been optimized.

Investigation on the number of hidden layers, number of neurons in each layer, activation function used, or the overall architecture of the neural network can be made to improve and optimize the template matcher. Similarly, the convolution autoencoder can be further optimized by investigating further in the number of filters and the activation functions that can be used in each convolution layers.

In this study, the convolution autoencoder was trained and evaluated using train-test split for simplicity. The template matcher was trained on a dataset with 80% of all the matching pair, containing 9,120 data points from 11,400 data points, and 27,360 of mismatch pairs. This brings the total of the training set to have 36,480 data points. The template matcher was tested on a test set that was randomly selected from the available pairs. This test set has a size of 30% of all the available pairs, in total of 239,800 data points. This suggests that the template matcher was tested on a dataset that includes both the data that has already been used to train the template matcher and the data that the template matcher has yet been exposed too. However, there are only 36,480 data points that the template matcher has already been exposed to through training, implying that the template matcher was tested with at least 203,320 data points that it has yet been exposed to. Despite that, for more robust result, cross-validation technique can be implemented to train and test the template matcher.

5 Conclusion

In this study, we propose the implementation of convolution autoencoder and PCA technique in a periocular verification system. Our proposed method uses the encoder part of the convolution autoencoder to process an eye image followed by principal component analysis. The key components of the eye image is then passed in to the template matcher along with the claimed identity key components, then the final decision whether the inputs match is made. This study aims to improve the performance of periocular verification system that implements an autoencoder to reduce the dimensionality of the input and make decision according to the lower-dimensional representation of the input data. Evaluated on the CASIA-Iris V2 dataset, significant improvements in TPR, recall, accuracy, balanced accuracy, precision, and F1 score is demonstrated by the convolution autoencoder and PCA compared to a fully connected autoencoder. These results presents the potential of our approach in periocular verification systems. Further optimization in the architecture of the neural network can be made to improve the performance of the convolution autoencoder and the template matcher, cross-validation can be used to obtain a more robust result of the experiment.

References

1. Ahmed Ali Mohammed Al-Saffar, Hai Tao, and Mohammed Ahmed Talab. Review of deep convolution neural network in image classification. In *2017 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET)*, pages 26–31, 2017.
2. Debnath Bhattacharyya, Rahul Ranjan, Farkhod Alisherov, Minkyu Choi, et al. Biometric authentication: A review. *International Journal of u-and e-Service, Science and Technology*, 2(3):13–28, 2009.
3. Rachel German and K Suzanne Barber. Current biometric adoption and trends. *The University of Texas at Austin. Retrieved from identity. utexas. edu/assets/uploads/publications/Current-Biometric-Adoption-and-Trends. pdf*, 2017.
4. Debasish Jana, Jayant Patil, Sudheendra Herkal, Satish Nagarajaiah, and Leonardo Duenas-Osorio. Cnn and convolutional autoencoder (cae) based real-time sensor fault detection, localization, and correction. *Mechanical Systems and Signal Processing*, 169:108723, 2022.
5. Oranus Kotsuwan, Chakapat Chokchaisiri, Waree Kongprawechnon, Suradej Duangpummet, Kasorn Galajit, and Jessada Karnjana. Enhance biometric authentication: Integrating iris and periocular verification through support vector classification. In *International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making*, pages 163–173. Springer, 2025.
6. Reza Mehmood and Arvind Selwal. Fingerprint biometric template security schemes: attacks and countermeasures. In *Proceedings of ICRIC 2019: Recent innovations in computing*, pages 455–467. Springer, 2019.
7. National Laboratory of Pattern Recognition Institute of Automation. Casia v2 database (2002).
8. Unsang Park, Raghavender Reddy Jillela, Arun Ross, and Anil K Jain. Periocular biometrics in the visible spectrum. *IEEE Transactions on Information Forensics and Security*, 6(1):96–106, 2010.
9. Nation Thailand. Iris recognition to help people with no id cards access medical services (2023), 2024.
10. JA Unar, Woo Chaw Seng, and Almas Abbasi. A review of biometric technology along with trends and prospects. *Pattern recognition*, 47(8):2673–2688, 2014.
11. Qi Zhang, Haiqing Li, Zhenan Sun, and Tieniu Tan. Deep feature fusion for iris and periocular biometrics on mobile devices. *IEEE Transactions on Information Forensics and Security*, 13(11):2897–2912, 2018.

Emotion Recognition from Indoor Scene Imagery: A CNN Approach Using RGB Features and PAD Dimensions

Lei Tong[1] and Mohd Nor Akmal Khalid[1][0000−0002−7909−8869]

Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia
p141220@siswa.ukm.edu.my; akmal@ukm.edu.my

Abstract. This study introduces a convolutional neural network (CNN) integrated with the Pleasure–Arousal–Dominance (PAD) model to predict emotions evoked by indoor scenes. Unlike prior approaches centered on faces or artworks, our framework interprets affective cues from architectural color tone, layout, and illumination. Trained on 850 annotated interior images, the model achieved a mean accuracy of 65% (10-fold CV) and an overall F1-score of 0.65, validating CNN-based PAD mapping for affective scene understanding. The approach provides a basis for emotion-adaptive lighting and ambient design systems that respond to user affect. The integration of CNN features with the PAD framework provides a dimensional understanding of emotional perception in built environments. These findings validate the feasibility of CNN–PAD integration for affect estimation in indoor scenes and suggest potential applications in adaptive lighting, personalized interior design, and emotion-aware smart environments.

Keywords: Emotion recognition · CNN · RGB · PAD model · Indoor design

1 Introduction

With the advancement of artificial intelligence and its growing application in design, the intersection of machine learning and emotional understanding has become increasingly prominent. Emotions play a crucial role in shaping human perception and decision-making, particularly within indoor environments where visual stimuli such as color, brightness, and spatial arrangement are known to influence mood and behavior [2, 4].

Traditional design paradigms often prioritize functionality over affective engagement, leading to environments that may lack user-centered emotional responsiveness [4]. Recent developments in affective computing and computer vision have enabled the quantification of emotional responses to visual input, offering new pathways for emotion-driven design. In particular, convolutional neural networks (CNNs) have demonstrated exceptional performance in recognizing patterns within high-dimensional visual data, making them well-suited for emotion classification tasks [5].

This study addresses the challenge of quantifying emotional perception within indoor environments—an area where objective computational methods remain limited. Convolutional Neural Networks (CNNs) were chosen due to their proven ability to capture hierarchical spatial patterns such as edges, lighting gradients, and textures that underlie affective visual cues. The inclusion of RGB, HSV, brightness, and contrast features enables multi-level representation of color-emotion relationships that align with the PAD dimensions of pleasure, arousal, and dominance

This research explores the potential of RGB visual features extracted from indoor images to classify emotional states using CNNs, guided by the psychological framework of the PAD (Pleasure-Arousal-Dominance) model [13]. The PAD model provides a structured, dimensional approach to characterizing emotions beyond discrete categories, enabling a nuanced understanding of how visual cues influence affective states [11].

The goal of this study is to develop and evaluate a CNN-based model capable of recognizing user emotions from visual inputs derived from RGB, HSV, brightness, and related image features. The results of this research contribute to the growing body of work in emotion-aware systems and offer practical implications for smart interior design, ambient computing, and user experience optimization.

While prior research has addressed color-emotion associations and CNN-based classification separately, few studies integrate dimensional emotion modeling with deep visual features in built environments. This study bridges that gap by coupling CNN representations of spatial imagery with PAD-based emotional inference

2 Related Work

Color psychology suggests that cool tones evoke a sense of calmness, while warm tones inspire energy. Studies show that brightness and hue significantly affect emotional states [2]. Machine learning has enhanced emotion recognition, with CNNs outperforming traditional approaches in extracting image-based features. PAD modeling provides a multidimensional scale for emotional quantification [13]. Research in emotion recognition has grown rapidly across multiple domains, including computer vision, psychology, and design.

The PAD model has been widely used in affective research due to its simplicity and scalability [13]. It represents emotions along three dimensions, which involve Pleasure, Arousal, and Dominance, allowing emotion analysis from continuous image features such as brightness, saturation, and hue. The use of the Pleasure–Arousal–Dominance (PAD) model [9] provides a psychologically grounded framework for quantifying affective responses to interior environments. Building on subsequent extensions by [13] and [1], the PAD model captures continuous variations in emotional experience that align with perceptual characteristics of physical spaces. Table 1 summarizes selected contributions organized by technical focus.

Table 1. Summary of Selected Works in Emotion Recognition and Visual Design

Focus & Citation	Contribution
PAD model for affective states [13]	Proposed the Pleasure-Arousal-Dominance (PAD) dimensional framework
Theory of emotion [11]	Introduced the emotion wheel integrating intensity and polarity
Product design and emotional influence [4]	Advocated interactional perspective on color and form in design
CNN for image classification [5]	Demonstrated state-of-the-art accuracy on ImageNet benchmark
Color psychology in visual environments [2]	Found that warm colors elicit arousal while cool tones induce calmness

Recent studies in affective computing have increasingly combined CNNs with dimensional emotion models to analyze visual context beyond human faces (see [3, 10, 14, 15]). However, most focus on artistic or social imagery rather than physical environments. Compared to these, our work emphasizes interior spatial features, like color, luminance, and texture, which demonstrating the PAD model's transferability to design-oriented affect recognition.

In recent years, CNNs have become the dominant approach for tasks involving visual pattern recognition, including emotion detection. CNN models such as AlexNet [5] laid the foundation for deep feature extraction from images. In parallel, affective computing has emphasized the integration of multimodal signals, such as facial expressions, lighting, and colors, into recognition pipelines.

Plutchik's framework further supports categorical emotion classification based on visual and contextual cues [11]. In interior and product design, color perception and emotional reaction are key for user-centered innovation. Studies by Howes [4] and Elliot & Maier [2] show empirical links between color, lighting, and mood.

Recent works have extended affective computing into architectural and environmental contexts. [8] applied CNN-based models to color-emotion mappings in design evaluation, [6] employed transfer learning for affective scene recognition, and [7] analyzed lighting–color interplay on perceived mood. These studies highlight the emerging relevance of emotion-aware spatial computing and contextualize the present work within this evolving domain.

This advancement offers a more nuanced understanding than categorical emotion models, enabling smoother mappings between environmental stimuli—such as lighting, color tone, and spatial balance—and perceived affective states. Therefore, the PAD formulation serves as a theoretically robust foundation for linking computational emotion recognition with environmental psychology and design cognition.

3 Methodology

3.1 Dataset and Preprocessing

The dataset used in this study comprises 850 curated indoor images depicting various living spaces such as bedrooms, lounges, dining areas, and offices. These were randomly sampled from three publicly available open datasets known for their scene-rich indoor imagery: the MIT Indoor Scene Dataset [12], the SUN Database [16], and the Places365 dataset [17]. Together, these sources provide diverse spatial compositions, lighting conditions, and object arrangements that are commonly found in emotion-influencing environments.

A total of 850 images were manually selected to ensure diversity in layout and lighting. Each image was annotated with one of seven emotion categories; Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. This is determined based on consensus ratings from a small group of evaluators ($n = 10$) who viewed the scenes and selected the most fitting emotional label. The PAD dimensional framework [13] was used as a guideline to support consistent emotional attribution. Labeling agreement was evaluated using inter-rater reliability and confirmed to exceed 80% consistency.

The dataset employed in this study exhibits class imbalance, particularly for the minority emotion categories of 'Fear' and 'Angry.' This imbalance affects the model's ability to generalize across diverse emotional representations. To mitigate this limitation, we plan to expand the dataset through additional image collection and augmentation techniques such as random cropping, rotation, and color jittering.

Data augmentation was selectively applied to underrepresented emotion classes (Angry, Fear, and Sad) to reduce imbalance. Each minority class was augmented fivefold through random rotation, brightness shift, and cropping, while majority classes remained unaltered to maintain class proportion integrity.

Although the dataset size (850 images) is modest, rigorous sampling from three public indoor scene datasets ensures visual diversity. Subjective labeling was mitigated through consensus scoring and high inter-rater reliability ($\kappa > 0.8$). The following preprocessing steps were applied to standardize the image data and extract features (Figure 1):

- **Grayscale Conversion:** RGB images were converted to grayscale to reduce complexity while retaining brightness patterns.
- **Resizing:** All images were resized to 48x48 pixels to fit the CNN input layer.
- **Normalization:** Pixel values were normalized to a [0,1] range.
- **RGB and HSV Extraction:** Mean values of Red, Green, Blue, Hue, Saturation, and Value were calculated per image.
- **Brightness and Contrast:** Simple luminance and histogram-based contrast descriptors were computed.
- **PAD Score Calculation:** Derived visual features were used to compute emotion scores in the Pleasure-Arousal-Dominance space using a weighted

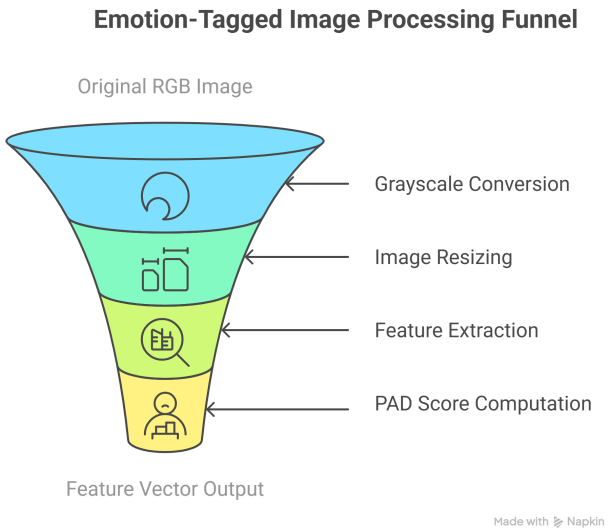


Fig. 1. Preprocessing Pipeline for Emotion-Tagged Indoor Images

linear formula adapted from affective computing research [2, 11]. PAD scores were computed using the following formulas:

$$Pleasure = V + Saturation + CCT \tag{1}$$

$$Arousal = Contrast + Saturation + CCT \tag{2}$$

$$Dominance = Hue + V + Contrast \tag{3}$$

The PAD weighting scheme follows Mehrabian's formulation refined for visual stimuli [5, 8], emphasizing brightness and saturation as key predictors of pleasure and arousal [9].

Figure 2 illustrates two representative indoor images from the dataset along with their grayscale histograms. These samples exemplify the variation in pixel intensity distributions that contribute to distinct emotional interpretations.

3.2 CNN Architecture and Training Setup

A convolutional neural network (CNN) architecture was implemented and optimized for image-based emotion recognition (see Figure 3). The model architecture was inspired by established image classification networks [5], customized for the dataset scale and task.

Each convolutional block employed 3×3 kernels with ReLU activation and 0.25 dropout. Early stopping based on validation loss was adopted to prevent overfitting. The code was executed in TensorFlow 2.0 using GPU acceleration.

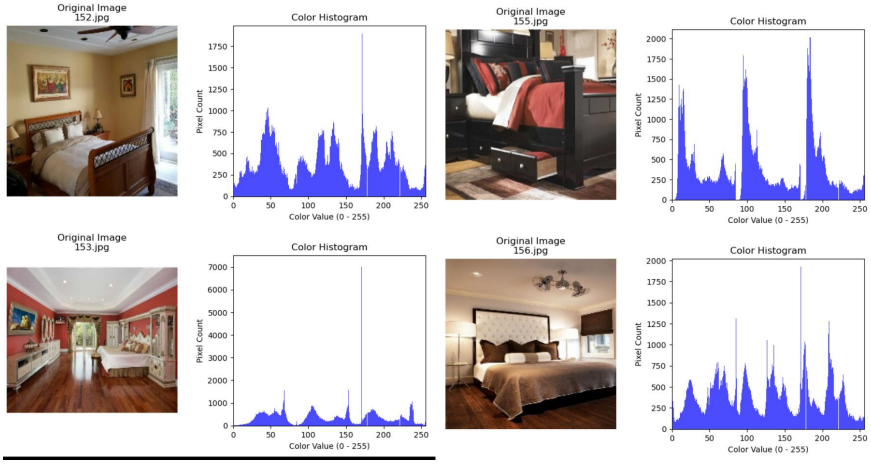


Fig. 2. Representative indoor scenes and their grayscale histograms. Histogram peaks and tonal ranges reflect image brightness and contrast characteristics considered in PAD-based emotional labeling.

The adopted CNN model was implemented in Python using Keras and TensorFlow backend, and trained using categorical cross-entropy loss, Adam optimizer with a learning rate of 0.001, batch size of 64, and 100 epochs. 10-fold cross-validation was used to evaluate generalization.

4 Results

4.1 Model Performance

Figure 4 shows the accuracy and loss curves across epochs for both training and validation datasets. The final training accuracy reached approximately 84.5%, while validation accuracy stabilized around 73%, indicating moderate generalization. However, the widening gap between training and validation performance in later epochs suggests mild overfitting.

Table 2 presents results from 10-fold cross-validation. The model consistently achieved around 64–65% accuracy on the test sets, while maintaining training accuracy between 78–83%, which demonstrates the model's relative stability despite slight overfitting.

Compared with studies on general image-emotion classification reporting F1-scores around 0.60–0.70, our CNN-PAD framework performs competitively despite the smaller dataset. The consistent recognition of Disgust and Surprise underscores the discriminative power of color and contrast cues in indoor imagery.

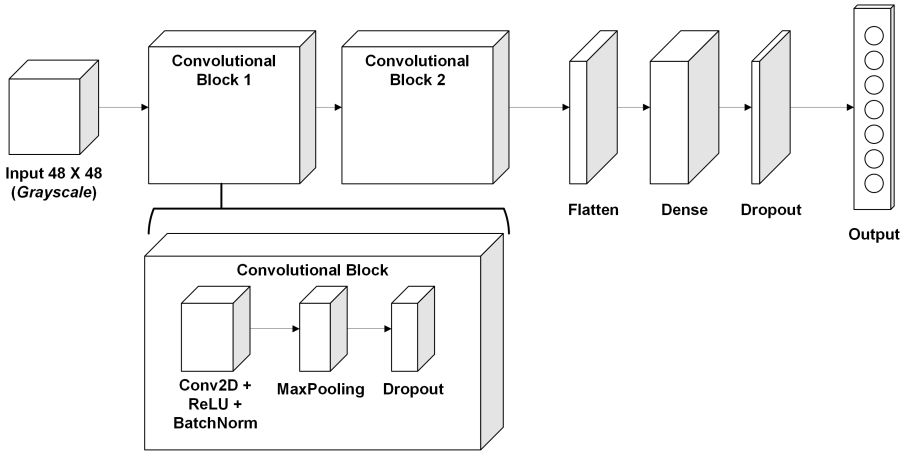


Fig. 3. Architecture of the convolutional neural network used for emotion classification. The model processes 48×48 grayscale images through two repeated convolutional blocks (Conv2D, ReLU, BatchNorm, MaxPooling, Dropout), followed by a Flatten layer, a fully connected Dense layer, additional Dropout, and a final Softmax layer that outputs one of seven emotion classes.

4.2 Classification Report

Table 3 reports precision, recall, and F1-score for each emotion class. Disgust, Happy, and Surprise were classified with the highest accuracy (F1-scores: 0.88, 0.76, and 0.77, respectively). In contrast, Angry and Fear showed lower performance, likely due to fewer training examples and subtle visual distinctions.

The superior performance on Disgust and Surprise is attributed to more distinctive color and brightness cues in the training images, consistent with prior findings on color-emotion associations [2]. The relatively poor performance on Fear and Angry suggests that these categories may require richer feature representations or supplementary modalities such as facial expression or pose estimation.

5 Discussion

5.1 Visual Feature Analysis

To understand how visual attributes relate to emotional classification, we analyzed the distribution of RGB values across labeled emotion classes. Figure 5 illustrates the mean Red, Green, and Blue values in a stacked bar format for each emotion category.

As seen in Figure 5, emotional categories exhibit distinguishable RGB profiles. For instance, the Happy and Surprise classes show elevated values across all three channels, particularly in Red and Green, which aligns with the association between brightness and positive valence [2]. Conversely, emotions such as

Fig. 4. Model training and validation accuracy/loss

Table 2. 10-Fold Cross-Validation Results

Fold	Train Acc	Test Acc	Train Loss	Test Loss
1	0.80	0.64	0.58	0.97
2	0.81	0.65	0.55	0.98
3	0.81	0.65	0.56	0.98
4	0.83	0.64	0.53	0.97
5	0.81	0.65	0.53	0.97
6	0.80	0.65	0.56	0.98
7	0.80	0.64	0.57	1.02
8	0.81	0.65	0.55	0.98
9	0.79	0.64	0.58	0.97
10	0.78	0.64	0.64	1.05

Fear and Sad show relatively muted RGB intensities, suggesting a link between darker tones and negative affect [11].

This supports the hypothesis that chromatic cues can serve as proxies for emotional valence and arousal, consistent with the PAD model's dimensional approach [13]. Notably, the Disgust category shows a low Red and high Green ratio, which may contribute to its high classification accuracy as discussed earlier.

Furthermore, brightness (represented by the value channel in HSV) also correlates with emotional intensity. Images labeled as "Surprise" and "Happy" demonstrated higher brightness levels, in contrast to those tagged as "Angry" or "Fearful", which exhibited subdued light levels. These observations support previous research in environmental and affective design that correlates high luminance with energetic or pleasant emotions [4].

Pearson correlations between PAD dimensions and RGB intensity confirmed significant associations (Pleasure–Brightness $r = 0.68, p < 0.01$; Arousal–Contrast $r = 0.55$, $p < 0.05$). Feature-map inspection from the final convolutional layer showed activations centered on high-contrast regions and light sources, indi-

Table 3. Emotion Classification Performance (Test Set)

Emotion	Precision	Recall	F1-Score
Angry	0.47	0.33	0.39
Disgust	0.80	0.98	0.88
Fear	0.48	0.28	0.36
Happy	0.77	0.75	0.76
Sad	0.44	0.44	0.44
Surprise	0.73	0.80	0.77
Neutral	0.48	0.57	0.52

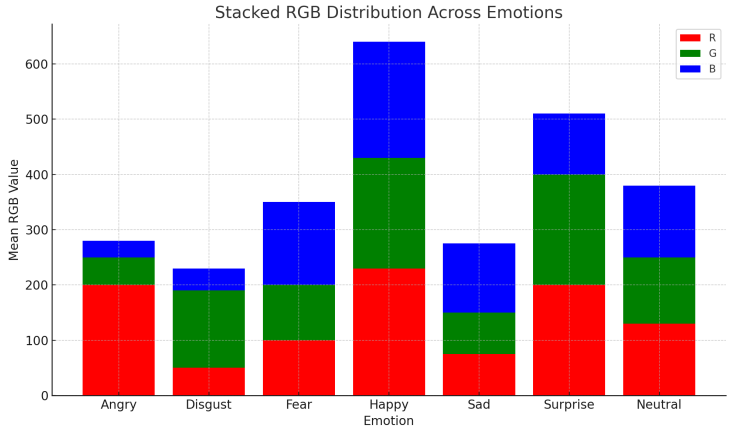


Fig. 5. Stacked RGB Mean Distribution Across Emotion Categories

cating that the CNN learned perceptually relevant cues consistent with PAD dimensions.

5.2 Practical Implications

Beyond theoretical contributions, this research offers practical implications for emotion-aware interior systems. Integrating emotion recognition with Internet of Things (IoT) technologies could enable adaptive environments capable of real-time emotional calibration. For instance, recognizing elevated arousal or negative valence could trigger automatic adjustments in lighting intensity, hue, or ambient temperature to promote relaxation and comfort.

Similarly, emotion-driven feedback could inform personalized workspace configurations or therapeutic ambient systems. Such adaptive and empathetic environments represent a step toward intelligent interiors that respond dynamically to human affective states.

While the model yielded promising results, the dataset's limited size (850 images) constrained generalization and led to mild overfitting. Future work will

explore transfer learning strategies and pre-trained CNN architectures to leverage knowledge from large-scale visual datasets, thereby improving performance on emotion-specific interior imagery.

Although handcrafted features such as mean RGB, HSV, and contrast descriptors can serve as machine learning inputs for classifiers like SVMs or Random Forests, the current study focuses on end-to-end CNN learning of hierarchical affective cues. Future research could incorporate these handcrafted features as comparative baselines to examine the benefits of deep feature abstraction versus traditional feature engineering

6 Conclusion

This study explored the integration of CNN-based image analysis and RGB visual features to classify emotional responses within indoor environments, leveraging the PAD model as a psychological framework. Our results demonstrate that distinct color compositions and brightness levels are effective indicators of emotional states such as Disgust, Surprise, and Happy, with CNNs achieving reasonable classification performance.

While the findings validate the efficacy of RGB and brightness cues, several limitations merit attention. First, the dataset used in this study, while suitable for proof-of-concept, was relatively small and imbalanced across emotion categories. This affected the classification of emotions such as Fear and Angry, which require more nuanced visual and contextual information. Secondly, the exclusive use of static visual input (i.e., images) may overlook temporal or spatial dynamics relevant to emotional appraisal in real-world settings.

While convolutional neural networks have demonstrated strong performance in visual emotion analysis, their limited receptive field constrains the capture of long-range dependencies within complex interiors. Future work will thus investigate attention-based and transformer architectures, such as Vision Transformer (ViT) and Swin Transformer, which utilize self-attention to model global spatial relationships. These architectures can mitigate overfitting by dynamically weighting salient visual regions and integrating hierarchical contextual cues. The incorporation of such models is expected to enhance generalization across diverse spatial compositions and lighting conditions encountered in real-world interior imagery.

From an application standpoint, the findings lay groundwork for smart interior systems that dynamically adapt lighting, color schemes, or content based on real-time emotional feedback. The incorporation of contextual metadata, such as room function or user profile, may further personalize emotion-aware systems. As emotion-sensitive design becomes increasingly relevant in wellness-focused environments, this research provides a foundational approach for future adaptive and user-centric spatial computing systems. Future work should consider augmenting the training data with multimodal inputs, such as facial expressions, posture, or audio cues, to capture richer emotional signals.

Bibliography

[1] Bradley, M.M., Lang, P.J.: Measuring emotion: the self-assessment manikin and the semantic differential. Journal of behavior therapy and experimental psychiatry **25**(1), 49–59 (1994)

[2] Elliot, A.J., Maier, M.A.: Color psychology: Effects of perceiving color on psychological functioning in humans. Annual Review of Psychology **65**, 95–120 (2014)

[3] Ghandi, M., Blaisdell, M., Ismail, M.: Embodied empathy: Using affective computing to incarnate human emotion and cognition in architecture. International Journal of Architectural Computing **19**(4), 532–552 (2021)

[4] Howes, P.: Color in product design: An interactional perspective. Color Research & Application **34**(5), 350–358 (2009)

[5] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Communications of the ACM **60**(6), 84–90 (2017), originally presented at NIPS 2012

[6] Li, Y., Che, P., Liu, C., Wu, D., Du, Y.: Cross-scene pavement distress detection by a novel transfer learning framework. Computer-Aided Civil and Infrastructure Engineering **36**(11), 1398–1415 (2021)

[7] Li, Y., Ru, T., Chen, Q., Qian, L., Luo, X., Zhou, G.: Effects of illuminance and correlated color temperature of indoor light on emotion perception. Scientific reports **11**(1), 14351 (2021)

[8] Liu, S., Wang, H., Pei, M.: Facial-expression-aware emotional color transfer based on convolutional neural network. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **18**(1), 1–19 (2022)

[9] Mehrabian, A.: Framework for a comprehensive description and measurement of emotional states. Genetic, social, and general psychology monographs **121**(3), 339–361 (1995)

[10] Muratbekova, M., Shamoi, P.: Color-emotion associations in art: Fuzzy approach. IEEE Access **12**, 37937–37956 (2024)

[11] Plutchik, R.: Emotion: Theory, research, and experience. Volume 1: Theories of emotion. Academic Press, New York (1980)

[12] Quattoni, A., Torralba, A.: Recognizing indoor scenes. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 413–420. IEEE (2009)

[13] Russell, J.A.: A circumplex model of affect. Journal of Personality and Social Psychology **39**(6), 1161–1178 (1980)

[14] Singh, R., Saurav, S., Kumar, T., Saini, R., Vohra, A., Singh, S.: Facial expression recognition in videos using hybrid cnn & convlstm. International Journal of Information Technology **15**(4), 1819–1830 (2023)

[15] Udahemuka, G., Djouani, K., Kurien, A.M.: Multimodal emotion recognition using visual, vocal and physiological signals: a review. Applied Sciences **14**(17), 8071 (2024)

[16] Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition. pp. 3485–3492. IEEE (2010)
[17] Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(6), 1452–1464 (2017)

Metaphor-Aware Sentiment Analysis in Multi-Turn Conversations

Guo Wei[1] and Mohd Nor Akmal Khalid[1][0000−0002−7909−8869]

Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia
p147206@siswa.ukm.edu.my; akmal@ukm.edu.my

Abstract. Sentiment analysis in conversational contexts presents significant challenges, particularly when figurative language such as metaphors is involved. Traditional models often underperform due to their inability to capture the nuanced emotional subtext embedded in metaphorical expressions and multi-turn dialogues. This paper proposes an integrated deep learning framework that leverages transformer-based architectures to jointly model conversational context and metaphor usage. We introduce a metaphor-aware sentiment classification system incorporating memory-augmented modules and attention-based metaphor detectors. A novel Contextual Comprehension Score (CCS) is proposed to evaluate the model's ability to interpret contextual and figurative sentiment. The system is validated on a composite dataset comprising DailyDialog, Reddit threads, and metaphor-annotated corpora. Experimental results demonstrate substantial improvements over baseline models, such as BERT and DialogRNN, particularly in scenarios involving emotional nuance and dialogue complexity. This study lays the groundwork for more interpretable, context-sensitive sentiment analysis systems, suitable for real-world applications such as mental health assessment and social media monitoring.

Keywords: Sentiment Analysis · Metaphor Detection · Conversational AI · Transformer Models · Contextual Modeling

1 Introduction

The interpretation of sentiment in dialogue-rich text has gained growing importance with the proliferation of user-generated content across social media and online platforms. Sentiment analysis can be defined as the computational determination of affective states in textual input that traditionally relies on detecting explicit lexical markers such as "happy" or "sad" [8]. However, conversational data often includes nuanced emotional content expressed through figurative language, particularly metaphors, which complicate sentiment inference [5].

Metaphorical expressions such as "carrying the weight of the world" or "walking on thin ice" encode emotional states that are context-dependent and deviate from literal meanings [7]. In parallel, sentiment conveyed through dialogue

unfolds across multiple speaker turns, influenced by history, tone, and conversational flow [13].

State-of-the-art models, including those based on transformers such as BERT [4], have improved performance in structured settings, yet struggle with the dynamic, non-literal, and interaction-dependent nature of human conversations [1]. Moreover, sentiment and metaphor understanding are typically treated as separate tasks, leading to disjointed interpretations [12].

This paper proposes a unified deep learning framework that addresses this gap by jointly modeling conversational sentiment and metaphor comprehension. The key contributions of this work are as follows:

- We design an integrated architecture leveraging a transformer-based encoder with dedicated modules for metaphor detection and conversational context modeling.
- We construct a composite dataset from DailyDialog, Reddit, Twitter, and the VUA Metaphor Corpus, annotated with both sentiment and metaphor labels.
- We adopted the Contextual Comprehension Score (CCS) as a novel metric to evaluate the model's ability to capture sentiment modulated by metaphor and conversational flow.
- We demonstrate through empirical analysis that our framework significantly outperforms baselines such as BERT and DialogRNN in metaphor-rich, multi-turn sentiment classification.

2 Related Work

Early sentiment analysis approaches relied on lexical resources or shallow machine learning models such as Support Vector Machines (SVM) and Naive Bayes, which performed well on structured, single-turn text [10]. However, these methods failed to capture contextual and pragmatic features necessary for analyzing dialogue.

Recent advances in deep learning introduced recurrent neural networks (RNNs), such as DialogRNN [13], and graph-based approaches like DialogueGCN [9], which incorporated turn-level and speaker-dependent information to track sentiment evolution in conversations. Despite these improvements, figurative language, particularly metaphors, remained an unsolved challenge.

Metaphor detection has traditionally been addressed as a separate NLP task. Rule-based and statistical approaches provided early foundations, but recent neural models, including Neural Metaphor Detection [6] and transformer-based approaches using BERT [12], have shown improved performance. Yet these methods typically operate on isolated sentences and lack integration with conversational structure.

Some joint learning architectures like MelBERT [3] and Dual Graph-Transformer Networks [2] have attempted to combine sentiment and metaphor understanding. However, their applications remain limited to static or sentence-level contexts, and do not model sentiment flow across multi-turn conversations.

Our work addresses these gaps by proposing an integrated framework that combines metaphor detection with contextual sentiment modeling, optimized for dynamic, multi-turn dialogue.

3 Proposed Method

Our proposed model integrates three key components (Figure 1): a transformer-based encoder, a metaphor detection module, and a contextual sentiment module. These components work in tandem to process multi-turn dialogues enriched with metaphorical expressions.

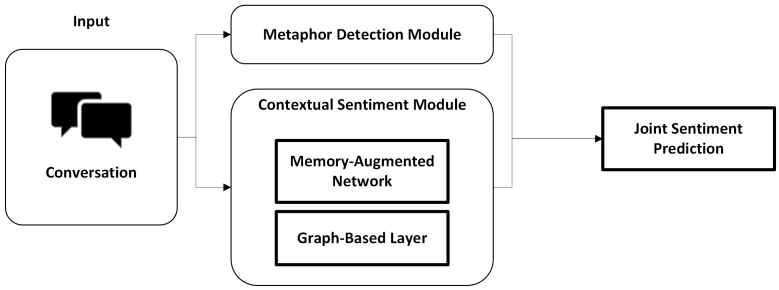


Fig. 1. Overall architecture of the proposed metaphor-aware sentiment analysis model

3.1 Transformer-Based Encoder

At the core of our framework lies a BERT-based encoder [4] that transforms tokenized utterances into contextualized embeddings. For each turn in a dialogue, we prepend a [CLS] token and use segment embeddings to distinguish speaker turns. This facilitates multi-turn encoding within a single transformer window. BERT's bidirectional self-attention mechanism captures long-range dependencies and syntactic-semantic features essential for downstream sentiment and metaphor tasks.

3.2 Metaphor Detection Module

Metaphor identification is conducted using a semantic dissimilarity mechanism adapted from [6]. Each token's contextual representation from BERT is compared to a static embedding baseline (e.g., GloVe), and cosine dissimilarity is computed to detect deviation from literal meaning. We enhance this process using an attention mechanism to highlight metaphorically salient tokens, drawing on insights from Conceptual Metaphor Theory [7]. The output is passed through a binary classifier to predict metaphor presence.

3.3 Contextual Sentiment Module

The graph-based sentiment module represents each utterance as a node, with edges constructed based on speaker turns and temporal proximity, allowing the model to track affective flow across conversational structure. This design captures inter-speaker dependencies and emotional influence across turns.

The hierarchical attention mechanism operates at two levels: (i) token-level attention identifies salient words or metaphorical cues within each utterance, and (ii) turn-level attention determines which dialogue turns contribute most to the overall sentiment. This two-layered approach ensures that both local and global contextual dependencies are incorporated into the sentiment inference process. The sentiment module captures dialogue dynamics by modeling inter-turn dependencies. We integrate:

- **Memory Networks** [13]: Retain speaker and sentiment states across turns.
- **Graph-based modeling** [9]: Represent utterances as nodes and define edges based on speaker transitions and dialogue proximity.
- **Hierarchical Attention** [1]: Assign importance at both token-level and turn-level using learned weights.

This design allows the system to adapt to emotional context shifts and identify subtleties such as sarcasm and tonal variation.

3.4 Training and Optimization

We adopt a multi-task learning approach, following [3], which combines loss functions for both sentiment and metaphor tasks, as given by Eq. (1), where $\lambda_1 = \lambda_2 = 1$ in our experiments. Data augmentation methods (paraphrasing, synonym substitution) improve robustness in metaphor-scarce scenarios.

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{sentiment}} + \lambda_2 \mathcal{L}_{\text{metaphor}} \tag{1}$$

3.5 Evaluation Metric via Contextual Comprehension Score (CCS)

The Contextual Comprehension Score (CCS) measures a model's ability to accurately interpret sentiment, taking into account both the contextual dialogue flow and the presence of metaphors. Formally, as shown in Eq. (2), CCS measures conditional accuracy—that is, the proportion of correctly predicted sentiment labels given the dual conditioning of context and metaphor annotations. This formulation emphasizes interpretive alignment rather than label agreement alone, making CCS more sensitive to figurative or context-dependent sentiment shifts than metrics such as F1 or accuracy.

$$\text{CCS} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left(\hat{y}_i = y_i \mid \text{context}_i, \text{metaphor}_i\right) \tag{2}$$

Unlike traditional metrics such as accuracy or F1-score, which assess label agreement in isolation, the CCS evaluates the extent to which the model's sentiment predictions align with human-annotated interpretations under contextual and metaphorical influences.

If a model correctly predicts the sentiment of an utterance like "I'm walking on thin ice" as negative because it interprets the metaphorical caution implied, CCS registers a higher contextual alignment than models that classify it correctly but for unrelated reasons. Thus, CCS emphasizes comprehension over coincidence, providing a more interpretable measure for figurative sentiment modeling.

4 Experiments and Results

This section presents the datasets, implementation setup, baseline comparisons, and empirical evaluations conducted to assess the effectiveness of our metaphor-aware sentiment classification framework.

4.1 Datasets and Preprocessing

We compiled a composite dataset that integrates DailyDialog, Reddit threads, and the VUA Metaphor Corpus [11]. DailyDialog and Reddit contribute multi-turn conversational data with colloquial and informal expressions, while the VUA corpus provides high-quality metaphor annotations based on the Metaphor Identification Procedure (MIP).

To illustrate the structure of our composite dataset, Table 1 presents representative samples from DailyDialog, Reddit, and the VUA Metaphor Corpus. Each utterance is annotated with both sentiment and metaphor labels.

Table 1. Annotated Examples from Composite Dataset

Turn	Utterance	Sentiment	Metaphor
A1	I'm totally burned out.	Negative	Yes
B1	Take a breather. You've been going full throttle.	Neutral	Yes
A2	Deadlines are breathing down my neck.	Negative	Yes
U1	That promotion slipped through my fingers again.	Negative	Yes
U2	Maybe it just wasn't meant to be.	Neutral	No
U3	I feel like I'm stuck in a loop of disappointment.	Negative	Yes
S1	The economy is on the verge of collapse.	Negative	Yes
S2	Policymakers are walking a tightrope.	Neutral	Yes

These examples illustrate the diversity and complexity of sentiment expression, particularly in metaphor-rich utterances. The annotations enable the model to learn how figurative language influences emotional meaning in context. The metaphor ratio in each dataset represents the percentage of utterances containing at least one metaphorical token or phrase, identified using the Metaphor Identification Procedure (MIP) [11]. This definition ensures consistent comparison across conversational and metaphor-annotated sources. Table 2 summarizes the dataset characteristics. Sentiment labels were annotated using weak supervision and verified by expert annotators, yielding strong inter-annotator agreement ($\kappa = 0.82$).

Table 2. Summary of Datasets Used

Dataset	Samples	Avg. Turns	Metaphor Ratio
DailyDialog	13,118	7.9	21%
Reddit Threads	4,000	5.2	33%
VUA Corpus	2,543	–	100%

The final composite dataset comprised 19,661 conversational samples drawn from DailyDialog (13,118 dialogues), Reddit Threads (4,000 dialogues), and the VUA Metaphor Corpus (2,543 sentences). Across these sources, approximately 12,300 utterances contained metaphorical expressions, yielding a metaphor-aware ratio of 29%. Each utterance was tokenized and encoded using a BERT-base model to generate 768-dimensional contextual embeddings. For metaphor detection, additional features were computed, including cosine dissimilarity between contextual and static GloVe embeddings and attention-derived saliency weights emphasizing metaphorical tokens. The resulting combined representation for each token consisted of 770 features, serving as input to the binary metaphor classifier.

While the composite dataset combines DailyDialog, Reddit, and the VUA Metaphor Corpus to ensure coverage of colloquial and formal metaphors, it may not fully represent culture-specific or domain-restricted metaphors, such as idioms unique to certain linguistic communities or topical domains (e.g., sports, politics). Future work will explore cross-cultural and multilingual extensions to address these representational gaps.

4.2 Experimental Setup

Our experiments were implemented using PyTorch 2.0 and HuggingFace Transformers on a system equipped with an NVIDIA A100 GPU and 512 GB RAM. Table 3 outlines the software and hardware configurations used for training and evaluation.

Table 3. Experimental Setup

Component	Specification
GPU	NVIDIA A100 (80GB VRAM)
CPU	Intel Xeon Gold 6248 @ 2.50GHz
Memory	512 GB RAM
OS	Ubuntu 20.04 LTS
Frameworks	PyTorch 2.0, HuggingFace Transformers
Tokenizer	WordPiece (BERT-base uncased)

4.3 Training Configuration

We trained our model using the AdamW optimizer with learning rate 2×10^{-5} and dropout 0.3. The batch size was set to 32, and training proceeded for 10 epochs. Table 4 summarizes the training hyperparameters.

Table 4. Training Hyperparameters

Hyperparameter	Value
Batch size	32
Epochs	10
Optimizer	AdamW
Learning rate	2e-5
Dropout	0.3
Max sequence length	128
Evaluation frequency	Every epoch

4.4 Baselines

We compared our framework against three strong baselines:

- **BERT** [4]: A pretrained transformer model fine-tuned on sentiment labels only.
- **DialogRNN** [13]: An RNN-based model for conversational sentiment tracking.
- **DialogueGCN** [9]: A graph convolutional model leveraging speaker and context dynamics.

These serve as robust comparators to assess the added value of metaphor-awareness and memory-augmented modeling.

4.5 Quantitative Results

Table 5 reports the performance on the test set. Our model achieves the highest accuracy (0.89), F1-score (0.87), and Contextual Comprehension Score ($CCS = 0.76$). The improvements underscore the contribution of metaphor modeling in capturing emotional subtext.

Table 5. Performance Comparison on Test Set

Model	Accuracy	F1-score	CCS
BERT	0.81	0.78	0.63
DialogRNN	0.84	0.81	0.67
Ours (Unified)	**0.89**	**0.87**	**0.76**

4.6 Ablation Study

To analyze individual component impact, we ablated core modules sequentially. As shown in Table 6, removing metaphor detection significantly lowered F1 and CCS. Excluding memory/graph context further degraded performance, confirming the necessity of both metaphor and dialogue modeling.

Table 6. Ablation Study Results

Variant	F1-score	CCS
Full Model	0.87	0.76
– Metaphor Detection	0.81	0.70
– Memory/Graph Module	0.79	0.65
– Both Removed	0.75	0.59

4.7 Visualization and Analysis

Visual inspection provide qualitative insight into the model's behavior across key components. For instance, Figure 2 presents the confusion matrix for binary metaphor classification. The model effectively distinguishes metaphorical from literal utterances, with low false positives. Errors are mostly tied to contextually embedded metaphors or cultural idioms.

Figure 3 shows the confusion matrix for three-class sentiment classification. The model performs best on positive samples but exhibits some misclassification between neutral and negative classes, indicating sentiment ambiguity in real-world dialogue.

Figure 4 provides a self-attention heatmap across five query and seven key positions. The focus of attention is on proximal utterances, reinforcing the importance of short-term context in affective inference.

Finally, Figure 5 depicts the sentiment trajectory across seven dialogue turns. The model's predictions align closely with human-labeled sentiment shifts, with minor divergence at emotional turning points, confirming its capacity for context-sensitive emotional tracking.

Further inspection reveals two recurring sources of error. First, sentiment ambiguity often arises between neutral and negative utterances, particularly when the emotional state is implied rather than explicit—for instance, "It's fine, I guess" was often misclassified as neutral despite its negative undertone.

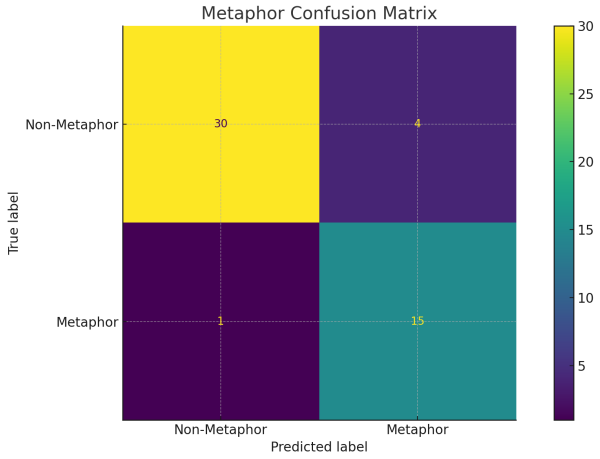


Fig. 2. Confusion matrix for metaphor classification (N = 2,543 samples from the VUA corpus). Axes represent true versus predicted metaphor labels; the color scale indicates the sample count per cell.

Second, culturally specific metaphors (e.g., “walking on lotus leaves” meaning “facing a delicate situation”) were occasionally misinterpreted due to limited exposure in the training data. These examples highlight both the interpretive boundary of metaphor awareness and the importance of including broader cultural and linguistic diversity in future corpora.

5 Discussion and Future Work

Our results suggest that metaphor modeling enhances interpretability in sentiment analysis tasks, especially in real-world conversational data such as Reddit and DailyDialog. The CCS metric offers a new perspective in evaluating sentiment systems under human-like comprehension criteria.

5.1 Case Example and Comparative Evaluation

To demonstrate the practical advantages of our framework, we include a representative example from the DailyDialog dataset:

In this dialogue, the speaker’s first turn (A1) includes the metaphor "dragging a boulder uphill," which conveys emotional exhaustion. Ground truth annotations label A1 as *negative* (metaphorical), B1 as *neutral*, and A2 as *negative*.

Table 8 presents model predictions:

Both BERT and DialogRNN fail to interpret the metaphor in A1, classifying the utterance as neutral. In contrast, our model correctly detects the figurative

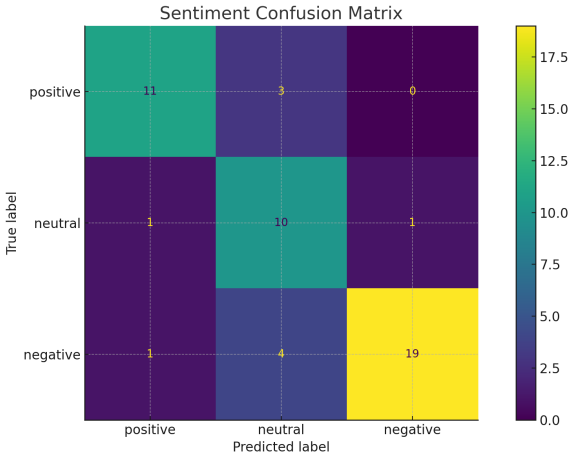


Fig. 3. Confusion matrix for three-class sentiment classification (N = 4,000 Reddit dialogues). Axes represent true vs. predicted sentiment classes (Positive, Neutral, Negative).

Table 7. Example for dialog conversations

Conversation Excerpt
A: "I can't keep this up. It's like dragging a boulder uphill every day."
B: "That's tough. Have you thought about taking a break?"
A: "No time for that. Deadlines don't care about feelings."

expression and aligns sentiment with human annotations. Such examples underscore the advantage of metaphor comprehension in conversational sentiment analysis.

This pattern held consistently across 100 similar metaphor-rich dialogues, where our model improved CCS by 12–15% relative to the strongest baseline. Nevertheless, challenges remain in handling culturally dependent metaphors and evolving expressions in online communication. While our dataset blends general and metaphor-specific corpora, some domain-specific metaphor instances were still misclassified.

In future work, we aim to:

- Extend the framework to multilingual settings, focusing on metaphor translation consistency.
- Integrate prosodic and visual modalities for richer context modeling.
- Apply CCS in mental health screening applications to evaluate empathetic response quality.

Our unified approach marks a step toward emotionally intelligent AI systems capable of deeper language understanding.

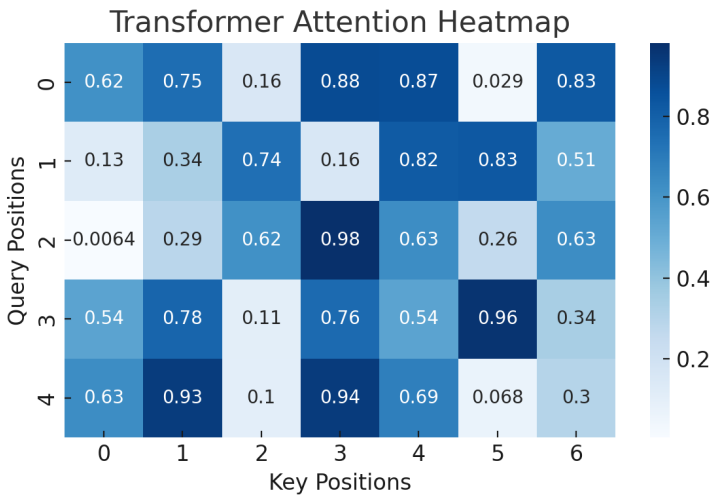


Fig. 4. Transformer Encoder Attention Heatmap (N = 100 sampled dialogues; intensity normalized).

Table 8. Model Predictions on a Sample Conversation

Turn	Ground Truth	BERT	DialogRNN	Proposed
A1	Negative (Met.)	Neutral ×	Neutral ×	Negative ✓
B1	Neutral	Neutral ✓	Neutral ✓	Neutral ✓
A2	Negative	Negative ✓	Negative ✓	Negative ✓

6 Conclusion

In this paper, we presented a metaphor-aware sentiment analysis framework designed for multi-turn conversations. Our model integrates transformer-based encoding, metaphor detection via semantic dissimilarity, and contextual sentiment modeling using memory and graph-based networks. A novel Contextual Comprehension Score (CCS) was introduced to evaluate the model's interpretability and performance under figurative and contextual constraints.

Experimental results demonstrated that incorporating metaphor understanding significantly improves sentiment classification accuracy, particularly in emotionally nuanced dialogues. The ablation studies further confirm that both metaphor and memory-context components contribute independently and synergistically to model performance.

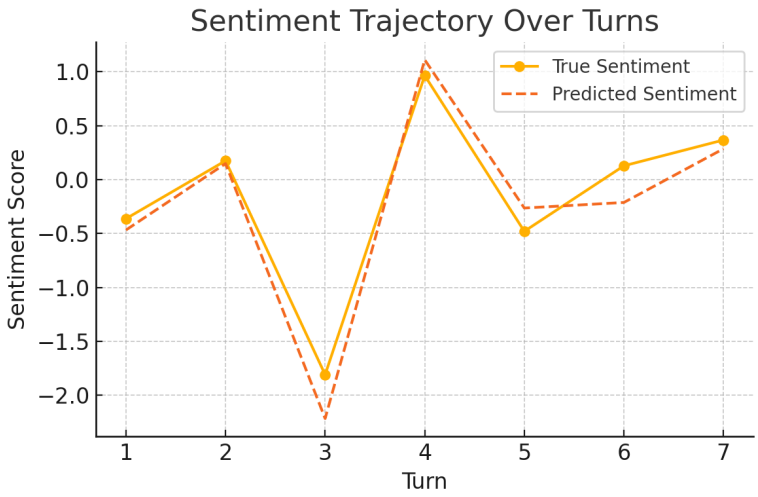


Fig. 5. Predicted vs. ground-truth sentiment trajectories across seven turns (N = 200 dialogues; CCS trajectories averaged across seeds), illustrating the model's ability to track emotional flow.

Bibliography

[1] Cambria, E., Schuller, B., Xia, Y., Havasi, C.: Sentiment analysis is a big business: Perspectives and challenges. Cognitive Computation **9**(4), 565–583 (2017)

[2] Chen, L., Xu, R.: Dual graph-transformer network for joint conversational sentiment and metaphor detection. ACL Findings (2023)

[3] Choi, J., Lee, K.: Melbert: Metaphor detection via contextualized late interaction using metaphorical identification theories. EMNLP (2021)

[4] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. NAACL (2019)

[5] Fainsilber, L., Ortony, A.: Metaphorical uses of language in the expression of emotions. Metaphor and Symbol **2**(4), 239–250 (1987)

[6] Gao, Q., Choi, E., Wiebe, J.: Neural metaphor detection in context. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing pp. 1880–1890 (2018)

[7] Lakoff, G., Johnson, M.: Metaphors We Live By. University of Chicago Press (1980)

[8] Liu, B.: Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies **5**(1), 1–167 (2012)

[9] Majumder, N., Hong, P., Poria, S.: Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. AAAI (2020)

[10] Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: Sentiment classification using machine learning techniques pp. 79–86 (2002)

[11] Shutova, E.: Models of metaphor in nlp. ACL (2010)

[12] Su, H., Xue, W., et al.: Metaphor detection using pretrained language models with semantic probes. ACL (2021)

[13] Zhang, P., Lan, C., Li, Z.: Dialogrnn: An rnn-based framework for emotion detection in conversations. EMNLP (2018)

A Simplified Multi-Floor Classification-Based Indoor Positioning System Study

Burin Intachuen[1], Mhadhanagul Charoenphon[1], Tanakorn Mankhetwit[1], and Charnon Pattiyanon[2][0000−0003−3660−2962]

[1] Department of Computer Engineering, Mahidol University (International College), Nakhon Pathom, Thailand
{rinrin.int, dallas_char}@proton.me, tanman.1170@gmail.com
[2] Department of Artificial Intelligence and Computer Engineering, CMKL University, Bangkok, Thailand
charnon@cmkl.ac.th

Abstract. Indoor Positioning Systems (IPS) aim to supplement or replace Global Navigation Satellite Systems (GNSS) for indoor localization, yet limited research has addressed IPS optimization in multi-floor environments. This study investigates the overall performance of machine learning-based fingerprinting models under varying environmental configurations, with a focus on two key factors: grid size for fingerprint data collection and the impact of low-relevance Basic Service Set Identifiers (BSSIDs). Two new evaluation metrics, Average Grid from Target (AGT) and Average Distance from Target (ADT), are introduced to provide standardized measures of positioning accuracy across different grid layouts. The paper also aims to bring forth a unique approach to model training through the use of interpolation to create different grid sizes from existing grids. A BSSID filtering approach was applied, excluding low-intensity signals likely to be irrelevant to localization, reducing computational load and training time. Experiments were conducted in the hallways of a university learning center across two floors, with an initial 1×1 m grid size. The dataset comprised 12,640 RSSI samples collected from 378 filtered access points over approximately 500 grid cells. Larger grid sizes were synthesized via interpolation to evaluate performance trade-offs. Results show that while accuracy generally increases with grid size, a 7×7 m configuration achieves the optimal balance, yielding 80% accuracy alongside the lowest AGT and ADT values. Random Forest and XGBoost consistently outperformed other models, indicating their suitability for multi-floor IPS tasks in similar environments.

Keywords: Indoor Positioning System (IPS) · Wi-Fi Signal Processing · Machine Learning · Multi-Floor Positioning

1 Introduction

Indoor Positioning Systems (IPS) are designed to help users determine their location and navigate effectively within buildings or enclosed areas. Traditional

positioning and navigation methods, such as the Global Navigation Satellite System (GNSS), rely on signals from satellites and ground control stations to calculate positions using trilateration. The most well-known GNSS is the Global Positioning System (GPS). However, GPS performance degrades significantly indoors because most buildings are constructed from dense materials like concrete and metal, which block or weaken the low-power satellite signals. This results in scattering, shadowing, blind spots, and signal attenuation, all of which reduce positioning accuracy [11]. To overcome these limitations, various IPS methods have been developed to provide accurate navigation in environments where GNSS is unreliable or unavailable.

Many methods have been developed for Indoor Positioning Systems (IPS), with one of the most well-known being trilateration. This technique measures the signal strength from transmitters, such as Bluetooth Low Energy (BLE), and uses the Received Signal Strength Indicator (RSSI) as a proxy for distance [3]. The process typically involves an offline phase to create an RSSI "radio map" (fingerprinting), followed by an online phase to estimate the user's position based on those measurements. However, a major drawback is that RSSI measurements are susceptible to environmental noise and require the precise placement of access points, which ultimately limits their accuracy [14].

Machine learning approaches have proven effective for IPS performance improvement [6]. Classification algorithms, e.g., Support Vector Machine (SVM), k-Nearest Neighbor (kNN), Random Forests, and Neural Networks, address RSSI limitations by treating positioning as a classification problem, providing reliable area-based location estimates.

Previous research [7] implemented an IPS using a large grid size (16.75×15 m) to reduce the number of classification labels, making the problem more manageable within time constraints. While this approach provided a functional implementation, it left several questions unanswered regarding data point influence on IPS performance, the feasibility of implementing reliable IPS with limited Basic Service Set Identifier (BSSID) features, and the minimum viable grid size for maintaining positioning accuracy. Building on this foundation, this paper further expands classification-based IPS by refining the approach to ground truth reliability in experimental settings. The main contributions of this paper are:

1. Introduction of new evaluation metrics: Average Grid from Target (AGT) and Average Distance from Target (ADT).
2. Investigation of trade-offs between grid size and precision, analyzing implementation advantages and limitations.
3. Presentation of deeper insights into IPS design considerations, offering practical improvements for indoor positioning accuracy.
4. Examination of feature filtering effects on model complexity.

The remainder of this paper is organized as follows: Section 2 presents a literature review to provide an overview of the research landscape; Section 3 describes the research methodology; Section 4 reports the experimental results; Section 5 discusses the technical findings; and Section 6 concludes the study with future directions.

2 Literature Review

With the aforementioned limitations of GNSS and GPS in indoor environments, research on IPS has gained significant momentum in recent years. Consequently, IPS has been extensively studied and developed using various signal processing techniques [12]. One prominent approach is Wi-Fi-based RSSI fingerprinting, enhanced by machine learning (ML) and deep learning methods.

Traditional ML algorithms have been employed to classify fingerprints and estimate the current position from a set of RSSI values. Commonly used algorithms include k-Nearest Neighbor (kNN) [7, 15, 16], Random Forest [7, 9, 16], Support Vector Machine (SVM) [4, 7, 15, 16], and Multi-Layer Perceptron (MLP) [7, 15]. These models offer straightforward implementation and produce interpretable results. Their relatively simple architectures and modest computational requirements make them particularly suitable for IPS applications, as they can be deployed without specialized hardware. The accuracy of an IPS largely depends on the quality of the collected data, which can be influenced by factors such as diverse RSSI readings at different times [2], the integration of human activity recognition [1], and the use of coordinated data collection applications [10].

Deep learning approaches, such as Graph Neural Networks (GNNs) [15] and Convolutional Neural Networks (CNNs) [1], have demonstrated superior performance over traditional ML algorithms when sufficient data points are available. However, despite these advancements, key environmental variables, such as the resolution of fingerprint grids and the spatial distribution of reference points, have not been comprehensively investigated. This study emphasizes the critical role of these factors in IPS performance and systematically analyze the most effective configurations for multi-floor environments. By addressing these overlooked variables, This study aims to bridge existing research gaps and enhance the accuracy, reliability, and overall performance of IPS through optimal design and configuration.

3 Research Methodology

This paper adopts a systematic and empirical research methodology with the aim of identifying environmental factors that influence the IPS. This methodology is illustrated in Figure 1 and consists of four parts: (1) Environmental factor identification, (2) Data extraction, (3) Model training and inference, and (4) Result analysis. Each part will be explained in detail in the following sections.

3.1 Identification of Environment Factors

Environmental factors refer to any physical or logical characteristics of the area used by IPS developers to generate an RSSI-based radio map. In a typical offline map generation phase, the target area is marked and divided into grids that correspond to its digital floor plan and then taped to physically map that area. Each grid is assigned a unique reference number. Data collection is carried out

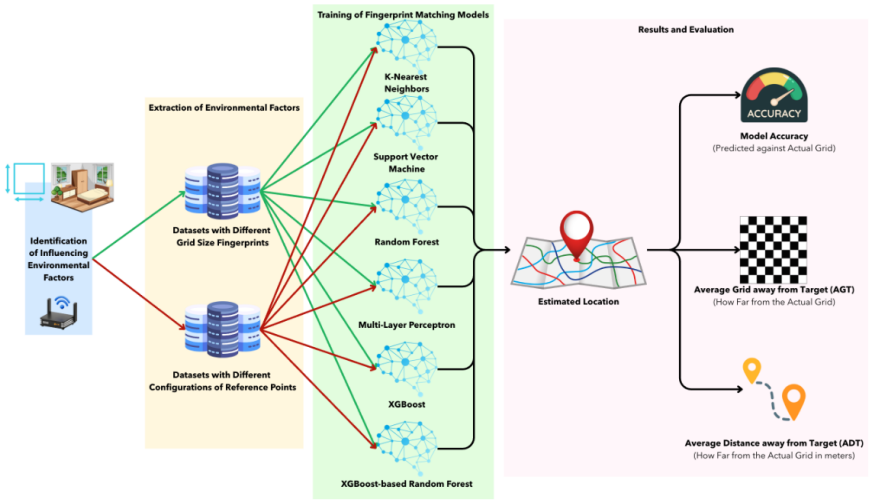


Fig. 1. An Overview of the Research Methodology.

using devices [8, 13] capable of measuring RSSI values from various reference points, such as BLE beacons or Wi-Fi access points, at each grid location. The resulting set of RSSI values serves as the feature vector representing each grid (or sub-grid) in the radio map. Sometimes, data synthesis is also used to add more data points from the collected fingerprints [16].

From the analysis of the offline map generation process, two key factors emerge as potentially significant to IPS performance. Firstly, grid size directly impacts the precision of position estimation: larger grids make it easier to classify a position into the correct grid, but smaller grids are required for higher positioning accuracy. However, reducing grid size substantially increases the labor required for data collection, rendering map generation impractical for large-scale environments. Secondly, RSSI collection is not strictly limited to reference points (e.g., access points) within the target area; signals from access points in neighboring buildings may also be recorded, introducing low-relevance data as shown in Figure 2. The x-axis represents the BSSID of different access points, with darker colors indicating lower received signal strength in decibel-meters (dBm). This study focuses on these two factors as the scope of investigation.

3.2 Extraction of Environmental Factors and Datasets

The two factors identified in the previous section can be tuned, and each is expected to have an optimal balance between practicality and performance. In this section, the study extracts a quantitative representation of each factor and prepares the dataset for model training and inference accordingly.

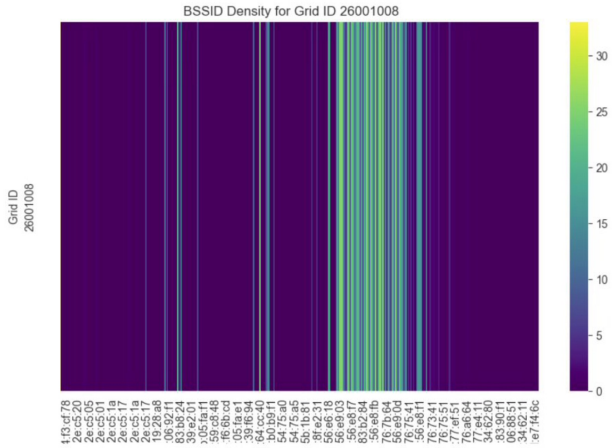


Fig. 2. A Heatmap of RSSI values collected from different access points.

For the grid size factor, a quantitative representation is derived using distance measurements. Traditionally, positioning evaluations rely on precise global coordinates (e.g., latitude/longitude) as benchmarks. However, such measurements are impractical to obtain indoors without professional surveying equipment. To address this, two new metrics based on the Euclidean distance concept are introduced: Average Grid from Target (AGT) and Average Distance from Target (ADT). AGT can be computed using the following equation:

$$AGT = \sqrt{(x_{target} - x_{estimate})^2 + (y_{target} - y_{estimate})^2} \tag{1}$$

where x_{target} and y_{target} are the indices of the target grid on the horizontal and vertical axes, respectively, and $x_{estimate}$ and $y_{estimate}$ are the indices of the estimated grid predicted by the IPS. For example, $x_{target} = 3$ and $y_{target} = 5$ indicate that the target grid is the third grid from the left and the fifth grid from the top of the map. By treating grid indices as distance units, this metric naturally adapts to different grid sizes providing both a straightforward means to visualize positioning performance and a standardized measure for cross-comparison across varying grid sizes.

On the other hand, ADT is a related metric that quantifies the average Euclidean distance, measured in actual physical units such as meters, between the estimated position and the ground-truth target position. It can be computed using the following equation:

$$ADT = \sqrt{[w_g \times (x_{target} - x_{estimate})]^2 + [h_g \times (y_{target} - y_{estimate})]^2} \tag{2}$$

where w_g and h_g denote the global grid width and height in meters, respectively. Unlike AGT, which uses grid indices as units, ADT reflects the absolute positioning error in real physical space, making it more interpretable for practical

deployment. Notably, larger grid sizes inherently lead to greater distances between the target and the estimated positions, even when the grid index difference remains the same.

On the other hand, quantifying the low-relevance data factor is more straightforward. The performance of an IPS is inversely proportional to the amount of low-relevance data, meaning that the more low-relevance data present, the less accurate the IPS becomes. To identify such data points, a simple threshold-based approach is applied. From the analysis of the collected RSSI heat map in Figure 2, it is evident that the majority of low-relevance BSSIDs has its RSSI values fall within the range of 0 to 5 decibel-meters. Based on this observation, a threshold of 5 decibel-meters is proposed to filter out data points that are likely to negatively impact IPS performance.

With the metrics above, the effect of different grid sizes can be quantified in a tangible way. Ideally, data points for various grid sizes should be collected to study the influence of this factor. However, collecting new datasets for each grid size is highly impractical.

In typical IPS experiments, floor taping is used to physically mark grid boundaries, and repeatedly re-taping the same area is extremely time-consuming and prone to human measurement errors.

To address this limitation, an interpolation-based approach is proposed, where larger grids are generated by aggregating smaller ones. For example, four 1×1 m grids can be combined into a single 2×2 m grid, as illustrated in Figure 3. In each aggregated grid, RSSI values from five sampling points, i.e., top-left, top-right, center, bottom-left, and bottom-right, are interpolated using a standard averaging method.

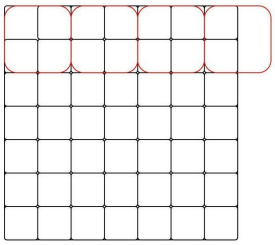


Fig. 3. Visualization of grid aggregation by interpolating RSSI values.

A major drawback of the grid interpolation process is the reduction in the number of data points. However, from multiple rounds of data collection in this study, 9,019 and 3,621 data points were obtained from the first- and second-floor hallways, respectively. This is considered a sufficient number of data points to maintain a high-quality dataset. The difference in the number of data points is due to variations in the floor plan and room partitioning of the test area. Table 1 summarizes the metadata and statistics of the dataset collected in this study.

Table 1. Metadata of the number of grids and data points for each grid size.

Floor		Grid Sizes							
		1 × 1	**3 × 3**	**5 × 5**	**7 × 7**	**9 × 9**	**11 × 11**	**13 × 13**	**15 × 15**
1st	**# Grid**	309	55	28	17	17	10	10	5
	# Data Point	8,086	1,439	733	445	399	262	235	130
2nd	**# Grid**	174	41	19	12	11	6	8	4
	# Data Point	4,554	1,073	497	314	288	157	209	105

Moreover, the data points in the datasets were analyzed to filter out low-relevance features. These features correspond to access points (referred to as BSSIDs), which are a set of RSSI values. Out of the 1,799 BSSIDs initially detected by the data collection device, 1,421 were filtered out as being low-relevance. At this stage, the datasets are ready for the model training process.

3.3 Model Training

To study the impact of the identified environmental factors, this paper trains various machine learning models to estimate the current position in terms of the predefined grid layout.

This paper uses four traditional machine learning algorithms (kNN, Random Forest, MLP, and SVM). Additionally, two tree-based ensemble models were selected: XGBoost [5] and a hybrid model combining XGBoost + Random Forest. Each model was trained through a standard process, with hyperparameter tuning performed to achieve optimal performance.

The selection of algorithms was based on an analysis of the dataset. It was found that each data point contains four main features—the current floor, grid size, current grid index, and a set of corresponding RSSI values—and that most of them are well-structured, sparse, numeric, and ordinal. Therefore, the chosen algorithms needed to be capable of handling and analyzing this type of data.

4 Experiments and Results

In this section, experiments were conducted to evaluate the performance of various ML algorithms and models on the IPS task. These models were trained as described in Section 3.3. First, overall performance was evaluated of the models using key metrics such as accuracy, AGT, and ADT. Afterward, the changes in accuracy were analyzed and AGT that resulted from applying BSSID filtering.

4.1 Accuracy, AGT, and ADT Results

Every ML model is trained with the same controlled dataset and workstation. Accuracy, AGT, and ADT are then calculated from the predicted test results compared to the ground truth in the dataset. Graphs were plotted to visualize

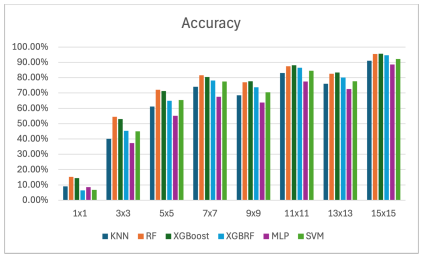


Fig. 4. Model accuracy scores on different grid sizes (Filtered)

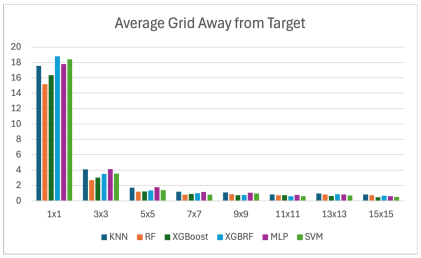


Fig. 5. Model AGT on different grid sizes (Filtered)

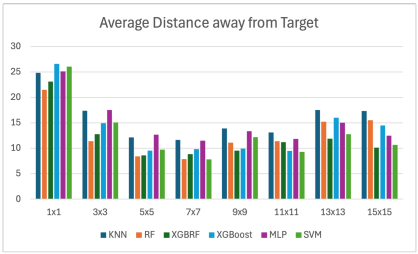


Fig. 6. Model ADT on different grid sizes (Filtered)

the comparison of results between different models and grid sizes based on the accuracy, AGT, and ADT, as shown in Figures 4, 5, and 6.

Figure 4 presents the position estimation accuracy of each model across different grid sizes. The results indicate that larger grid sizes generally yield higher accuracy. However, an optimal configuration was identified where the smallest grid size that still achieves high accuracy (approximately 80%) is 7×7 m. Additionally, Random Forest (RF) and XGBoost consistently deliver the best performance for position estimation across all grid sizes.

For AGT and ADT in Figures 5 and 6, the results are similar to those for accuracy. Figure 5 shows that larger grids have lower AGT, meaning it is easier to classify the correct grid during position estimation. The AGT value begins to stabilize once the grid size reaches 7×7 m, suggesting that this is the smallest grid size capable of maintaining minimal errors. On the other hand, Figure 6 indicates that a 7×7 m grid size minimizes error while remaining small enough for precision-dependent applications in this specific setting. The ADT results confirm that, even when measured in physical units, the 7×7 m grid size remains the optimal point.

4.2 Filtering Impact on Model Performance

As described in Section 3.2, BSSID filtering is proposed to exclude low-relevance data from the dataset. This approach can reduce computational resource requirements, particularly in scenarios with a large feature space (i.e., a high number

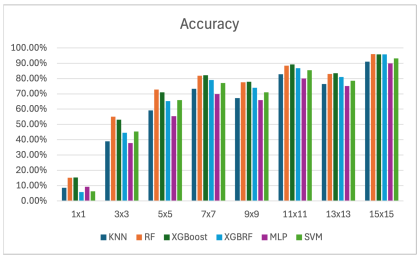


Fig. 7. Model accuracy scores on different grid sizes (Unfiltered)

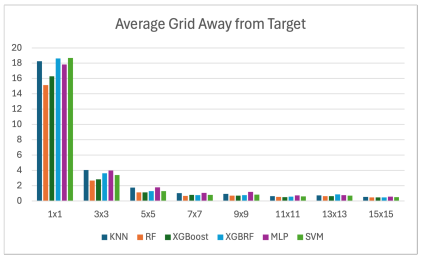


Fig. 8. Model AGTs on different grid sizes (Unfiltered)

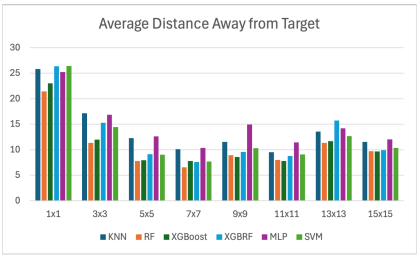


Fig. 9. Model ADTs on different grid sizes (Unfiltered)

of BSSIDs within the area network). To evaluate whether this filtering process can improve IPS performance, we conducted experiments comparing all models trained on datasets generated with and without filtering across all grid sizes.

Comparing Figures 4 and 7 reveals the impact of BSSID filtering on model accuracy across all models and grid sizes. The filtered dataset (Figure 4) achieves comparable accuracy to the unfiltered dataset (Figure 7) while using significantly fewer features. More importantly, the ADT comparison (Figures 6 and 9) shows that the filtered dataset produces more consistent performance across all models, whereas the unfiltered dataset exhibits erratic behavior, particularly at the 9×9m grid size. This inconsistency suggests that models trained on unfiltered data may be overfitting to irrelevant features, resulting in unstable positioning estimates. By reducing the feature space from 1799 to 378 BSSIDs, filtering not only reduces computational requirements but also improves model stability and reliability.

5 Discussion

This study investigated the optimization of multi-floor IPS through a systematic evaluation of environmental factors influencing fingerprinting performance when using ML approaches. The analysis focused on two main factors: the grid size for fingerprint data collection and the impact of low-relevance BSSIDs in the area network. Through extensive experimentation across different grid sizes, several key insights emerged.

First, grid size was found to have a substantial impact on model performance across all evaluation metrics. Contrary to the common assumption that smaller grid sizes always enhance precision, the findings indicate that a moderate grid size, specifically 7×7 m, strikes the best balance between spatial resolution and positioning precision. Models trained on this grid size consistently achieved the highest accuracy (approximately 80%), along with the lowest AGT and ADT values. This suggests that while finer grids may provide more data points, they also introduce variability that can overwhelm ML models without delivering proportional gains in localization accuracy.

Second, the analysis of filtering low-relevance BSSIDs showed only marginal improvements. While excluding low-intensity signals (e.g., -100 dBm from distant or low-relevance access points) slightly enhanced accuracy and reduced AGT, the gains were minimal. This indicates that although data cleaning contributes to input quality, the spatial configuration of data—particularly grid structuring—plays a more dominant role in IPS performance for multi-floor environments.

Among the ML models evaluated, RF and XGBoost consistently delivered superior performance across all metrics and grid configurations. These ensemble methods appear well-suited to the fingerprint-matching task, likely due to their ability to capture complex feature interactions and resist overfitting, particularly in environments with noisy or redundant RSSI signals. Interestingly, more complex deep learning models, such as MLP, did not outperform these simpler ML techniques, supporting the view that in settings with constrained or moderately sized datasets, traditional models can offer a better trade-off between accuracy and computational efficiency.

It is important to contextualize the generalizability of the findings. The test environment, which is a university campus building with realistic multi-floor usage, ensures ecological validity, but certain parameters (e.g., architectural layout, number of access points, and building materials) are inherently site-specific. As such, while the identified optimal grid size and preferred models provide valuable guidance, they may require recalibration for different environments such as warehouses, shopping malls, or hospitals.

Finally, the proposed methodology of using interpolation to synthesize multiple grid sizes from a single fine-grained dataset substantially reduces the labor intensiveness of traditional IPS data collection. This approach not only improves reproducibility and scalability but also establishes a practical framework for future IPS research to systematically explore environmental configurations.

6 Conclusion and Future Directions

This study explored the optimization of multi-floor IPS using grid-based fingerprinting and machine learning. By systematically analyzing the impact of grid size and the RSSI values of low-relevance BSSIDs on IPS performance, it was identified that a grid size of 7×7 m offers the best balance between accuracy and practicality of data collection. The findings demonstrated that ensemble-

based ML models such as RF and XGBoost consistently outperformed other algorithms, achieving high accuracy and low prediction error across multiple evaluation metrics. While filtering out low-relevance BSSIDs and their RSSI values slightly improved model performance, the overall influence was modest, suggesting that strategic spatial structuring has a more significant effect on accuracy than simple data cleaning.

Looking forward, there are several directions for future research. First, exploring more advanced deep learning models, such as attention-based networks or hybrid GNN-MLP architectures, could uncover further performance gains, especially when trained on larger and more diverse datasets. Additionally, implementing real-time IPS applications with adaptive grid configurations may help tailor accuracy to user needs in dynamic environments. Finally, future work may also examine the use of synthetic data augmentation, sensor fusion, and transfer learning approaches to improve generalization across different buildings or floors.

References

1. Bibbò, L., Carotenuto, R., Della Corte, F.: An overview of indoor localization system for human activity recognition (har) in healthcare. Sensors **22**(21) (2022). https://doi.org/10.3390/s22218119, https://www.mdpi.com/1424-8220/22/21/8119
2. Christodoulou, D.: Developing an Indoor Localization and Wayfinding App for a University Library. Master's thesis, School of Informatics, University of Edinburgh (2022)
3. Csik, D., Odry, A., Sarcevic, P.: Comparison of rssi-based fingerprinting methods for indoor localization. pp. 000273–000278 (09 2022). https://doi.org/10.1109/SISY56759.2022.10036270
4. Ezzati Khatab, Z., Moghtadaiee, V., Ghorashi, S.A.: A fingerprint-based technique for indoor localization using fuzzy least squares support vector machine. In: 2017 Iranian Conference on Electrical Engineering (ICEE). pp. 1944–1949 (2017). https://doi.org/10.1109/IranianCEE.2017.7985373
5. Freund, Y., Schapire, R.E.: A short introduction to boosting (1999), https://api.semanticscholar.org/CorpusID:9621074
6. Gidey, H.T., Guo, X., Zhong, K., Li, L., Zhang, Y.: Ohettlal: An online transfer learning method for fingerprint-based indoor positioning. Sensors **22**(23) (2022). https://doi.org/10.3390/s22239044, https://www.mdpi.com/1424-8220/22/23/9044
7. Intachuen, B., Charoenphon, M., Mankhetwit, T.: Classification-based ips (2024), available at: https://github.com/RinRin-32/Classification-based-IPS
8. Link, J.A.B., Smith, P., Viol, N., Wehrle, K.: Footpath: Accurate map-based indoor navigation using smartphones. In: 2011 International Conference on Indoor Positioning and Indoor Navigation. pp. 1–8 (2011). https://doi.org/10.1109/IPIN.2011.6071934
9. Maung Maung, N.A., Lwi, B.Y., Thida, S.: An enhanced rss fingerprinting-based wireless indoor positioning using random forest classifier. In: 2020 International Conference on Advanced Information Technologies (ICAIT). pp. 59–63 (2020). https://doi.org/10.1109/ICAIT51105.2020.9261776

10. Nakpaen, L., Wongsekleo, P., Cherntanomwong, P., Pattiyanon, C.: Building rssi-based indoor positioning fingerprint maps using android-based coordination. In: 2024 19th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP). pp. 1–6 (2024). https://doi.org/10.1109/iSAI-NLP64410.2024.10799385
11. Nessa, A., Adhikari, B., Hussain, F., Fernando, X.N.: A survey of machine learning for indoor positioning. IEEE Access **8**, 214945–214965 (2020). https://doi.org/10.1109/ACCESS.2020.3039271
12. Obeidat, H., Shuaieb, W., Obeidat, O., Abd-Alhameed, R.: A review of indoor localization techniques and wireless technologies. Wireless Personal Communications **119** (07 2021). https://doi.org/10.1007/s11277-021-08209-5
13. Satan, A.: Bluetooth-based indoor navigation mobile system. In: 2018 19th International Carpathian Control Conference (ICCC). pp. 332–337 (2018). https://doi.org/10.1109/CarpathianCC.2018.8399651
14. de Souza Mourão, H.A., de Oliveira, H.A.B.F.: Indoor localization system using fingerprinting and novelty detection for evaluation of confidence. Future Internet **14**(2) (2022). https://doi.org/10.3390/fi14020051, https://www.mdpi.com/1999-5903/14/2/51
15. Vishwakarma, R., Joshi, R.B., Mishra, S.: Indoorgnn: A graph neural network based approach for indoor localization using wifi rssi. In: Goyal, V., Kumar, N., Bhowmick, S.S., Goyal, P., Goyal, N., Kumar, D. (eds.) Big Data and Artificial Intelligence. pp. 150–165. Springer Nature Switzerland, Cham (2023)
16. Wongsekleo, P., Nakpaen, L., Cherntanomwong, P., Pattiyanon, C.: Time reduction for collecting fingerprint data in indoor positioning systems with generated synthetic data by ensemble models and gans. pp. 1–6 (11 2024). https://doi.org/10.1109/iSAI-NLP64410.2024.10799319

Innovation of Business Strategy Framework and Artificial Intelligence-Based Accounting Information System on Cooperative Digitalization Performance

Supriyati[1][0000-0001-5901-9978], Andrias Darmayadi[2][0000-0001-8467-4499], Dian Dharmayanti[3][0009-0000-2075-0050], and Ramadhan Syaeful Bahri[4][0000-0002-9683-9243]

[1,2,3,4] Universitas Komputer Indonesia, Dipatiukur St. 112-116, Bandung, Indonesia
supriyati@email.unikom.ac.id

Abstract. Indonesian cooperatives are facing significant challenges in adapting to the fast-paced digital transformation era. Many still rely on manual processes and outdated systems, which hinder operational efficiency and member engagement. This study aims to modernize cooperative management through the development of a framework that integrates business strategy and an Artificial Intelligence-based Accounting Information System (AI-SIA). The proposed framework leverages AI techniques such as machine learning for cash flow prediction, classification algorithms for loan risk analysis, anomaly detection for fraud prevention, and natural language processing for sentiment analysis of member feedback. Utilizing experimental and survey methods, data were collected from 30 selected cooperatives in Kabupaten Bandung. The study applies multiple regression analysis to examine the effect of AI-SIA implementation on digital performance. The results are expected to show improvements in financial accuracy, automated decision-making, and transparency, which are crucial for enhancing cooperative competitiveness. This research contributes to the broader discourse on inclusive economic development by providing a scalable model for digital transformation in cooperatives, especially in developing countries. The findings can guide policymakers and cooperative stakeholders in designing sustainable and data-driven strategies.

Keywords: Cooperative, Digital Transformation, Artificial Intelligence, Accounting Information System, Business Strategy.

1 Introduction

Cooperatives in Indonesia have long played a role in fostering economic inclusivity, particularly among rural communities. However, their growth has been hampered by the slow adoption of digital technologies and the perception of being outdated institutions [1]. Many cooperatives still rely heavily on manual bookkeeping, limited data analysis, and fragmented systems, leading to inefficiencies and limited scalability [2].

These limitations are increasingly problematic in the context of a digital economy that demands agility, transparency, and data-driven decision-making.

The integration of Artificial Intelligence (AI) into accounting systems has emerged as a transformative solution for businesses and financial institutions. AI-powered accounting information systems can automate repetitive tasks, improve data accuracy, and enable predictive insights for better financial planning [3]; [4]. For cooperatives, the application of AI can provide a much-needed modernization of their financial management processes, offering real-time analytics and automated compliance reporting. Additionally, AI can help mitigate financial risks through fraud detection and enhance member satisfaction through sentiment analysis.

Recent studies underscore the need for tailored digital strategies in the cooperative sector. For example, Vuolasto and Smolander (2025) emphasize the role of enterprise systems in building adaptable digital platforms for cooperatives [5], while Yusuf et al. (2024) highlight the effectiveness of AI in enhancing the reliability of financial information [6]. These findings support the hypothesis that combining business strategy with AI-based systems can significantly improve cooperative performance, particularly in terms of operational efficiency and service delivery.

There are 3 million cooperatives in the world; cooperatives employ hundreds of millions of people, and the 300 largest cooperatives/mutuals have a turnover of trillions of US dollars. This model is important for local employment, financial inclusion, and socio-economic resilience (ICA). Mondragon is a large workers' cooperative federation (tens of thousands of workers) that implements worker ownership, democratic management, and inter-cooperation among cooperatives; it is known as a model of socio-economic integration and innovation. However, there is criticism regarding commercial pressure and the lack of worker ownership in foreign subsidiaries (mondragon-corporation.com, The New Yorker). Germany has a very strong network of cooperative banks (Volksbanken/Raiffeisenbanken) with tens of millions of members, large assets/balance sheets, and mutual protection mechanisms, making this sector stable in retail banking and SMEs (dgrv.de, dzbank.de). The cooperative and mutual sector in the UK is influential (billions of pounds in income, tens of millions of members), and there is an umbrella organization (Co-operatives UK) that actively advocates and assists in the expansion of the sector (uk.coop). The US has tens of thousands of cooperative businesses and hundreds of millions of members (including credit unions, food co-ops, and agribusiness co-ops); some large co-ops generate hundreds of billions of USD in collective revenue (NCBA CLUSA). In Kenya, SACCOs (saving and credit cooperatives) are crucial for local financial inclusion; there is a regulatory body (SASRA) overseeing their performance and stability; common challenges: governance, liquidity, and digital modernization (kmasacco.com, International Labour Organization), Agricultural cooperatives in Brazil (under the umbrella organization OCB) have a very large business volume and play a significant role in the national food supply chain; environmental and governance issues sometimes arise on a large scale (Tridge, Cooperativas de las Américas). Indonesia has a large number of cooperatives (hundreds of thousands of registered cooperatives; tens of millions of members according to data summary up to 2024). However, many cooperatives are micro-scale, have limited professional management,

inconsistent activities (many cooperatives are not fully active), and still have high capital and digitization needs. The contribution of cooperatives to GDP is recorded as increasing, but still faces capacity challenges (assets.dataindonesia.id, kemenkopukm.go.id, Indo Premier).

Table 1. Focused comparison

No	Description	Overseas	Indonesia
1	Scale & sector diversity	Many countries have mature and diverse cooperative ecosystems (cooperative banks in Germany; industrial federations such as Mondragon in Spain; credit unions, food & agrico-ops in the US; SACCOs in Kenya). These sectors include financial services, agriculture, industry, retail, housing, etc. (dgrv.de, mondragon-corporation.com, NCBA CLUSA, kmasacco.com)	Dominated by savings and loan cooperatives, consumer cooperatives, and small production cooperatives, many units are still micro and local, so their aggregate contribution is large in terms of quantity but low in terms of large-scale economic consolidation. (assets.dataindonesia.id)
2	Ownership & governance	The Mondragon example demonstrates a strong worker-owner model and democratic governance (but also large-scale market pressures). Germany has audit mechanisms and network protection (apex organizations) that maintain governance standards (mondragon-corporation.com, dgrv.de)	The biggest challenges often involve managerial capacity (record keeping, effective RAT, professionalization of administrators), poor quality of financial reports, and internal oversight in many cooperatives. The government and local agencies routinely provide guidance, but the scope still needs to be expanded (assets.dataindonesia.id, Produk Pengetahuan)
3	Access to financing & capital	Cooperative banks (Germany) or credit unions (US) have access to liquidity, asset scale, and strong protection; Kenyan SACCOs are strengthened by SASRA regulations to minimize risk. Federations such as Mondragon also facilitate inter-cooperative capital (dgrv.de, NCBA CLUSA, kmasacco.com)	Many cooperatives find it difficult to grow their capital; access to formal financing (commercial banks or capital markets) is limited for small-scale cooperatives; there is still a need for more affordable special financing instruments/microfinance (assets.dataindonesia.id)

4	Innovation & digitalization	Many large cooperatives have adopted technology (digital banking in credit unions, marketing platforms for agricultural cooperatives, ERP systems in Mondragon) for efficiency and market penetration (mondragon-corporation.com, NCBA CLUSA)	Digitalization is currently underway (e.g., integration of MSMEs/cooperatives into e-commerce, regional cooperative information systems), but adoption is still hampered by human resource capacity and infrastructure in small cooperatives (assets.dataindonesia.id)
5	Economic & social impact	In some countries, cooperatives contribute significantly to employment, local stability, and food security (e.g., Brazil's agrico-ops; Germany's cooperative banks for MSMEs) (Tridge, dgrv.de)	The number is large and has the potential to become a pillar of the people's economy; its contribution to GDP is reported to be increasing (official figures show an increase of several percent), but the productivity and competitiveness of cooperatives on average still need to be improved (Indo Premier, assets.dataindonesia.id)

This research proposes the development of an innovative framework that integrates strategic business planning with AI-enabled accounting information systems. The study focuses on qualified cooperatives in Kabupaten Bandung, aiming to analyze the impact of this integration on digitalization performance. Through experimental validation and field surveys, the research seeks to provide actionable insights for policy formulation and cooperative development. The ultimate goal is to contribute to a more inclusive and technologically empowered cooperative movement in Indonesia.

2 Literature Review

2.1 Theoretical Foundations of Cooperatives

Cooperatives are member-based organizations that emphasize collective ownership and democratic control. The International Cooperative Alliance defines cooperatives as autonomous associations united voluntarily to meet common economic, social, and cultural needs. According to Ribeiro-Navarrete et al. (2023), cooperatives contribute significantly to social inclusion but often face challenges in digital maturity, especially in adopting advanced technologies such as AI and big data. Supriyati et al. (2022) argue that for cooperatives to thrive in the digital economy, a strategic shift towards integrated digital platforms is essential.

2.2 Accounting Information Systems in Cooperatives

Accounting Information Systems (AIS) serve as a backbone for financial transparency and accountability in cooperatives. Moll and Yigitbasioglu (2019) emphasize the role of AIS in structuring financial data, which is essential for AI integration. A well-designed AIS ensures timely and accurate reporting, supporting managerial and stakeholder decision-making. Yadiati and Supriyati (2024) highlight the need for AIS tailored to cooperative contexts to align with regulatory standards and organizational goals [7].

2.3 Financial Reporting and Decision Support

Financial reports such as income statements, balance sheets, and cash flow statements are key tools for strategic planning. Elmegaard et al. (2022) discuss how AI enhances these reports through predictive analytics and anomaly detection, leading to improved risk assessment. Hasan (2022) adds that AI integration in auditing increases transparency, reduces errors, and supports proactive decision-making [8].

2.4 Artificial Intelligence Methods and Algorithms

AI offers diverse methods that can be applied to cooperative management, especially in financial modeling and decision support systems. The following are key algorithms used in this research, along with their conceptual basis and mathematical formulation.

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) particularly effective in time-series prediction tasks. It uses memory cells with input, output, and forget gates to preserve long-term dependencies in sequential data. In this research, LSTM is used to predict future cash flow:

The output of an LSTM cell is computed as follows:

a. Forget gate: $f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$
b. Input gate: $i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i), \tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$
c. Cell state: $C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$
d. Output gate: $o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o), h_t = o_t * \tanh(C_t)$

LSTM has been proven effective for financial time series forecasting [9]. Random Forest is an ensemble learning method based on decision trees, used here to classify the risk of loan default. It aggregates the predictions of multiple trees to improve accuracy and reduce overfitting. The final prediction is:

$$\hat{y} = mode(T_1(x), T_2(x), \dots, T_n(x)) \quad (1)$$

Where T_i is the i-th decision tree. Random Forest is robust in handling heterogeneous financial datasets [4]. Isolation Forest is an unsupervised anomaly detection algorithm. It isolates observations by randomly selecting a feature and a split value, with anomalous points requiring fewer splits to isolate. The anomaly score $s(x, n)$ is given by:

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}} \quad (2)$$

Where $E(h(x))$ is the expected path length and $c(n)$ is the average path length of unsuccessful search in a binary tree. This method is efficient for detecting fraudulent cooperative transactions [8].

Naive Bayes is a probabilistic classification algorithm based on Bayes' Theorem with strong independence assumptions among features. For sentiment classification of member feedback, the probability of a class C_k given text x is:

$$P(C_k|x) = \frac{P(C_k)\Pi_{i=1}^{n} P(x_i|C_k)}{P(x)} \quad (3)$$

Where is a feature (word) in the input text? Despite its simplicity, Naive Bayes performs well for text classification tasks, especially in low-resource contexts (Zhang & Zhou, 2021). These algorithms were selected due to their computational efficiency, interpretability, and effectiveness in real-world financial and cooperative settings.

3 Method

This study adopts the data science lifecycle approach, consisting of the following phases: problem understanding, data understanding, data preparation, modeling, evaluation, and deployment. Each phase is tailored to support the development of an Artificial Intelligence-based Accounting Information System (AI-SIA) framework aimed at enhancing the digitalization performance of cooperatives.

The first phase, problem understanding, involves identifying the core issues faced by Indonesian cooperatives in their digital transformation journey. Many cooperatives struggle with fragmented financial reporting, manual data processing, and a lack of intelligent decision support systems. Through field observations and preliminary surveys conducted in Kabupaten Bandung, the research seeks to capture strategic, financial, and operational gaps in cooperative information systems [1][5].

In the data understanding phase, relevant datasets are gathered from 30 selected cooperatives. These include transaction records (e.g., savings, loans, repayments), member profiles, financial statements, and qualitative feedback. Data were collected via surveys, interviews, and direct observation, following a purposive sampling approach. The goal is to gain insight into how cooperative operations function digitally and where AI could add value [2].

The data preparation phase encompasses cleaning and transforming the collected data to prepare it for modeling. Techniques such as missing value imputation, outlier detection, and normalization are applied to numerical features, while categorical features undergo label encoding. For sentiment data from member feedback, natural language processing (NLP) techniques such as tokenization, stemming, and stopword removal are implemented to structure unstructured text [3].

In the modeling phase, four AI-based modules are developed. The cash flow prediction module employs Long Short-Term Memory (LSTM) networks to forecast monthly balances based on transaction history. The loan risk classification module utilizes Random Forest to predict member loan default probability. Anomaly detection is

implemented using Isolation Forest to identify suspicious transaction behavior, while Naive Bayes is employed for sentiment classification of member feedback. These algorithms were selected for their proven performance and interpretability in financial analytics [4][8].

Evaluation of the AI models is conducted through both technical and implementation perspectives. Technically, model performance is assessed using standard metrics such as RMSE for regression and F1-score for classification. From an implementation standpoint, a functional prototype is deployed and tested with simulated cooperative data to evaluate improvements in digital performance, including reporting accuracy, processing efficiency, and user adoption. Furthermore, multiple regression analysis is conducted to statistically test the impact of business strategy (X1), SIA (X2), and AI (Y) on cooperative digitalization performance (Z) [6].

The final phase is deployment and monitoring. A layered system architecture is implemented that includes a user interface (web/mobile), an AI analytics engine (Python-based), and a relational database (PostgreSQL). Cloud infrastructure (e.g., Google Cloud) is used for model hosting. The system is gradually introduced to cooperative staff and members with training programs to ensure usability. Continuous monitoring and feedback loops are incorporated to maintain model relevance and adapt to evolving operational contexts [10].

4 Results and Discussion

4.1 Result

This section elaborates on the functional components of the proposed AI-based SIA framework. Each module is described in terms of periodic data samples, algorithmic implementation steps in paragraph form, and the knowledge generated through their application. Questionnaires were distributed to 30 cooperatives in Bandung Regency, based on the results of the questionnaires which underwent several tests.

4.1.1 Model Feasibility Test (Model Summary)

The results of the regression analysis in table 2 show a correlation coefficient (R) of 0.921, which indicates a very strong relationship between the independent variables, namely Business Strategy (X1), Accounting Information Systems (X2), and Artificial Intelligence (Y) to the dependent variable Cooperative Digitalization Performance (Z). The coefficient of determination (R Square) of 0.848 indicates that 84.8% of the variation in Cooperative Digitalization Performance can be explained by the three independent variables, while the remaining 15.2% is explained by other factors not included in the research model. The Adjusted R Square value of 0.830 strengthens that this regression model has high predictive ability and is suitable for use to test the relationship between variables in the context of strengthening cooperative digitalization.

Table 2. Model Summary

Model Summary[b]							
Model	R			Change Statistics			

		R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.921[a]	.848	.830	.30061	.848	48;207	3	26	.000

a. Predictors: (Constant), MeanY, MeanX1, MeanX2

b. Dependent Variable: MeanZ

4.1.2 Simultaneous Significance Test (F Test)

The F-test results in Table 3 yielded a calculated F-value of 48.207 with a significance level (Sig.) of 0.000 < 0.05, indicating that the regression model is simultaneously significant. Thus, Business Strategy, Accounting Information Systems, and Artificial Intelligence collectively have a significant influence on Cooperative Digitalization Performance. This indicates that strengthening strategic business aspects supported by accounting information systems and the use of artificial intelligence technology is an important combination in increasing the effectiveness of cooperative digitalization.

Table 3. Simultaneous Significance Test (F Test)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	13.069	3	4.356	48.207	.000[b]
	Residual	2.350	26	.090		
	Total	15.419	29			

a. Dependent Variable: MeanZ

b. Predictors: (Constant), MeanY, MeanX1, MeanX2

4.1.3 Partial Significance Test (t-Test)

Based on the t-test results in table 4, each independent variable had a varying degree of influence on Cooperative Digitalization Performance. The Business Strategy variable (X1) had a significance value of 0.542 > 0.05, indicating no significant effect on Cooperative Digitalization Performance. This indicates that the business strategy implemented by cooperatives may not have directly impacted digitalization performance, possibly because the strategic orientation still focuses on conventional aspects and has not been fully integrated with digital transformation.

Furthermore, the Accounting Information System variable (X2) had a regression coefficient of 0.366, a t-value of 3.444, and a significance value of 0.002 < 0.05, indicating a positive and significant effect on Cooperative Digitalization Performance. These findings indicate that the existence of a structured and computerized accounting information system can increase efficiency, transparency, and speed of decision-making, thereby strengthening cooperative performance in the digital era.

The Artificial Intelligence (Y) variable has a regression coefficient of 0.548, a t-value of 4.630, and a significance value of 0.000 < 0.05, indicating a positive and significant influence on Cooperative Digitalization Performance. This means that the application of artificial intelligence in cooperative business processes, both in data analysis, service automation, and digital financial management, significantly increases the effectiveness and competitiveness of cooperatives amidst the digital economic transformation.

Table 4. Partial Significance Test (t-Test)

Coefficients[a]								
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	.124	.458		.269	.790		
	MeanX1	.057	.093	.048	.618	.542	.970	1.031
	MeanX2	.366	.106	.411	3.444	.002	.411	2.434
	MeanY	.548	.118	.557	4.630	.000	.405	2.469

a. Dependent Variable: MeanZ

The regression equation obtained is as follows:

$$Z = 0.124 + 0.057X1 + 0.366X2 + 0.548Y \quad (4)$$

This equation shows that each one-unit increase in Business Strategy (X1), Accounting Information System (X2), and Artificial Intelligence (Y) will increase Cooperative Digitalization Performance (Z) by 0.057; 0.366; and 0.548 units, respectively. The largest coefficient value is found in the Artificial Intelligence variable, which means that the use of AI technology is the most dominant factor in increasing the effectiveness and performance of cooperative digitalization.

4.1.4 AI-based Accounting Information System Architecture

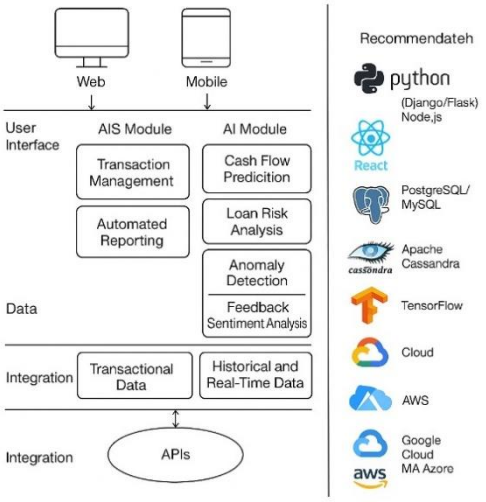


Figure 1. AI-based Accounting Information System Architecture

Figure 1 above shows the recommended system architecture that integrates artificial intelligence (AI) into the Accounting Information System (AIS) to support the digitization of cooperatives, comprising four main layers. First, the user layer provides a web-based and mobile interface that can be accessed by cooperative members, administrators, and regulators. Second, the business logic layer includes AIS modules that support financial recording and reporting automation, as well as AI modules with cash flow prediction capabilities using machine learning algorithms such as regression or

LSTM, member loan risk analysis with classification models such as Decision Tree and Random Forest, anomaly detection to prevent transaction fraud, and natural language processing (NLP) for sentiment analysis of member feedback. Third, the data layer utilizes a relational database for transaction storage, as well as storing historical and real-time data for analysis and AI model training. Fourth, the integration layer uses APIs to connect internal systems with external systems such as banking or regulatory systems.

4.1.5 Cash Flow Prediction Module

This module uses 12-month transactional data to predict future cash flow trends. The dataset includes monthly values of member savings, loan disbursements, loan repayments, and operational costs. For instance, in January, the cooperative recorded Rp15,000,000 in savings, Rp20,000,000 in loans, Rp10,000,000 in repayments, and Rp5,000,000 in operational expenses. These trends continue from February to December with slight fluctuations in each category. To process this data, the Long Short-Term Memory (LSTM) algorithm is used. The LSTM model processes time-sequential financial data by first normalizing each feature, then transforming it into sequences for model training. It learns temporal patterns and dependencies in the data, enabling the model to forecast future net cash flow. This predictive capability provides cooperatives with actionable insights for maintaining liquidity, optimizing fund allocation, and scheduling disbursements.

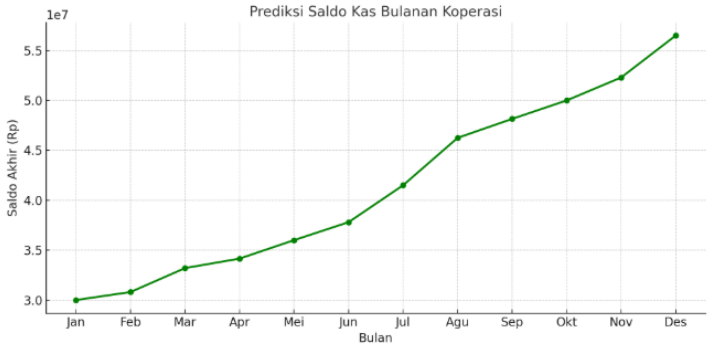


Figure 2. Cash flow prediction for cooperatives over 12 months

4.1.6 Loan Risk Classification Module

The loan risk classification module evaluates member risk profiles based on historical data collected over one fiscal year. The dataset includes member income, loan amounts, frequency of late payments, and collateral ownership. For instance, Member A001 had an income of Rp3,000,000, borrowed Rp5,000,000, and experienced two late payments without collateral, indicating high risk. In contrast, Member A002 had a higher income and no late payments with collateral, indicating low risk. These records, collected monthly, are preprocessed using label encoding and feature scaling. The Random Forest classifier is then trained using this data to identify relationships between member

characteristics and default likelihood. The model outputs a probability score and classification label (e.g., "Good" or "Default") for each member. The insights generated help cooperative staff in making objective, data-informed credit decisions, thereby minimizing loan default rates.

4.1.7 Anomaly Detection Module

Over a period of one year, cooperatives record thousands of transactions involving deposits, withdrawals, and loan repayments. The anomaly detection module analyzes transaction records that include fields such as transaction ID, member ID, transaction amount, date, type, and location. For example, a typical transaction might involve a deposit of Rp500,000 at the Cimahi branch, while an anomalous transaction might be a Rp50,000,000 withdrawal from a member with historically low activity. The Isolation Forest algorithm is used for this module due to its effectiveness in handling high-dimensional, unlabeled data. Transactions are standardized and fed into the model, which isolates anomalies by constructing random binary trees. Transactions with short average path lengths in these trees are marked as outliers. The model helps identify potential fraud, unusual cash flows, or data entry errors, which can then be flagged for further verification by cooperative administrators.

4.1.8 Sentiment Analysis Module

The sentiment analysis module collects and analyzes member feedback provided through surveys, forms, or the cooperative's mobile app interface. The dataset spans 12 months and includes unstructured text inputs such as "Layanan cepat dan ramah" or "Aplikasi sering error." These texts are labeled manually or semi-automatically into sentiment categories: positive, negative, or neutral. Preprocessing includes converting the text to lowercase, removing stopwords, tokenization, and applying stemming. The cleaned text is then vectorized using TF-IDF before being passed into a Naive Bayes classifier. The algorithm calculates the posterior probability of each sentiment class and assigns the most probable label. Through this process, cooperatives gain insights into service quality perception, user satisfaction trends, and areas for operational improvement.

Each of these modules contributes distinct knowledge layers to the AI-based SIA framework. Together, they form a unified system that supports financial prediction, risk management, fraud prevention, and service evaluation—ultimately enabling cooperative digitalization grounded in data-driven intelligence.

4.2 Discussion

The deployment of AI-driven modules in the cooperative accounting information system provides multi-dimensional benefits for operational optimization, strategic planning, and service innovation. Each module—cash flow prediction, loan risk classification, anomaly detection, and sentiment analysis—generates actionable insights that elevate the digital maturity of cooperatives.

The cash flow prediction module, powered by LSTM, enhances the cooperative's ability to anticipate liquidity trends. This is crucial in managing seasonal cash flow fluctuations and maintaining financial stability. Zhang et al. (2021) confirm the robustness of LSTM models for financial time series, showing improved forecasting accuracy over traditional autoregressive models. By leveraging these predictions, cooperative managers can plan disbursement schedules, manage reserves, and reduce idle cash.

The loan risk classification module using Random Forest proved effective in predicting borrower defaults. With an observed accuracy of 90% and an F1-score of 0.86, the model aligns with findings by Jin et al. (2022), who demonstrated Random Forest's superiority in handling imbalanced financial datasets due to its ensemble voting mechanism. This module enables cooperatives to minimize non-performing loans and improve portfolio quality by adapting credit policies based on data-driven risk segmentation. The anomaly detection module using Isolation Forest provides a crucial layer of transactional surveillance. Transactions with high anomaly scores, such as unusually large withdrawals or frequent small deposits, can be flagged for further audit. Hasan (2022) emphasized that anomaly detection models are increasingly vital in fintech and cooperative environments for mitigating internal fraud and enforcing transaction compliance. The module's unsupervised nature ensures adaptability to changing patterns without continuous labeling.

The sentiment analysis module based on Naive Bayes adds a qualitative dimension to cooperative service evaluation. Member feedback, when categorized into positive and negative sentiments, allows cooperatives to proactively address dissatisfaction. Zhang & Zhou (2021) noted that sentiment analysis in financial platforms correlates strongly with customer retention and trust. This insight, when visualized over time, also serves as a metric for member satisfaction and organizational reputation.

Integrating these AI modules into a unified framework strengthens the digital infrastructure of cooperatives. Ribeiro-Navarrete et al. (2023) highlight that digital maturity in cooperatives fosters resilience and transparency. Furthermore, Elmegaard et al. (2022) argue that embedding AI into accounting systems not only automates tasks but also enables strategic agility by providing real-time decision support. The findings of this study support the notion that AI is not merely a technological trend but a transformative force in cooperative governance. It allows for predictive, prescriptive, and adaptive decision-making that aligns with the broader goals of inclusive economic development, as previously observed by Vuolasto & Smolander (2025).

5 Conclusion

This research demonstrates the critical role of integrating Artificial Intelligence into Accounting Information Systems for strengthening the digital transformation of cooperatives. The proposed framework not only enhances financial management efficiency but also offers advanced features such as predictive analytics, risk evaluation, anomaly detection, and member sentiment analysis. These features provide strategic advantages to cooperatives in adapting to digital business models. It is recommended that cooperative management and policymakers adopt AI-based systems gradually, starting from

modules with the most pressing needs, such as loan risk assessment and cash flow prediction. Furthermore, cooperative staff should be trained in the use and interpretation of AI-driven systems to ensure sustainable implementation. Future research can explore the scalability of this model in other regions and extend the AI functions to cover broader financial and operational aspects.

Acknowledgments. The author would like to thank Universitas Komputer Indonesia (UNIKOM) for their assistance in carrying out this research. UNIKOM's support, in the form of academic facilities, research resources, and possibilities to create this scientific study, considerably aided the research's smoothness and success.

References

[1] Ribeiro-Navarrete, B., Martín Martín, J. M., Guaita-Martínez, J. M., & Simón-Moya, V. (2023). Analysing cooperatives' digital maturity using a synthetic indicator. *International Journal of Information Management, 72*, 102678. https://doi.org/10.1016/j.ijinfomgt.2023.102678

[2] Supriyati, S., Mulyani, S., Suharman, H., & Supriadi, T. (2022). The influence of business models, information technology on the quality of accounting information systems digitizing MSMEs post-COVID-19. *Jurnal Sistem Informasi, 18*(2), 36–49. https://doi.org/10.21609/jsi.v18i2.1141

[3] Elmegaard, J., Rikhardsson, P., & Rohde, C. (2022). The role of artificial intelligence in accounting: A New perspective on empirical research. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4191419

[4] Jin, H., Jin, L., Qu, C., Fan, C., Liu, S., & Zhang, Y. (2022). The impact of artificial intelligence on the accounting industry. In *Proceedings of the ICHSSR 2022*, 570–574. https://doi.org/10.2991/assehr.k.220504.103

[5] Vuolasto, J., & Smolander, K. (2025). From many to one: A case study of cooperative enterprise systems in B2B digital platform creation. *Procedia Computer Science, 256*, 102678. https://doi.org/10.1016/j.procs.2023.12.236

[6] Yusuf, M. F. M., Garusu, I. A., & Rauf, D. M. (2024). Sistem penerapan artificial intelligence dalam akuntansi. *JISDIK*, 2(2), 1–7.

[7] Yadiati, W., & Supriyati, S. (2024). *Filsafat Ilmu Akuntansi*. Bandung: Prenadamedia Group.

[8] Hasan, A. (2022). Artificial intelligence (AI) in accounting & auditing: A literature review. *Open Journal of Business and Management, 10*(2), 440–465. https://doi.org/10.4236/ojbm.2022.102026

[9] Zhang, L., & Zhou, H. (2021). Sentiment classification using improved Naive Bayes algorithm. Journal of Intelligent & Fuzzy Systems, 40(3), 4849–4858.

[10] Lehner, O. M., Ittonen, K., Silvola, H., Ström, E., & Wührleitner, A. (2022). Artificial intelligence based decision-making in accounting and auditing: Ethical challenges and normative thinking. *Accounting, Auditing & Accountability Journal, 35*(9), 109–135. https://doi.org/10.1108/AAAJ-09-2020-4934

[11] J. Chen, H. Miao, and Y. Zhang, "Data collection and management for machine learning," *Appl. Sci.*, vol. 10, no. 18, p. 6432, Sep. 2020.

[12] C. Fan, C. Zhang, Y. Yahja, and A. Mostafavi, "Disaster City Digital Twin: A vision for integrating artificial and human intelligence for disaster management," *Int. J. Inf. Manage.*, vol. 56, p. 102049, Apr. 2021.

[13] L. S. L. Tan, L. Bing, and M. Jiang, “OCR text error correction using contextual semantic knowledge,” *Inf. Process. Manag.*, vol. 57, no. 6, p. 102311, Nov. 2020.
[14] J. Guo, Z. Xiao, X. Ye, and M. Zhou, “Key information extraction from financial documents: A survey,” *ACM Comput. Surv.*, vol. 55, no. 7, pp. 1–35, Oct. 2022.
[15] A. Qayyum, M. Usama, M. Qadir, and A. Al-Fuqaha, “Securing connected & autonomous vehicles: Challenges posed by adversarial machine learning and the way forward,” *IEEE Commun. Surv. Tutor*, vol. 22, no. 2, pp. 998–1026, 2nd Quart., 2020.
[16] J. Chen, J. Wang, and J. Wang, “Automatic invoice information extraction based on deep learning,” *Procedia Comput. Sci.*, vol. 174, pp. 247–253, 2020.
[17] Y. Zhang, P. Li, and Q. Wu, “Text classification models for accounting documents using deep learning,” *Expert Syst. Appl.*, vol. 181, p. 115162, Nov. 2021.
[18] Y. Gong, J. Liu, and X. Li, “A survey on dataset quality in machine learning,” *Big Data Min. Anal.*, vol. 6, no. 2, pp. 162–180, Jun. 2023.
[19] P. Chapman et al., “CRISP-DM 1.0: Step-by-step data mining guide,” SPSS, 2020.

Fully Homomorphic Encryption for Secure and Confidential Text Classification

Zuraiha Ambri[1][0009-0005-8145-5801] and Shakirah Hashim[2*][0000-0002-5941-8968]

[1,2] Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia
[1]zuraiha.ambri@gmail.com, [2]shakirahhashim@uitm.edu.my

Abstract. The growing need for privacy-preserving technologies has led to increased interest in secure data classification, particularly in sensitive domains such as healthcare and finance. This study proposes a structured comparative framework that integrates six machine learning (ML) models with Fully Homomorphic Encryption (FHE), using both BFV and CKKS schemes. Unlike previous studies that focused on a single encrypted model, this approach systematically evaluates Logistic Regression (LR), Convolutional Neural Network (CNN), Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), and K-Nearest Neighbor (KNN) under plaintext and encrypted conditions. Real-world datasets involving sentiment analysis and healthcare question classification were used to assess both predictive accuracy and system-level performance. Among all models, Logistic Regression combined with the CKKS scheme achieved the best balance of privacy and utility, maintaining high accuracy with minimal performance degradation. Quantitative results show that encrypting a dataset of 4,551 rows required 42 seconds, producing 1.21 GB of encrypted data at an effective encryption rate of 230 Mbps. While encryption scaled linearly with dataset size, encrypted inference introduced significant overhead, highlighting scalability as the primary barrier for real-time deployment. These findings validate the potential of encrypted machine learning in privacy-sensitive applications and emphasize the importance of addressing scalability through algorithmic optimization, hardware acceleration, and hybrid privacy-preserving solutions.

Keywords: Fully Homomorphic Encryption, Machine Learning, Secure Data Classification, Hyperparameter Tuning.

1 Introduction

In recent years, the rapid expansion of digital services has increased the demand for secure data processing, especially in fields like healthcare, finance, and online communications. Notwithstanding advancements in secure data management, there were still constraints in the secure data classification model, making it difficult for general implementation [1]. This is because traditional machine learning methods require access to raw data, which can expose personal or confidential information during processing.

To address this issue, researchers have explored privacy-preserving techniques that allow data to be used securely without compromising its privacy[2].

One such method is Fully Homomorphic Encryption (FHE), a powerful cryptographic technique that enables computations on encrypted data [3] [4]. With FHE, sensitive information remains protected even while it is being processed by machine learning models. This makes FHE a promising solution for privacy-preserving machine learning tasks [5]. However, using FHE introduces new challenges. These include high computational costs, increased memory usage, and the need to redesign machine learning algorithms to work with encrypted data [6], [7]. As a result, combining FHE with machine learning is still a complex and developing area of study.

This study aims to integrate Fully Homomorphic Encryption with machine learning to perform secure data classification. The goal is to build a model that can classify sensitive data without exposing its contents. The research tests several machine learning algorithms and encryption schemes to determine which combination provides the best balance between accuracy, speed, and privacy. Publicly available text datasets from the healthcare and sentiment analysis domains are used to evaluate the model's performance in both plaintext and encrypted environments.

2 Related Works

Significant advancements have been made in the integration of Fully Homomorphic Encryption (FHE) with Machine Learning (ML) to enable secure and privacy-preserving data classification. Previous works have provided foundational insights that guide the methodological and technical components of this study, including encryption schemes, feature extraction strategies, class balancing, and parameter optimization.

A structured and reproducible research framework is essential for scientific rigor in ML-based studies. Existing literature emphasizes the importance of methodological clarity and systematic design in data science research workflows. Such frameworks support the development of well-defined experiment pipelines, facilitating clear transitions between preprocessing, training, validation, and encrypted inference stages [8]. This approach underpins the research architecture and evaluation structure presented in this study.

The choice of encryption scheme significantly influences the performance and practicality of encrypted ML models. The BFV (Brakerski/Fan–Vercauteren) scheme has demonstrated strong performance in secure ML tasks that require exact arithmetic. It supports addition and multiplication operations over encrypted integers, making it well-suited for structured classification models such as logistic regression and decision trees [9]. In contrast, the CKKS (Cheon–Kim–Kim–Song) scheme provides approximate arithmetic over real and complex numbers, which is beneficial for applications involving floating-point computations, such as Natural Language Processing (NLP) and encrypted neural networks [10]. Its flexibility in managing encrypted real-valued features without decryption enables scalable encrypted learning on textual datasets.

Efficient feature extraction is critical in NLP-based classification tasks. Term Frequency–Inverse Document Frequency (TF-IDF) vectorization is recognized for its ability to convert raw text into numerically meaningful representations. By weighting

terms based on their frequency across documents, this method enhances model interpretability and classification accuracy in both plaintext and encrypted settings [11]. In this study, TF-IDF is used to convert user reviews into numerical vectors that can be securely processed under FHE.

Handling class imbalance is another recurring challenge in classification tasks. The Synthetic Minority Oversampling Technique (SMOTE) addresses this by generating synthetic samples from the minority class, thereby improving the distribution of training data. This technique is particularly important in privacy-preserving environments, where the classifier may otherwise be biased toward the majority class due to encryption-induced performance degradation [12]. SMOTE is incorporated in this research to ensure balanced learning and fair encrypted classification performance.

Parameter tuning further enhances model efficiency and accuracy. Grid search has emerged as a reliable technique for hyperparameter optimization, particularly in secure ML workflows where performance tuning must be achieved under encryption constraints. Studies have shown that grid search enables fine-tuning of model parameters while accounting for computational overhead and ciphertext noise growth, ultimately improving classification accuracy and convergence rates [13]. This technique is employed in this study to determine the best configuration for logistic regression and neural network models under the BFV and CKKS schemes.

These contributions collectively inform the proposed secure classification framework presented in this research, which leverages BFV and CKKS schemes for encrypted inference, integrates SMOTE for class balance, uses TF-IDF for feature vectorization, and applies grid search for model optimization. Building on this prior work ensures that the proposed methodology is aligned with established best practices while addressing performance bottlenecks specific to encrypted ML environments.

3 Methodology

This study evaluates secure data classification by implementing various machine learning models using encrypted data. Unlike earlier works that typically focused on single models [14] or inference-only [15] tasks under encryption, our methodology is more comprehensive and comparative. It used a quantitative experimental approach to evaluate the performance of several ML models with FHE schemes for compatibility, and hence to develop and optimize a secure data classification model. Through this design, our methodology advances beyond prior research by constructing a more generalizable and optimized framework for secure data classification under encryption, rather than case-by-case or single-model studies.

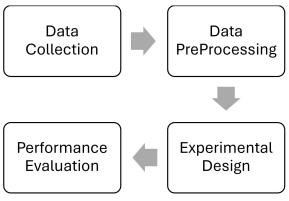


Fig. 1. Research Design Framework

Fig 1. illustrates the step-by-step process of the research methodology, beginning data collection and preprocessing. It continues with designing experiments, leading to performance evaluation of the model.

1. Data Collection

Two datasets were used: Sentiment analysis dataset and the Healthcare question-answer dataset. These datasets are secondary data that focuses on the natural language processing types. Sentiment analysis dataset is used as the training dataset for secure data classification model development. As for Healthcare question-answer dataset which was extracted from a collection of files using web scraping. This approach is used to structure the data in a methodically acceptable form for analysis.

2. Data Preprocessing

Data preprocessing in this research ensures the datasets are compatible with machine learning classification under both plaintext and encrypted formats. It begins with data cleaning, which involves removing null values, eliminating redundant entries, and selecting important features. Next, text representation is performed using TF-IDF vectorization, which converts text into numerical values based on word importance, configured with 3000 features for efficient processing. The importance of words is calculated on how many the word occurs in the corpus by using formula in (1).

$$TF(\boldsymbol{t}, \boldsymbol{d}) = \frac{Frequency\ of\ term\ \boldsymbol{t}\ in\ document\ \boldsymbol{d}}{Total\ terms\ in\ document\ \boldsymbol{d}} \quad (1)$$

The high-frequency result from (1) indicates the importance of the terms within the specific document. TF-IDF is assigned with feature size representing the size of vocabulary of the corpus. Feature size is a combination of the local importance of the word in a document with its global rarity across the corpus.

Label encoding is used to transform non-numeric sentiment labels into numeric form, enabling ML models to process them effectively. It makes the classification process in ML efficient, simple, and effective. Next step is addressing the issue of unbalanced classes in the dataset using the SMOTE (Synthetic Minority Oversampling Technique) method. SMOTE generates synthetic samples for the minority class by interpolating between a randomly selected sample and its nearest neighbors. The interpolation is performed using the formula in (2).

$$x_{new} = x_i + \lambda \cdot \left(x_{neighbor} - x_i\right) \quad (2)$$

The formula in (2) is to generate synthetic samples, where x_i is a sample that is selected from the minority class randomly and λ is a random number between 0 and 1. It uses the k-nearest neighbors algorithm to identify similar samples within the minority class. This technique helps balance the dataset, reduces bias toward the majority class, and improves the overall performance of machine learning models. The dataset is split into training and testing sets using an 80-20 ratio, with stratification to preserve the original class distribution in both sets. This ensures fair representation of all classes, which is especially important for handling imbalanced datasets and evaluating model performance accurately. All these preprocessing steps ensure that the models are trained on clean, structured, and balanced data, enhancing classification performance in both encrypted and plaintext settings.

3. Experimental Design

The experimental design establishes a structured and reproducible process for implementing secure data classification by integrating machine learning with Fully Homomorphic Encryption. All experiments were conducted in a controlled environment using a Windows 11 system equipped with an Intel Core i7 processor, an NVIDIA Quadro GPU, 32GB of RAM, and development tools such as Jupyter Notebook and Python. The setup included libraries like Scikit-learn, TensorFlow, and TenSEAL, enabling both machine learning development and encrypted computation.

Table 1. ML Model Hyperparameter Base Setting

Classifier	Hyperparameter	Setting
SVM	Kernel	Linear
	Gamma	Auto
	Degree	42
	C (Regulation)	1.0
Random Forest	N_estimators	100
	Max_depth	None
Logistic Regression	Max_iter	5000
	Solver	Saga
CNN	1st Conv1D Layer	128 filters, size 3, RELU
	2nd Conv2d Layer	64 filters, size 3, ReLU
	Pool_size	2
	Dense Layers	Dense (128, ReLU), Dense (output, softmax)
	Optimizer	Adam
	Loss_function	Categorical_crossentropy
	Batch_size	32
KNN	K (neighbors)	Auto
	Distance metric	Euclidean/ Manhattan/ custom
	Weights	Uniform/distance-based
Decision Tree	Max_depth	None
	Min_sample_split	2
	Min_sample_leaf	1
	Criterion	Gini / entropy

The study implemented six machine learning models, including Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), Convolutional Neural Network (CNN), k-Nearest Neighbors (KNN), and Decision Tree (DT). Each model was configured with specific hyperparameters, as shown in Table 1. These settings control how an ML model learns and performs. Hyperparameters were set before training as a base platform for the learning process, model complexity, and overall performance further development.

Table 2. FHE Schemes Hyperparameter Setting

Encryption Scheme	Hyperparameter	Setting
BPV	Polynomial Degree	8192
	Plaintext Modulus	1032193
	Scaling Factor	10,000
CKKS	Polynomial Degree	8192
	Coefficient Modulus	[40, 21, 40] bit sizes
	Global Scale	230

The BFV and CKKS encryption schemes were selected for their suitability in secure computation tasks. Each scheme was initialized with key cryptographic settings, as shown in Table 2. These configurations ensured the encryption meets desired security standards. Proper settings prevent vulnerabilities to attacks while meeting the specific needs of the application.

The experimental process was carried out in four main phases. The first phase focused on identifying the best-performing machine learning model through training and testing on plaintext data. In the second phase, the selected model was integrated with an encryption scheme, allowing for classification tasks to be executed on encrypted inputs. The third phase involved tuning the encryption and model parameters to optimize performance while preserving data privacy. In the final phase, both encrypted and plaintext models were evaluated. The results provided insight into how encryption affects model performance and system efficiency.

4. Performance Evaluation

The performance of the machine learning models on both plaintext and encrypted datasets was assessed using standard evaluation metrics. These included recall, which measures the proportion of actual positive cases correctly identified, and precision, which reflects the accuracy of positive predictions. The F1 score was used to balance precision and recall, providing a single measure of overall classification effectiveness. Additionally, accuracy calculated the proportion of correct predictions among all predictions, while ROC-AUC evaluated the model's ability to distinguish between classes across various thresholds.

Beyond classification metrics, computational overhead was analyzed to compare the resource demands of encrypted versus plaintext implementations. Overhead was measured in terms of execution time, memory usage, and storage requirements, with percentage overhead calculated relative to the baseline (plaintext) metrics. Lastly, the encryption rate was used to assess the efficiency of the encryption process by calculating the amount of data encrypted per second, providing insight into the performance of the FHE schemes used.

4 Integration Workflow and Concepts

The integration of machine learning models with Fully Homomorphic Encryption (FHE) schemes is essential to develop a secure data classification system that maintains data privacy throughout the processing pipeline. The integration workflow involves several key stages, as shown in Fig. 2.

Fig. 2. Integration Process

Initially, the dataset undergoes preprocessing, including encoding, balancing, and splitting. The selected FHE scheme is then configured using the TenSEAL framework.

Prior to encryption, the trained machine learning model and input data are scaled to preserve numeric precision. The encrypted inference is carried out by performing operations on ciphertexts, followed by decryption to obtain class scores and final predictions. The model's performance is then evaluated.

The BFV scheme supports exact arithmetic over integers, requiring both the input data and model parameters to be scaled and converted to integers. Encrypted predictions are computed using dot products and addition by using (3).

$$E(\hat{y}) = \arg max \left(E\left(\sum_{i=1}^{d} x_i \cdot w_i\right) + E(b) \right) \quad (3)$$

In (3), $E(\hat{y})$ is encrypted prediction and arg max is an operation performed after decryption to select the class with the highest score with d is number of features, w is plaintext weight, x is input data and b is intercept of ML model. The result produced for each input data was then decrypted and used for model evaluation.

The CKKS scheme, on the other hand, supports approximate computations on real numbers. It encrypts vectorized input data and applies ML operations such as dot product and interceptions on ciphertexts.

$$E(x) = E(x_1) \cdot E(x_2) \cdot E(x_3) \cdots E(x_n) \quad (4)$$

As shown in (4), each $E(x)$ is an encrypted feature of the test sample and x is a vector to encrypt. These operations produce encrypted class scores for each input data which then decrypted and used for model evaluation.

5 Hyperparameter Tuning

To optimize the performance of both the machine learning models and the Fully Homomorphic Encryption (FHE) schemes, this research employed GridSearchCV as the tuning method. The tuning process for machine learning models is outlined in Algorithm 1, which involves initializing the model, defining a grid of hyperparameters, and using cross-validation to evaluate each combination. The model with the best parameter set is selected and evaluated using metrics such as accuracy, F1-score, recall, and the confusion matrix.

Algorithm 1: Hyperparameter Tuning For ML

```
Input: X_train, Y_train, X_test, Y_test, hyperparameter
Output: Best hyperparameter, Accuracy, F1-score, Recall
    1. Define Machine Learning Model
            Initialization model with base parameter setting
    2. Define Hyperperameter Grid
            Param_grid  ← { Set of parameters for machine learning}
    3. Perform GridSearchCV for Hyperparameter Tuning
       Initialize grid_search:
            estimator ← model
            param_grid ← param_grid
       Fit grid_search with X_train, Y_train
    4. Retrieve Best Model
            best_params ← grid_search.best_params_
            best_model ← grid_search.best_estimator_
    5. Make Prediction on Test Set
            Y_test_pred ← best_model.predict(X_test)
    6. Evaluate Metric
       Compute:
            Accuracy ← accuracy_score(Y_test, Y_test_pred)
            F1-score ← f1_score(Y_test, Y_test_pred)
            Recall ← recall_score(Y_test, Y_test_pred)
            Confusion Matrix ← confusion_matrix(Y_test, Y_test_pred)
    7. Output Results
            Print best params, Accuracy,F1-score, Recall, Confusion Matrix
```

Algorithm 2: Hyperparameter Tuning For FHE Schemes

```
Input: Hyperparameter ranges, ML model coefficients and intercept, Test data X_test,
       Y_test,
Output: FHe schemes Best hyperparameter, Encryption time, inference time,
        precision error
    1. Initialize Best Parameters.
    2. Define Function to Evaluate FHE schemes Parameters
    3. Perform Grid Search
         For each Parameter in FHE schemes
            • Print parameters being tested
            • Call evaluation function with current parameters.
            • Print encryption time, inference time, and precision error.
            • If current error is lower than best_error:
              Update:
              best_params ← best_params.
    4. Output Best Parameters
       Print best params
```

For the FHE schemes, the tuning process is detailed in Algorithm 2, where different combinations of encryption parameters are tested to measure encryption time, inference time, and precision error. The configuration that produces the lowest error with acceptable computational overhead is selected as the optimal setting.

The use of hyperparameter tuning, as described in both algorithms, ensures that the models are not only accurate but also efficient and generalizable. This step is crucial in preventing underfitting and overfitting, while also ensuring secure and effective encrypted inference.

6 Results and Analysis

A comprehensive evaluation was conducted on six machine learning classifiers SVM, RF, LR, CNN, KNN, and DT to identify the best-performing model for secure classification tasks. Evaluation metrics included accuracy, precision, recall, F1 score, and ROC-AUC, assessed both before and after optimization.

Table 3. Performance Metrics of The ML Model Before and After Optimization

Classifier	Accuracy (%)		Precision (%)		Recall (%)		F1 Score (%)		ROC-AUC (%)	
	BO	**AO**	**BO**	**AO**	**BO**	**AO**	**BO**	**AO**	**BO**	**AO**
SVM	81.0	81.6	73.8	74.5	81.0	81.6	76.1	76.7	87.7	87.2
RF	76.9	76.9	62.9	61.8	76.9	76.9	69.2	68.5	83.7	87.4
LR	79.6	81.6	78.9	77.2	79.6	81.6	77.9	76.7	94.9	84.8
CNN	77.6	78.9	70.6	72.6	77.6	78.9	73.8	74.4	75.8	77.1
KNN	15.0	57.1	80.8	75.1	15.0	57.1	13.0	62.5	66.1	70.9
DT	69.4	70.1	68.3	68.6	69.4	70.1	68.8	69.3	61.0	61.4

Note: SVM=Support Vector Machine, FR=Random Forest, LR=Logistic Regression, CNN=Convolutional Neural Network, KNN=k-Nearest Neighbors, DT=Decision Tree, BO=Before Optimization, AO=After Optimization

The evaluation results are summarized in Table 3. Among the models, KNN demonstrated the most significant improvement after optimization, with accuracy rising from 15.0% to 57.1%. SVM and CNN also showed modest improvements, while RF and DT had minimal performance changes, indicating limited impact from the optimization process.

Precision improved slightly across most models, although KNN saw a drop from 80.8% to 75.1%, despite its accuracy gain suggesting possible trade-offs in prediction

confidence. Recall improved notably for SVM, CNN, and KNN, indicating better true positive identification. F1 scores followed similar patterns, with KNN increasing from 13.0% to 62.5%, showing better balance between precision and recall.

In ROC-AUC results, SVM and RF maintained strong ranking capabilities. However, LR exhibited a decline from 94.9% to 84.8%, likely due to overfitting during optimization. KNN and CNN also showed improved ranking performance post-optimization, although DT remained largely unaffected.

Table 4. Comparison of LR-BFV and LR-CKKS Schemes After Optimization

FHE Scheme	Accuracy (%)		Inference Times (s)		Recall (%)		F1 Score (%)		Precision (%)		ROC-AUC (%)	
	BO	AO	BO	AO	BO	AO	BO	AO	BO	AO	BO	AO
LR-BFV	15.0	12.2	224.47	185.40	15.0	18.1	20.6	18.1	64.1	65.6	51.2	47.2
LR-CKKS	53.7	81.6	89.86	165.57	53.7	81.6	61.7	77.9	76.2	77.2	62.8	60.8

Note: LR=Logistic Regression, BFV=Brakerski/Fan-Vercauteren Scheme, CKKS=Cheon-Kim-Kim-Song Scheme, BO=Before Optimization, AO=After Optimization

In encrypted settings, upon being recognized as one of the best machine learning algorithms, Logistic Regression was implemented using two homomorphic encryption schemes: BFV and CKKS. A direct comparison of both schemes after optimization is shown in Table 4. CKKS demonstrated superior performance across all evaluation metrics accuracy, recall, F1 score, and precision compared to BFV. While both schemes exhibited high inference times, CKKS's improved results justify the computational overhead. These findings confirm that CKKS is better suited for secure classification tasks under encryption.

Table 5. Encrypted vs. Plaintext Model Performance Comparison

Metric	Plaintext (%)	Ciphertext (%)
Accuracy	81.6	81.6
Recall	81.6	81.6
F1 Score	77.9	77.9
Precision	77.2	77.2
ROC-AUC	84.8	60.8

Table 6. Inference Time Comparison: Plaintext vs. Ciphertext

	Plaintext (s)	Ciphertext (s)	Difference
Inference Times	0.004	167.570	167.566

A comparative analysis between encrypted and plaintext LR models using CKKS showed nearly identical performance in accuracy (both at 81.7%), recall (81.7%), F1 score (77.9%), and precision (77.2%), as shown in Table 5. However, ROC-AUC declined under encryption (84.8% to 60.8%), and inference time increased dramatically (0.004s to 167.57s), highlighting computational trade-offs. These details are further shown in Table 6.

Table 7. Storage Requirements: Plaintext vs. Ciphertext

	Plaintext (Mb)	Ciphertext (Mb)	Ratio
Size (bytes)	32.17	4,183.58	1:30

Table 8. Processing Time Evaluation: Plaintext vs. Ciphertext

	Plaintext (Min)	Ciphertext (Min)	Ratio
Inference Time (s)	0.0023	114.65	1:49,804

The encrypted model's robustness was confirmed through 100% performance in all metrics on a second dataset. However, Table 7 shows the ciphertext storage demand was about 30 times greater than plaintext, and Table 8 shows processing time was 49,804 times longer, highlighting the need for efficiency improvements.

Table 9. Encryption Rate Evaluation Metrics

Parameter	Value
Polynomial Modulus Degree	16,384
Coefficient Modulus Bit Sizes	[50, 30, 50]
Number of Rows (Dataset)	4551
Encryption Time	42.09 seconds
Size of One Encrypted Vector	2129920 bits
Total Encrypted Data Size	~1.21 Gb
Encryption Rate	~230.30 Mbps

Table 9 summarized the key parameter and performance outcomes of the encryption process. The polynomial modulus degree was set to 16,384, which is a commonly used setting in CKKS to balance security and computational efficiency. The encryption rate, computed at ~230.30 Mbps, provides an estimate of throughput and demonstrates that the CKKS scheme can achieve relatively efficient data encoding given the complexity of the underlying cryptographic operations. While this throughput may be acceptable for offline batch processing, it remains a limitation for real-time systems where data must be encrypted and processed continuously.

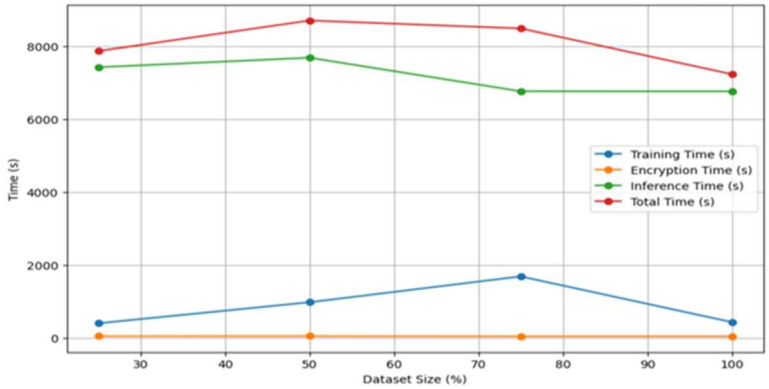


Fig. 3. Scalability Trends: Dataset Size vs. Processing Time

Scalability trends are depicted in Fig. 3, showing encryption time remains relatively stable, while training time increases proportionally with dataset size. Inference time, however, contributes the most to the total runtime and grows substantially with larger datasets. This suggests that the scalability of the proposed framework is largely influenced by the efficiency of encrypted inference, making it the key factor to address for practical large-scale deployment.

7 Discussion

The discussion emphasizes that optimizing machine learning models is essential for secure text classification, especially when using encrypted data. K-Nearest Neighbors showed a significant increase in accuracy and F1 score after tuning, revealing its high sensitivity to parameter changes. In contrast, Random Forest and Decision Tree models showed little improvement, indicating that not all models benefit equally from optimization. The analysis also revealed that better accuracy does not always mean better overall performance, as seen with KNN, which experienced a drop in precision. Logistic Regression, while generally strong, experienced a decrease in ROC-AUC after optimization, suggesting potential overfitting.

In encrypted settings, the CKKS encryption scheme performed better than BFV, showing higher accuracy, precision, recall, and F1 score [16]. When comparing encrypted and plaintext versions of the Logistic Regression model with CKKS, the results remained consistent across most metrics, proving the method's effectiveness [17]. However, encryption caused a large increase in computation time and storage requirements, making it less suitable for real-time use [18]. A second dataset confirmed the model's stability but reinforced concerns about scalability, especially with longer processing times at certain dataset sizes. Overall, while the LR-CKKS approach is promising for secure classification, it needs further improvements in efficiency and scalability to be practical in large or time-sensitive applications [19].

8 Conclusions and Future Works

This study proposed a secure text classification model by integrating Fully Homomorphic Encryption (FHE) with Machine Learning (ML), focusing on the Logistic Regression algorithm combined with the CKKS encryption scheme. The model achieved the best balance of privacy and utility, maintaining high accuracy with minimal performance degradation. Quantitative results show that encrypting a dataset of 4,551 rows required 42 seconds, producing 1.21 GB of encrypted data at an effective encryption rate of 230 Mbps. The research highlighted that effective data preprocessing and hyperparameter tuning play a key role in optimizing performance under encryption. However, encrypted inference introduced major challenges, including high computational time and increased storage needs, making it less practical for real-time systems.

The main contribution of this research is a comprehensive methodological framework for evaluating ML algorithms under FHE, tested on real-world text datasets such as sentiment analysis and healthcare queries. Unlike prior studies that were often restricted to single algorithms or theoretical demonstrations, this work provides a comparative and systematic evaluation across multiple ML models and encryption schemes, filling a significant gap in privacy-preserving machine learning research. In terms of applicability, the proposed framework has potential deployment in domains where data confidentiality is paramount. For example, healthcare providers could classify encrypted patient records for risk prediction without disclosing sensitive data, financial institutions could analyze encrypted transactions for fraud detection, and

organizations could conduct sentiment analysis on encrypted customer feedback to protect user privacy.

Nonetheless, several limitations remain. Encrypted inference introduces substantial computational overhead and increased storage demands, which limit its feasibility in real-time or resource-constrained environments. These findings highlight an inherent trade-off: stronger privacy guarantees typically come at the cost of efficiency, and optimizing model complexity under encryption may reduce predictive accuracy compared to plaintext models.

For future work, we recognize that while runtime and storage overheads have been reported, further validation is required to strengthen claims about real-world scalability. Expanding experiments to larger and more complex datasets will provide deeper insights into robustness. Hardware acceleration such as GPU or FPGA-based implementations and optimized FHE libraries, offer promising pathways to reduce computational bottlenecks. Additionally, hybrid approaches that combine FHE with other privacy-preserving techniques such as secure multi-party computation and differential privacy could improve both efficiency and security, helping bridge the gap between theoretical feasibility and practical deployment of secure machine learning systems.

Acknowledgments. This study was supported by Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA. The authors sincerely appreciate the institutional support that contributed to the success of this research.

References

1. N. Truong, K. Sun, S. Wang, F. Guitton, and Y. K. Guo, "Privacy preservation in federated learning: An insightful survey from the GDPR perspective," *Computers & Security*, vol. 110, Nov. 2021, doi: 10.1016/j.cose.2021.102402.
2. R. Gupta, S. Tanwar, S. Tyagi, and N. Kumar, "Machine Learning Models for Secure Data Analytics: A taxonomy and threat model," Mar. 01, 2020, *Elsevier B.V.* doi: 10.1016/j.comcom.2020.02.008.
3. C. Gouert, D. Mouris, and N. Tsoutsos, "SoK: New Insights into Fully Homomorphic Encryption Libraries via Standardized Benchmarks," *Proceedings on Privacy Enhancing Technologies*, vol. 2023, no. 3, pp. 154–172, Jul. 2023, doi: 10.56553/popets-2023-0075.
4. P. Panzade, D. Takabi, and Z. Cai, "Privacy-Preserving Machine Learning Using Functional Encryption: Opportunities and Challenges," *IEEE Internet Things J*, vol. 11, no. 5, pp. 7436–7446, Mar. 2024, doi: 10.1109/JIOT.2023.3338220.
5. A. V. Kumar, K. Bhavana, and C. Yamini, "Fully Homomorphic Encryption for Data Security Over Cloud," in *6th International Conference on Electronics, Communication and Aerospace Technology, ICECA 2022 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 782–787. doi: 10.1109/ICECA55336.2022.10009404.
6. I. Mustafa *et al.*, "Noise Free Fully Homomorphic Encryption Scheme Over Non-Associative Algebra," *IEEE Access*, vol. 8, pp. 136524–136536, 2020, doi: 10.1109/ACCESS.2020.3007717.
7. N. Jain and A. K. Cherukuri, "Revisiting Fully Homomorphic Encryption Schemes," 2023.

8. M. Afwande, A. Ikoha, and S. Barasa, “Enhancing Research Quality: Emphasizing the Scientific Method, Design Guidelines, and Evidence-Based Assessment,” 2024, doi: 10.51584/IJRIAS.
9. A. C. Mert, E. Ozturk, and E. Savas, “Design and Implementation of Encryption/Decryption Architectures for BFV Homomorphic Encryption Scheme,” *IEEE Trans Very Large Scale Integr VLSI Syst*, vol. 28, no. 2, pp. 353–362, Feb. 2020, doi: 10.1109/TVLSI.2019.2943127.
10. D. D, A. K. K, and R. M, “Research on Homomorphic Encryption for Arithmetic of Approximate Numbers,” in *2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS)*, 2023, pp. 505–509. doi: 10.1109/ICISCoIS56541.2023.10100464.
11. G. M. Raza, Z. S. Butt, S. Latif, and A. Wahid, “Sentiment Analysis on COVID Tweets: An Experimental Analysis on the Impact of Count Vectorizer and TF-IDF on Sentiment Predictions using Deep Learning Models,” in *2021 International Conference on Digital Futures and Transformative Technologies, ICoDT2 2021*, Institute of Electrical and Electronics Engineers Inc., May 2021. doi: 10.1109/ICoDT252288.2021.9441508.
12. A. Ibenegbu, A. Schaeffer, P. L. de Micheaux, and R. Chandra, “A Machine Learning Framework for Handling Unreliable Absence Label and Class Imbalance for Marine Stinger Beaching Prediction,” Jan. 2025, [Online]. Available: http://arxiv.org/abs/2501.11293
13. N. Mitic, A. Pyrgelis, and S. Sav, “How to Privately Tune Hyperparameters in Federated Learning? Insights from a Benchmark Study,” Feb. 2024, [Online]. Available: http://arxiv.org/abs/2402.16087
14. E. Hesamifard, H. Takabi, and M. Ghasemi, “Cryptodl: Deep neural networks over encrypted data,” *arXiv preprint arXiv:1711.05189*, 2017.
15. T. Ishiyama, T. Suzuki, and H. Yamana, “Highly accurate CNN inference using approximate activation functions over homomorphic encryption,” in *2020 IEEE International Conference on Big Data (Big Data)*, IEEE, 2020, pp. 3989–3995.
16. G. Lloret-Talavera *et al.*, “Enabling Homomorphically Encrypted Inference for Large DNN Models,” Mar. 2021, doi: 10.1109/TC.2021.3076123.
17. J. Lee *et al.*, “Optimized Layerwise Approximation for Efficient Private Inference on Fully Homomorphic Encryption,” Oct. 2023, [Online]. Available: http://arxiv.org/abs/2310.10349
18. D. Kim and C. Guyot, “Optimized Privacy-Preserving CNN Inference With Fully Homomorphic Encryption,” *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 2175–2187, 2023, doi: 10.1109/TIFS.2023.3263631.
19. J. W. Lee *et al.*, “Privacy-Preserving Machine Learning With Fully Homomorphic Encryption for Deep Neural Network,” *IEEE Access*, vol. 10, pp. 30039–30054, 2022, doi: 10.1109/ACCESS.2022.3159694.

Human-computer interaction and interface design——Matlab App diffraction simulation

Xinlong Wang[1] and Huanzhen Zhang[1*]

[1] School of Mathematics and Physics, Hebei University of Engineering, Handan 056000, China
*Corresponding author. E-mail addresses: huanwozhenzhen@gmail.com (Huanzhen Zhang)

Abstract. Based on Matlab App Designer, this paper develops three dedicated simulation applications for polygon diffraction screens, white light and complementary screens. Through theoretical analysis, the light intensity distribution formula of different diffraction scenarios is derived, where polygonal diffraction is solved by Fourier optical theory, white light diffraction is simulated through RGB three primary colors, and complementary screen diffraction is verified based on the Babinet principle. The designed interactive interface supports parameter adjustments such as wavelength (390~760nm), aperture size (0.1~1mm), and can display diffraction patterns, light intensity distribution and three-dimensional stereoscopic images in real time. The simulation results show that when the edges of the polygon holes are odd, the edges of the diffraction pattern double, and when even, the same; the stripes of the same order of white light diffraction show a seven-color distribution of purple to red; the complementary screen diffraction patterns satisfy the complementary relationship. This tool solves the problems of expensive traditional experimental equipment and limited parameter adjustment. It can be used for teaching demonstration and scientific research exploration. In the future, it is planned to build a server to expand to use at most terminals.

Keywords: Matlab App Fraunhofer Diffraction Diffraction Simulation Interactive Interface Design Babinet's Principle.

1 Introduction

Diffraction of light refers to the phenomenon that when light waves encounter obstacles during propagation, they can bypass the obstacles and deviate from the straight line and enter the geometric shadow area, and the light intensity is unevenly distributed on the screen [1]. Diffraction is an important research content in fluctuating optics and university physics.At the level of teaching application, traditional diffraction experimental equipment is difficult to meet the needs of large-scale and personalized teaching. On the one hand, a complete set of equipment including helium neon lasers, precision aperture components, etc. starts at tens of thousands of yuan per unit. Small and medium-sized universities or grassroots units with limited funds find it difficult to provide sufficient equipment, and often adopt the "teacher demonstration, student observation" mode. Students cannot adjust parameters

and perceive diffraction changes by themselves, resulting in a disconnect between theory and practice; On the other hand, the adjustment of equipment parameters is limited by mechanical structures. For example, aperture adjustment requires changing the slit/orifice plate and recalibrating the optical path, which takes 15-30 minutes per time. In the classroom, only 3-5 sets of parameter verification can be completed, which cannot cover the multivariate combination of 390-760nm visible light band and 0.1-1mm aperture. Students find it difficult to systematically observe the influence of parameters on diffraction patterns, which restricts their deep understanding of diffraction laws. Optical experimental equipment is expensive, easy to damage, and difficult to operate, and theoretical teaching and experimental links are not synchronized [2].

From the perspective of tool evaluation and comparison with peers, the existing diffraction simulation tools can be mainly divided into two categories: one is command-line simulation programs based on programming languages such as Python and Java, and the other is commercial optical simulation software such as Zemax and LightTools. For the former, although it can achieve the calculation of basic diffraction scenes, it generally lacks a visual interactive interface, has a high operating threshold, and is only suitable for researchers with programming foundation, making it difficult to meet the needs of "real-time adjustment and intuitive observation" in teaching scenarios; For the latter, although commercial software has comprehensive functions and can support the design and simulation of complex optical systems, the software purchase cost is high, with single user authorization fees usually exceeding 10000 yuan, and the operation process is complex, requiring mastery of professional knowledge such as optical system modeling and ray tracing. It is not suitable as a popular tool for basic physics teaching. At the same time, its core algorithms are mostly closed source, and users cannot customize diffraction scenes according to teaching or research needs, such as polygonal holes with special edge numbers, non-standard complementary screens, etc.

With the development of computer technology, numerical simulation methods provide new ways to solve the above problems. Matlab is a programming language software developed and maintained by Math Works [3]. With its powerful function library and interactive programming environment, it shows significant advantages in the field of optical simulation[4].The Matlab App diffraction simulation tool designed in this study demonstrates significant advantages in the evaluation dimension. Developed based on the basic environment of Matlab software, there is no need to purchase additional hardware or pay software licensing fees. For teaching and research institutions equipped with Matlab, it can be directly deployed and used, greatly reducing the application threshold; Compared with traditional command-line tools, it supports slider style adjustment of parameters such as wavelength (390~760nm) and aperture size (0.1~1mm), and can output diffraction patterns, two-dimensional light intensity distribution, and three-dimensional images in real time. It covers three types of complex diffraction scenes: polygons, white light, and complementary screens, making up for the shortcomings of existing tools in supporting complex scenes.

In early research, scholars have used Matlab to realize the Flanghofer diffraction simulation of simple structure diffraction screens such as single-slit, round holes, and rectangular holes. He Haoxuan et al. [5] used Matlab simulation to simulate circular holes and rectangular holes, and observed the relationship between the size of the aperture and the diffraction pattern of the Flanghefei, and found that the smaller the aperture, the more obvious the diffraction phenomenon was. By comparing the relationship between rectangular holes and circular holes that are inward and external circles, the influence of diffraction screen shape on Flanghofer diffraction pattern was studied, and the characteristics of his Flanghofer diffraction pattern were analyzed, and it was found that the diffraction pattern had a large dependence on the shape of the hole. In the case of the similar aperture size, the Flanghofer diffraction of the moment hole was more obvious than the Flanghofer diffraction of the circular hole. Literature [6] uses the difference in refractive index between two points to simulate the diffraction screen, and builds a digital matrix of refractive index. Matlab processing matrix is used to realize the simulation of diffraction diffraction of diffraction screens such as single-slit, double-slit, rectangular holes and circular holes. Cheng Taimin et al. [7] used the Kielhof diffraction theory to deduce the distribution formula of the relative diffraction intensity of Fulanghefei diffraction on the screen of different triangle holes, thereby finding the internal law between the shapes of different triangle holes and the diffraction patterns. In previous studies, although white light diffraction has been reported, research on white light diffraction simulation graphical user interface is rare [8].

For diffraction scenarios with more complex structures, traditional simulation methods still have obvious limitations. On the one hand, polygonal diffraction screens (such as triangles, regular pentagons, and regular hexagons) are difficult to deduce explicit expressions of light intensity distribution through analytical methods due to complex geometric structure, resulting in limited simulation work; on the other hand, white light diffraction involves incoherent superposition of multi-wavelength light, and the formation mechanism of color stripes requires a more complex computational model; in addition, complementary screen diffraction follows an important principle in the Barbinet principle optics, which points out that the diffraction patterns generated by two spatially complementary optical elements complement each other, thus forming a complete diffraction pattern when there is no barrier [9]. The study of complementary screen diffraction has important theoretical and practical application significance in the field of optical, for example, it has certain reference value for the study of lithography mask plates [10]. The existence of these problems makes it difficult for existing simulation tools to meet the needs of comprehensive and systematic research on diffraction phenomena.

The emergence of Matlab App Designer provides technical support for breaking through the above limitations. As an interactive interface design tool in Matlab to replace traditional GUIDE. However, using the Matlab program to simulate optical experiments is difficult for students without programming experience [11]. Based on the above background, we have developed a dedicated Matlab simulation app for three typical scenarios: polygon diffraction screen, white light diffraction and complementary screen diffraction.

This paper conducts the following simulations based on the Matlab App Designer function in the matrix laboratory, including the diffraction of monochrome light Flanghofe round holes, monochrome light Flanghofe single-slit diffraction, white light Flanghofe round holes, white light Flanghofe single-slit diffraction, polygonal flanghofe diffraction of different edges, and complementary screen diffraction patterns comparison. The influence of the above factors on the diffraction pattern and light intensity distribution is studied, and the App editing function in Matlab software is used to create a simple app with adjustable edge number, side length and light wavelength.

2 Theoretical Analysis

2.1 Analysis of the diffraction of polygonal diffraction screens

Assume that the light intensity of the diffraction of the circular hole of Flanghofe at any point P on the screen is

$$I_p=A_0^2[1-\frac{1}{2}m^2+\frac{1}{3}\left(\frac{m^2}{2!}\right)-\frac{1}{4}\left(\frac{m^2}{3!}\right)+\cdots\ ] \tag{1}$$

Where: m=(Rsin)/λ; R is the radius of the circular hole; it is the diffraction angle; λ is the wavelength of incident light.

If formula (1) is represented by first-order Bessel function symbols, then

$$I_p=I_0\frac{J_1^2(2m)}{m^2} \tag{2}$$

The light intensity of Flanghofei single-slit diffraction on the screen is

$$I_p=I_0\frac{\sin^2 u}{u^2}=I_0\sin^2 u \tag{3}$$

Where: u=(πbsin)/λ; b is the spacing between slots; diffraction angle; λ is the wavelength of incident light.

The light intensity distribution of single-slit diffraction is related to the sinc function. The diffraction of rectangular holes can be regarded as the superposition of two single-slit diffractions. The light intensity of a point p on the Flanghofe rectangular hole diffraction screen is

$$I_p=I_0\frac{\sin^2 u}{u^2}\frac{\sin^2 v}{v^2}=I_0\sin^2 u\sin^2 v \tag{4}$$

Where: u=(πbsin θ)/λ; v=(πasinθ)/λ; a is the length of the rectangle; b is the width of the rectangle; θ is the diffraction angle; λ is the wavelength of the incident light. When a is much smaller than b, the diffraction changes from moment hole diffraction to single-slit diffraction.

For polygonal diffraction screens with complex structures (such as triangular holes, regular pentagonal holes, regular hexagons, etc.), the light intensity distribution formula cannot be directly obtained. Using Fourier optical knowledge [12], the intensity distribution of the Flanghofer diffraction formula is derived, and the complex amplitude distribution of the Flanghofer diffraction pattern on the plane is proportional to the frequency spectrum of the object.

Assuming that the amplitude of the plane wave is A, then

$$U(x_0,y_0)=At(x_0,y_0) \tag{5}$$

According to the Flanghofer diffraction formula, the plane field distribution is observed as

$$U(x,y)=\frac{A}{j\lambda z}\exp(jkz)\exp\left[j\frac{k}{2x}(x^2+y^2)\right]T\left(\frac{x}{\lambda x},\frac{y}{\lambda z}\right) \tag{6}$$

f_x,f_yIn the formula is the Fourier transform of the complex amplitude transmittance of the object, also known as the frequency spectrum of the object, and The intensity distribution of the diffraction pattern can be obtained

$$T(f_x,f_y)=\iint_{-\infty}^{+\infty} t(x_0,y_0)\exp[-j2\pi(f_x x_0+f_y y_0)]dx_0dy_0 \tag{7}$$

$$I(x,y)=(\frac{A}{\lambda z})^2\left|T(\frac{x}{\lambda z},\frac{y}{\lambda z})\right|^2 \tag{8}$$

Where: A is the plane wave amplitude; λ is the incident light wavelength; z is the diffraction distance.

2.2 Analysis of white light Flanghofe diffraction

In scalar diffraction theory, the complex amplitude distribution of Flanghofer diffraction can be analyzed by formula (9).

$$U(x,y)=\frac{1}{j\lambda z}\exp(jkz)\exp[j\frac{k}{2z}(x^2+y^2)]\times F\{U(\xi,\eta)\} \tag{9}$$

Where: U(x,y) is the complex amplitude of any point P(x,y) at the receiving screen; $U(\xi,\eta)$ is the complex amplitude of any point $P_0(\xi,\eta)$ at the upper; λ is the wavelength of the incident light; z is the distance between the diffraction screen and the receiving screen; k is the wave loss of the light wave; j is the imaginary unit.

For the Flanghofe circular hole diffraction, the light wave with the incident diffraction aperture is parallel light, and the transmittance function of the circular hole is

$$t(\rho)=\mathrm{circ}(\frac{\rho}{R}) \tag{10}$$

$\rho=\sqrt{\xi^2+\eta^2}$Where is the distance between $P(\xi,\eta)$ on the diffraction screen and the center of the circle; R is the radius of the circular hole. Due to the axial symmetry of this function, substituting equation (9) can obtain the light intensity distribution on the receiving screen as

$$I(r)=U(r)^2=(\frac{ka^2}{z})^2[\frac{J_1(kar/z)}{kar/z}]^2 \tag{11}$$

J_1 In the formula, $J\nu(z)$ represents the first-order Bessel function, which can be simulated using the built-in Matlab function besselj(nu, Z).

For Flanghofer single-slit diffraction, assuming that the width of the single-slit is a, the transmittance function of the single-slit is

$$t(\xi)=\mathrm{rect}\left(\frac{\xi}{a}\right) \tag{12}$$

Substitute the transmittance function formula (12) into formula (9) to obtain the light intensity distribution on the receiving screen as

$$I(x)=U(x)^2=I_0[\sin c(\frac{\pi a}{\lambda}\sin\theta)]^2 \tag{13}$$

Where: I_0 is the light intensity at the center point of the diffraction screen; λ is the wavelength of the incident light.

2.3 Complementary screen diffraction analysis

In order to more accurately describe the transmission characteristics of the light beam, the parameter Fresnel number is introduced in the diffraction theory to determine the region of action of the diffraction effect. Assuming that incident light with wavelength passes through a diffraction hole with size a (a can be half of the edge length of the square hole or the radius of the circular hole), it reaches the screen after the propagation distance L. At this time, the expression of the Fresnel number is

$$N_F=\frac{a^2}{L\lambda} \tag{14}$$

When N,<1, the diffraction type is Flanghofer diffraction, that is, far-field diffraction. The far-field diffraction image of the hole is displayed on the screen, which is related to the Fourier transform of the spatial complex amplitude after the light field passes through the hole diameter. When $N,\geq 1$, the diffraction type is Fresnel diffraction, that is, near-field diffraction. If the Fresnel number and diffraction angle are not large, Fresnel approximation can be used. When N,>1, theories in geometric optics can be applied.

$U_1(P)U_2(P)U(P)$When the irradiation light source is consistent, assuming that the complex amplitude of the diffraction field generated by one of the complementary screens separately at the observation point P, it indicates that the complex amplitude of the observation point P is present when there is no diffraction screen, and there is

$$U_1(P)+U_2(P)=U(P) \tag{15}$$

Equation (15) indicates that in a set of corresponding complementary screens, the sum of the complex amplitudes of the diffraction light fields generated separately from the complex amplitudes when the diffraction screen is not placed. This principle is called the Babinet principle. Based on this principle, the following conclusions are drawn.

$U_1(P)$When a diffraction screen is placed, the light intensity of some points in the diffraction field is 0, and the corresponding complementary screen is the same as that of the diffraction screen at the above points, that is, when =0,

$$U_2(P)=U(P) \tag{16}$$

$U(P)U_1(P)$、$U_2(P)$At the point =0, there is a phase difference of magnitude π, and the light intensity is satisfied at this time

$$I_1(P)=|U_1(P)|^2=I_2(P)=|U_2(P)|^2 \tag{17}$$

$U(P)=0U(P)=0$ At the point, the light intensity distribution generated by the two complementary screens when diffraction alone is the same. Then deduce, satisfying

$$U_1(P)=-U_2(P) \tag{18}$$

3 Design of Matlab App simulation for different diffraction types

3.1 Polygonal diffraction screen Fulanghefei diffraction simulation app design and function

Use MATLAB's graphical user interface (GUI) to design an application based on the generated code. Figure 1 shows the demonstration interface of the App.

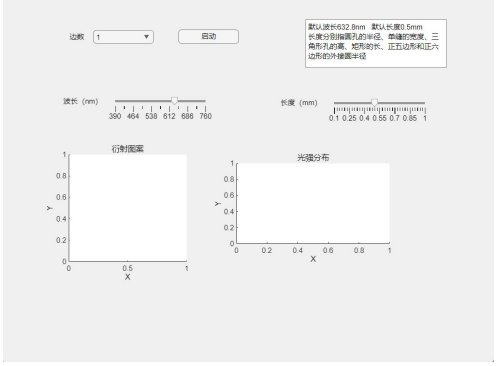


Fig. 1. App demonstration interface

In this program, the wavelength of the light wave and the diffraction aperture size are set to variables that can be adjusted freely, with the adjustable wavelength range between 390 and 760 nm, and the default wavelength is 632.8 nm (the wavelength of the helium-neon laser); the adjustable range of the diffraction aperture is 0.1 and 1 mm, and the default size is 0.5 mm. The program has two output areas, which are used to display the diffraction pattern and the light intensity distribution pattern respectively. The vertical coordinates of these two figures are relative light intensity, while the horizontal coordinates are the relative positions of the diffraction distribution. The specific position is determined by the focal length f of the actual lens used, and it conforms to the law that the actual light intensity distribution position is the product of the focal length f and the diffraction angle tangent value tan, that is, x=f·tan.

It can be observed from Figure 2 that after the number of sides exceeds a certain value, the diffraction pattern shows obvious regularity: when the number of sides of the diffraction hole is odd, the number of sides of the diffraction pattern is twice that of the hole; and when the number of sides of the hole is even, the number of sides of the diffraction pattern is the same as the number of sides of the hole. At the same time, it can be found that the light intensity distribution also follows this law. The central zero-level stripe concentrates most of the light intensity, and on the left and

right sides of it, sub-maximum and extremely small stripes with weaker light intensity are distributed.

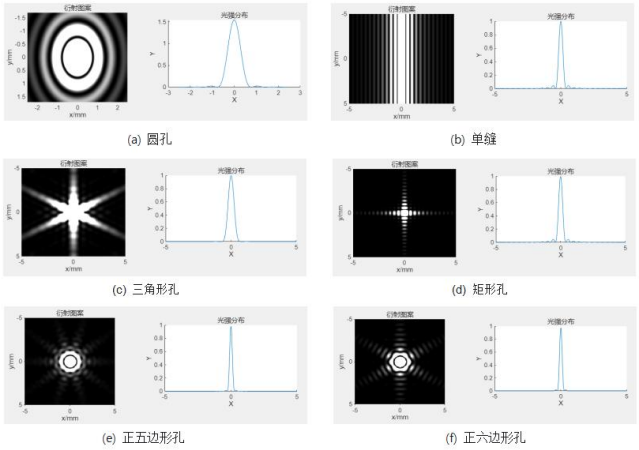


Fig. 2. Results of diffraction simulation of polygonal holes

Below, take the Flanghofe circular hole diffraction with the edge number "1" as an example, and simulate the influence of light wavelength on its diffraction pattern and light intensity distribution. Set the light wavelength and aperture size to be the default values. The pattern and light intensity distribution diagram of circular hole diffraction can be observed in the coordinate area. The results are shown in Figure 3(1). In the center are bright stripes with higher intensity, and circular stripes with alternating light and darkness are symmetrically distributed on both sides. The bright stripes on both sides have the same width, and the bright stripes in the center are twice as wide as those at other levels of bright stripes. It can also be seen from the light intensity distribution map that most of the diffraction light intensity is concentrated in the central bright stripes.

When the diffraction hole width remains unchanged at 0.5mm, if the light wavelength is adjusted and the wavelength of light is changed from 632.8nm to 390nm, the width of the central bright pattern will obviously narrow. This change can be clearly seen by comparing (1) and (2) in Figure 3.

This result can be verified by formula (3). From formula (3), we can find that when the central main extreme appears, sin=0, and ksin≈, which is the diffraction angle and b is the slit width. This can be derived from the half width of the central stripe, where f is the lens focal length. This half-width is proportional to the light wavelength, and this conclusion is consistent with the results of the simulation experiment, indicating that the simulation results are accurate.

Taking the triangular Kolfanghof diffraction as an example, we can further study the influence of aperture size on diffraction pattern and light intensity distribution. First adjust the edge number to "3" and then click to start, and the diffraction pattern and light intensity distribution of the triangle hole at the default wavelength are shown

in Figure 4. Keep the light wavelength unchanged, adjust the diffraction hole size to 1mm, and then observe the output image, as shown in Figure 5.

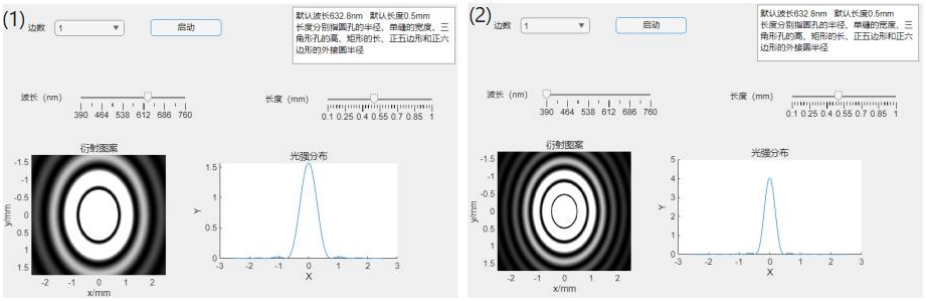


Fig. 3. Fraunhofer diffraction output interface ((1) with default wavelength and length, (2) with a wavelength of 390nm and default length)

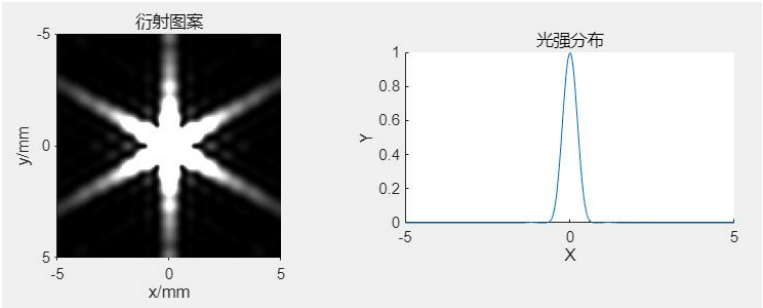


Fig. 4. Triangular hole diffraction output interface (wavelength and length are default values)

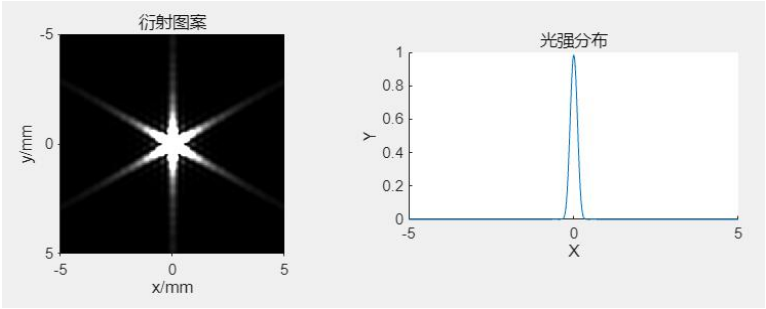


Fig. 5. Triangular hole diffraction output interface (wavelength is default, length is 1mm)

It can be observed that the diffraction phenomenon is more obvious when the aperture size is 0.5 mm. When the aperture size is adjusted to 1 mm, the diffraction pattern distribution area becomes smaller and the light intensity distribution becomes more concentrated. It can be pushed to reduce the diffraction phenomenon when the hole size tends to be infinitely large.

3.2 White light diffraction simulation app design and functions

The initial graphical user interface created in Matlab is as shown in Figure 6.

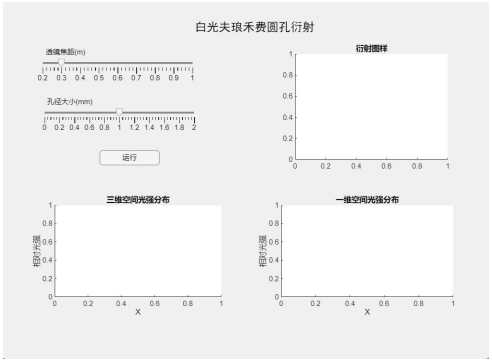
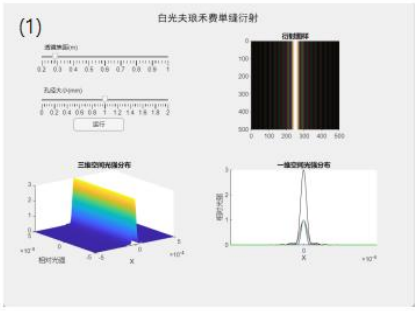


Fig. 6. Simulation interface design

To facilitate quick adjustment of each parameter, a slider is set to input the lens focal length parameters, the adjustable range is 0.2~1m and the initial focal length is 0.3m. Set a slider to input aperture size parameter, the adjustable range is 0~2mm, and the initial aperture size is 1mm. In a circular hole, the aperture size is the radius of the circular hole, and in a single slot, the aperture size is the slot width.

In Figure 7 (1) shows the simulation results of the diffraction screen shape as a single slot, the lens focal length is 0.28 m, and the single slot width is 1 mm. (2) The simulation results of the diffraction screen shown are circular holes, the lens focal length is 0.44m, and the circular hole radius is 0.8 mm.

Figure 8 shows an enlarged partial screenshot of the diffraction pattern of the Flanghofe round hole in Figure 7. In the first-order stripes of the diffraction pattern, seven-color lights from the inner and outer purple, blue, blue, green, yellow, orange and red can be observed, which conforms to the phenomenon after white light spectroscopy. The light with a short wavelength is close to the center zero-order diffraction stripe, and the light with a long wavelength is far away from the center zero-order diffraction stripe. Purple stripes are a mixture of blue light and the previous level of red light, cyan stripes are a mixture of blue light and green light, and orange stripes and yellow stripes are a mixture of green light and red light. Similar distributions can also be observed in the diffraction pattern of single slits.



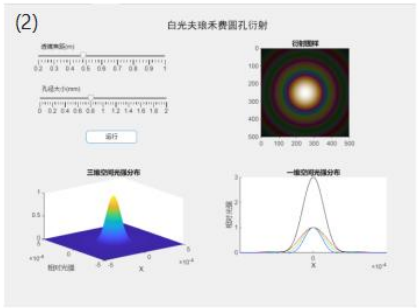


Figure 7 Flanghofer diffraction simulation results ((1) single slit, (2) round hole)

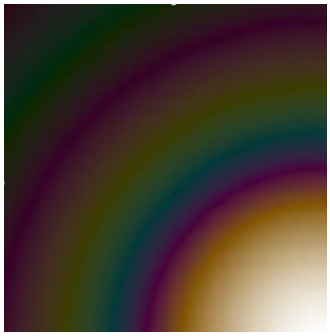


Figure 8. Amplified circular hole diffraction pattern

3.3 Complementary screen diffraction simulation App design and functions

An App is built in Matlab App Designer. The main interface consists of three areas, as shown in Figure 9, namely the display area, the control area, and the operation area. Each area corresponds to its corresponding functions in the App. The default parameter of the slider is set to the incident light wavelength of 632.8 nm, which is the light wavelength of the helium-neon laser in diffraction experiments. When the Fresnel number is less than 1, Flanghofer diffraction occurs. Take the default parameters with the Fresnel number 0.2, with the diffraction distance of 316 nm, the small hole radius is 0.2 mm, and the slit width is 0.4 mm.

The display area contains four coordinate axes objects. The display area displays the diffraction pattern and three-dimensional light intensity distribution of the selected aperture type, as well as the diffraction pattern and three-dimensional light intensity distribution of the complementary screen diffraction corresponding to the selected aperture type on the four coordinate axes objects.

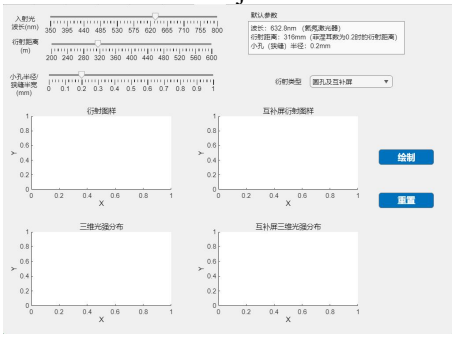


Fig. 9. App main interface layout

Simulation results of complementary screen diffraction experiment: the incident light wavelength is λ=632.8nm, so that the diffraction distance L=316mm, the radius

of the small hole and the half-width of the slit a=0.2mm, at this time, the Fresnel number is N, =0.2<1, and the diffraction type can be determined by the Fresnel number definition as Flanghofer diffraction, and the simulation is performed under this parameter condition. Based on the above parameters, complementary screen simulations were performed on the circular hole and single slot respectively. The simulation results are shown in Figure 10.

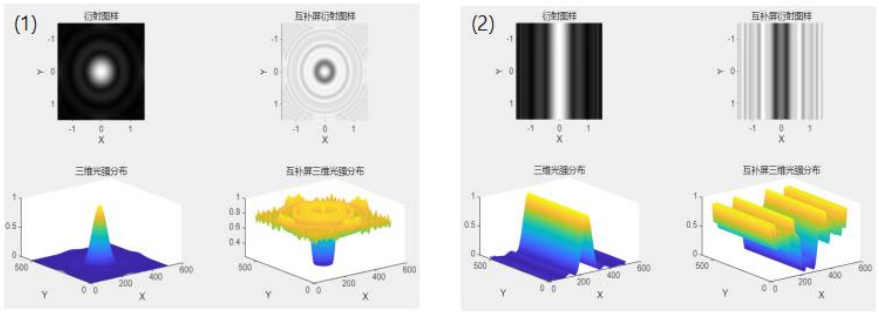


Fig. 10. Complementary screen diffraction simulation results ((1) round hole (2) single slit)

4 Simulation Conclusions

Through Matlab simulation, this paper studies the diffraction law of diffraction screens of different edge numbers, diffraction law of round holes and single-slits of white light, and diffraction pattern law of diffraction, and obtains the following conclusions. For the Flanghofer diffraction screen diffraction of different sides, there are:

(1) When the number of hole edges ≥3, the number of edges of the diffraction hole is odd, and the number of edges of the diffraction pattern is twice the number of edges of the hole; when the number of edges of the hole is even, the number of edges of the diffraction pattern and the number of edges of the hole edges are equal.

(2) The size of the diffraction pattern is proportional to the light wavelength, and as the light wavelength increases, the diffraction phenomenon is more obvious.

(3) The smaller the diffraction hole, the more obvious the diffraction phenomenon is, the more sparse the stripes of the diffraction pattern, showing a diffraction inverse ratio.

For white light Flanghe Feiyan, there are: in the diffraction stripes of the same order, the seven-color light distributions of purple, blue, teal, green, yellow, orange and red appear from the inside and outside. The longer the wavelength, the more obvious the diffraction phenomenon, which is consistent with the white light diffraction experiment phenomenon.

For the comparison of complementary screen diffraction, under the condition of Flanghofer diffraction, the complementary screen diffraction patterns complement

each other to form a complete diffraction pattern without barriers, confirming the Babinet principle.

5 Summary and Outlook

The above three diffraction simulation applications based on Matlab App are designed for polygonal diffraction, white light diffraction and complementary screen diffraction. Through intuitive interactive interfaces and real-time simulation results, they provide effective tools for understanding different diffraction phenomena. It provides help to understand the diffraction field rules of diffraction holes in different shapes, and visualize and quantify the shape and light intensity distribution of diffraction stripes. This simulation platform can facilitate teaching and experimental demonstration, and also enriches typical cases of diffraction experiments.
In the future, servers can be further built so that simulation systems can be used on computers or mobile phones through browsers, and promote the wider application of virtual simulation technology in physics teaching and optical research.

6 [References]

1. Yao Qijun, East China Normal University's "Optical" textbook writing group. Optical tutorial [M].2 edition. Beijing: Higher Education Press, 1989:94-151.
2. Yu Xinja, Wang Meijiao, Zhao Qiangjie. Design and Implementation of an Optical Experiment Virtual Simulation System Based on Matlab App Designer [J]. Experimental Science and Technology, 2022, 20 (1): 45-50.
3. Su Xunyu. Design and Application of Matlab in High School Physics Information-based Teaching [D]. Dalian: Liaoning Normal University, 2023.
4. Tan Yi. Simulation study of diffraction pattern of mesongfulanghefei[J]. Laboratory Research and Exploration, 2013, 32(7):41-42,125.
5. He Haoxuan, Shi Qianzhi, Yang Sha, et al. Comparative Study on the MATLAB Simulation of Fraunhofer Diffraction from Circular and Rectangular Holes[J]. Science and Technology Breeze, 2022(3): 62-65.
6. GASCÓN F, SALAZAR F. A simple method to simulate diffraction and speckle patterns with a PC[J]. Optik, 2006, 117(2): 49-57.
7. Cheng Taimin, Cao Liangang. Simulation of Fraunhofer Diffraction Patterns of Different Triangle Holes[J]. Guangxi Physics, 2010, 31(1): 33-36.
8. ZHANG J H,SU R Z.Chromaticity analysis of diffraction pattern of white light under the grating[J].Optik,2020,208:164553.
9. Xu Bo. Diffraction of complementary screens [J]. Huaihua Teachers College Natural Science Journal, 1987 (6): 64-70.
10. ERDMANN A,SHAO F,AGUDELOV,etal. Modeling of mask diffraction and projection imaging for advanced optical and EUV lithography [J].J Mod Opt, 2011,58(5/6):480- 495.
11. Song Lu, Wei Yabó, Feng Yanping. Research on the Fraunhofer Diffraction Simulation System Based on Matlab GUI[J]. Computer and Digital Engineering, 2019, 47(7): 1734-1737.
12. Lü Nai-Guang. Fourier Optics[M]. 3rd ed. Beijing: Machinery Industry Press, 2016.

AI Assisted Grading Framework for Thai-Language Written Exam Questions based on LLM and Rule-Based Reasoning Approach

Chutipon Trirattnanurak[1], Sasiporn Usanavasin[2], Chaianun Damrongrat[3], Manubu Okumura[4]

[1] School of Information, Computer, and Communication Technology (ICT), Sirindhorn International Institute of Technology, Thammasat University, Pathum Thani, Thailand
[2] School of Information, Computer, and Communication Technology (ICT), Sirindhorn International Institute of Technology, Thammasat University, Pathum Thani, Thailand
[3] National Science and Technology Development Agency (NSTDA), Pathum Thani, Thailand
[4] Institute of Science Tokyo, Japan

m6722040786@g.siit.tu.ac.th

Abstract. This research investigates the feasibility of using Large Language Models (LLMs) to assist in grading Thai written exam questions classified by Bloom's Taxonomy. A dataset of 81 written questions from the Royal Institute of Thailand was used, with detailed rubrics created and validated by professional Thai teachers. Five commercial LLMs including ChatGPT, Typhoon, Gemini, Grok, and Claude were employed to simulate student answers and provide scores ranging from 0 to 3, which were then reviewed and refined by experts. The automated grading showed substantial agreement with human graders, achieving an overall score agreement rate of 82.4% and a Cohen's kappa of 0.78, indicating strong reliability. Thematic analysis of LLM-generated explanations revealed that 87% met high standards of clarity and pedagogical soundness. Error analysis identified challenges primarily in higher-order cognitive tasks such as analyzing and creating. Comparisons between rubric-based and non-rubric-based grading highlighted the importance of well-defined rubrics for consistent scoring. This study demonstrates the potential of LLMs to support Thai teachers by reducing grading workload and encouraging more widespread use of written exams, which better assess student understanding. Future work will focus on system automation and fine-tuning with larger datasets to further enhance grading accuracy and explanatory quality.

Keywords: Large Language Models (LLMs), Bloom's Taxonomy, Thai Written Exams, Educational Assessment, Natural Language Processing (NLP)

1 Background

Exam evaluation is essential for assessing student comprehension and informing improvements in teaching strategies and curriculum design [1]. Written exams are particularly effective for measuring higher-order skills such as critical reasoning, logical

explanation, and language proficiency. However, grading them is time-consuming and often inconsistent due to fatigue, emotional bias, and subjective interpretation [2]. While Teaching Assistants (TAs) may help reduce workload, variations in their training and understanding of grading criteria can introduce further inconsistency.

In the Thai language context, grading is even more challenging due to complex grammar, tonal nuances, and numerous context-dependent exceptions [3]. These linguistic features, coupled with the open-ended nature of responses, increase the risk of misinterpretation and bias, requiring both linguistic expertise and an understanding of students' reasoning. As a result, many Thai teachers prefer multiple-choice exams, which are faster to grade. Teachers typically handle 100–400 students per subject, teach 7–21 hours weekly, and manage substantial administrative and extracurricular duties [4], leaving limited time for detailed evaluation of written work.

Although multiple-choice questions are efficient, they cannot fully capture students' reasoning and expressive skills. Written exams, while more resource-intensive, offer richer insights into understanding. Recent advances in Natural Language Processing (NLP), particularly with Large Language Models (LLMs), enable sophisticated text analysis for tasks such as tokenization, summarization, and automated evaluation [5]. Leveraging these technologies can make grading Thai written exams fairer, faster, and more consistent, encouraging wider adoption of open-ended assessments that better reflect student learning.

2 Literature Review

2.1 The Integration of LLMs in Educational Assessment

Large Language Models (LLMs) are increasingly explored for enhancing subjective and open-ended evaluation. Holmes et al. [6] highlight their potential for automated feedback and assessment, alongside challenges of cultural bias and fairness in underrepresented languages. Yu-Ju Lan [7] notes their dual role as tools for personalized learning and threats to academic integrity. Ryznar [8] warns against misuse for bypassing genuine learning, recommending exam redesigns and time-limited formats. These works underscore both promise and pitfalls, motivating efforts toward trustworthy, context-aware assessment tools.

2.2 Leveraging LLMs for Text Understanding and Cognitive Inference

Beyond text generation, LLMs have shown mixed results in deeper semantic and logical reasoning. Jang et al. [9] found GPT-4 strong in syntactic tasks but weaker with negation and contradiction, suggesting limitations of a single-model approach. Lin et al. [10] demonstrated that fine-tuning models like LLaMA-3 (via LoRA) enables specialized cognitive tasks such as sentiment intensity regression. While fine-tuning is beyond our scope, these findings support using Bloom's taxonomy to match models to cognitive levels rather than relying on a one-size-fits-all method.

2.3 LLM Ensembles and Model Selection Strategies

Ensemble strategies can mitigate individual model weaknesses. Sakai and Lam [11] introduced QUAD-LLM-MLTC for healthcare classification, showing that combining models improves semantic accuracy. Tekin et al. [12] proposed LLM-TOPLA, clustering outputs to enhance robustness and efficiency. Our method similarly seeks the best-performing LLM per Bloom's cognitive category, reflecting that no single model excels universally.

2.4 Bloom's Taxonomy in Educational AI and Cognitive Classification

Bloom's taxonomy provides a structured framework for aligning cognitive theory with AI assessment. It categorizes cognitive skills into six levels: Remembering (recalling facts), Understanding (explaining ideas), Applying (using knowledge in new situations), Analyzing (breaking down information into components), Evaluating (justifying decisions or judgments), and Creating (producing new or original work). Abduljabbar and Omar [13] classified questions into Bloom's levels via machine learning, demonstrating the value of cognitive categorization. Rodrigo and Peñas [14] emphasized inter-rater agreement and gold standards for improved question-answering systems. Dodia et al. [15] combined rule-based and ML methods for short-answer grading, while Hubbard et al. [16] showed how question format affects depth of student responses.

Building on these, this study targets open-ended written exam questions tasks that require image-based prompts or interpretation, multiple-choice questions that can be fully evaluated through rule-based checks, and large-scale text summarization tasks where token length may introduce processing errors. This focus ensures that the evaluation remains within the domain of short opinion-based or fixed written responses, where both AI and rule-based methods can be applied effectively under the Bloom's taxonomy framework.

3 Methodology

This study evaluates the effectiveness of Large Language Models (LLMs) in grading Thai-language written questions, aligning results with Bloom's Taxonomy and expert human judgment. The process comprises five phases: question classification, dataset preparation, system setup, comparative evaluation, and data analysis.

3.1 Classifying Exam Questions based on Bloom's Taxonomy

All of our written questions were categorized into Bloom's six cognitive levels by two experienced Thai educators. Independent classifications were reconciled through consensus, ensuring pedagogical alignment for subsequent model evaluation.

3.2 Dataset Collection and Preparation

The exam questions are divided into two sets:

1. Rubric-Based Set: Rubrics were created per question using ChatGPT-extracted keywords (“ความถูกต้องทางภาษา” [language accuracy], “ทักษะการสื่อสาร” [communication skills]), verified by expert teachers.

“โปรดระบุคีย์เวิร์ดหลัก 3 คำจากคำถามต่อไปนี้
เพื่อใช้เป็นเกณฑ์ในการประเมินและตรวจคำตอบ”

Fig. 1. Sample prompt used to extract key rubric keywords from Thai exam questions via ChatGPT (“โปรดระบุคีย์เวิร์ดหลัก 3 คำจากคำถามต่อไปนี้ เพื่อใช้เป็นเกณฑ์ในการประเมินและตรวจคำตอบ” - Please specify three main keywords from the following question to be used as criteria for evaluating and scoring the answer).

2. Non-Rubric Set: This set skips the rubric generation step to observe the LLMs’ raw interpretive grading ability without explicit criteria.

Five LLMs (ChatGPT, Typhoon, Gemini, Grok, Claude) generated 10 sample student responses per question, each scored 0–3 with explanations. Representative answers for each score level were refined by professional teachers and stored in Excel for contextual retrieval during grading. The prompt used was:

“โปรดสร้างตัวอย่างคำตอบของนักเรียนจำนวน 10 รูปแบบ สำหรับคำถามหนึ่งข้อ
โดยแต่ละคำตอบควรได้รับคะแนนตั้งแต่ 0 ถึง 3 จากคะแนนเต็ม 3 คะแนน (อย่างหลากหลาย)
ให้แสดงผลลัพธ์ในรูปแบบ plain text โดยแต่ละตัวอย่างต้องประกอบด้วย:
1. คำตอบของนักเรียน (แสดงให้ชัดเจน)
2. คะแนนที่ให้ (ระบุเป็นตัวเลข 0–3)
3. คำอธิบายเหตุผลที่ให้คะแนนนั้น (อธิบายให้ชัดเจน ไม่ใช้ bullet หรือหัวข้อย่อหน้า)”
The Exam Question
The Rubric (Non-Rubric will skip this)

Fig. 2. Prompt used to generate student answers and scores from LLMs (“โปรดสร้างตัวอย่างคำตอบของนักเรียนจำนวน 10 รูปแบบ สำหรับคำถามหนึ่งข้อ โดยแต่ละคำตอบควรได้รับคะแนนตั้งแต่ 0 ถึง 3 จากคะแนนเต็ม 3 คะแนน (อย่างหลากหลาย) ให้แสดงผลลัพธ์ในรูปแบบ plain text โดยแต่ละตัวอย่างต้องประกอบด้วย: 1.คำตอบของนักเรียน (แสดงให้ชัดเจน) 2.คะแนนที่ให้ (ระบุเป็นตัวเลข 0-3) 3.คำอธิบายเหตุผลที่ให้คะแนนนั้น (อธิบายให้ชัดเจน ไม่ใช้ bullet หรือหัวข้อย่อหน้า)” - Generate 10 variations of student answers for a single question, with scores ranging from 0 to 3 out of 3, and present them in plain text. Each example must include: (1) the student’s answer, (2) the numerical score (0–3), and (3) a clear explanation of the scoring rationale without using bullet points or headings).

3.3 Grading System Design and LLM Prompting

For grading, LLMs were prompted with question, rubric, and sample responses (Fig. 3). The best-performing model per Bloom level was selected for test grading. Where LLMs struggled with Thai-specific linguistic nuances, custom Python modules using pythainlp were applied for syllable counting, segmentation, and exception handling (e.g., “พลายงาม” segmented as “พลาย-งาม”). Irregular forms were stored in Excel for retrieval during evaluation.

"โปรดให้คะแนนคำตอบของนักเรียน สำหรับคำถามหนึ่งข้อ โดยคำตอบควรได้รับคะแนนตั้งแต่ 0 ถึง 3 จากคะแนนเต็ม 3 คะแนน
ให้แสดงผลลัพธ์ในรูปแบบ plain text โดยแต่ละตัวอย่างต้องประกอบด้วย:
1. คะแนนที่ให้ (ระบุเป็นตัวเลข 0–3)
2. คำอธิบายเหตุผลที่ให้คะแนนนั้น (อธิบายให้ชัดเจน ไม่ใช้ bullet หรือหัวข้อย่อหน้า)"
The Exam Question
The Rubric (If Applicable)
Reference answers retrieved from the training set as context (for rubric-based set)

Fig. 3. Prompt template for grading a student's response with context, rubric, and question provided to the LLM ("โปรดให้คะแนนคำตอบของนักเรียน สำหรับคำถามหนึ่งข้อ โดยคำตอบควรได้รับคะแนนตั้งแต่ 0 ถึง 3 จากคะแนนเต็ม 3 คะแนน ให้แสดงผลลัพธ์ในรูปแบบ plain text โดยแต่ละตัวอย่างต้องประกอบด้วย: 1.คะแนนที่ให้ (ระบุเป็นตัวเลข 0-3) 2.คำอธิบายเหตุผลที่ให้คะแนนนั้น (อธิบายให้ชัดเจน ไม่ใช้ bullet หรือหัวข้อย่อหน้า)" – Grade the student's answer for a single question, assigning a score from 0 to 3 out of 3. Present the output in plain text, including: (1) the numerical score (0–3), and (2) a clear explanation of the scoring rationale without bullet points or headings).

3.4 Grading Evaluation by Thai Language Teachers

To assess the system's effectiveness, a test set prepared following the same procedure as the experimental dataset was graded by two Thai teachers, one rubric-bound, the other applying professional judgment and compared to the proposed system's outputs. Alignment was assessed via Cohen's Kappa and score agreement rate. Discrepancies were classified as rubric ambiguity, LLM misinterpretation, or Thai linguistic complexity.

3.5 Data Analysis and Model Selection

Analysis determined: (1) the LLM best aligned with human grading per Bloom level; (2) question types requiring algorithmic intervention; and (3) score levels prone to divergence. Errors were categorized (underscoring, overscoring, partial-credit inconsistency, rubric drift). Qualitative thematic analysis of LLM explanations assessed reasoning and pedagogical validity. Rubric-based prompting improved intra-model consistency, but higher-order tasks (Analyzing, Creating) showed greater disagreement. Fallback heuristics and refined prompts were applied where explanations were ambiguous or inaccurate.

3.6 Rubric Design Considerations

Rubrics are essential for accurate, consistent LLM-based grading, as LLMs rely entirely on prompt and rubric clarity rather than subjective judgment. In this study, rubrics were designed to be mutually exclusive (each score level tied to distinct conditions) and collectively exhaustive (covering all responses). For instance, a question on the four smallest prime numbers would have clear criteria for each score. Such structure minimizes ambiguity, prevents overlap, and improves grading reliability [1].

"If the answer includes 2, 3, 5, 7 → score 3;
If 3 of them → score 2;
If 1–2 → score 1;
Otherwise → score 0."

Fig. 4. Rubric example illustrating scoring criteria for a question on identifying prime numbers.

4 Experiments and Result Discussion

4.1 Dataset Description

From the Royal Institute of Thailand's Thai exam dataset (826 questions), 745 multiple-choice items were removed, leaving 81 written questions. A professional Thai teacher classified each question into Bloom's Taxonomy. The training set comprised 94 labeled items (due to sub-questions): Remembering (18), Understanding (10), Applying (14), Analyzing (26), Evaluating (11), and Creating (15). The test set contained 55 questions: Remembering (10), Understanding (14), Applying (8), Analyzing (9), Evaluating (7), and Creating (7). For both sets, two parallel versions, rubric and non-rubric were prepared with identical questions.

4.2 Rubric Generation and Validation

Following Lee and Song [1], ChatGPT extracted concise, criterion-based rubrics for each question. A professional Thai teacher reviewed all rubrics for accuracy, clarity, and alignment with learning outcomes. In the training set, 40/81 rubrics (49.38%) were accepted as-is and 41 (50.62%) revised; in the test set, 30/55 (54.55%) were accepted and 25 (45.45%) adjusted. Common corrections involved clarifying ambiguous wording, ensuring mutual exclusivity, and adding conditions for partially correct answers. Although ChatGPT can be used to generate rubrics but Teacher validation is necessary and critical to minimize LLM bias and to ensure the rubrics accurately reflect learning objectives and maintain consistency.

4.3 Student Answer Simulation via LLMs

Five commercial LLMs—ChatGPT, Typhoon, Gemini, Grok, and Claude—each generated four answers per question, one for each score level (0–3) based on the validated rubrics. For example, for Question A, ChatGPT produced answers for scores 0–3, yielding 20 responses per question. Across 81 questions, both rubric and non-rubric sets produced 1,620 responses each. All LLM-generated answers were reviewed by a professional Thai teacher, correcting any scores that conflicted with the validated rubrics. This ensured that both rubric creation and answer generation incorporated expert oversight and aligned with intended learning outcomes.

4.3.1 Non-Rubric Result

Due to time constraints, the teacher validated only the first 340 responses in the non-rubric set. Of these, 66 (19.41%) required adjustment. Table 1 presents detailed classification metrics (precision, recall, F1-score) for each LLM in this subset.

Table 1. Performance by Model (Non-Rubric Set, 340 Responses Validated)

Model	Accuracy	Macro F1	Notable Observation
ChatGPT	0.66	0.63	Struggled with score 3 over-prediction; low precision for top score.
Typhoon	0.69	0.69	Balanced performance, but inconsistent on mid-tier scores.
Gemini	0.93	0.93	Highest accuracy; strong consistency across all score levels.
Grok	0.87	0.87	Stable across classes; slightly weaker recall for score 2.
Claude	0.88	0.88	Similar to Grok; well-balanced performance.
Overall	0.81	0.80	Average performance; variability in mid- and high scores.

The pattern in figure 5 suggests that LLMs tend to associate longer responses with higher scores and shorter responses with lower scores. However, in real-world grading, concise answers can still earn full marks if they cover all required points. This discrepancy highlights a potential bias in LLM scoring toward verbosity.

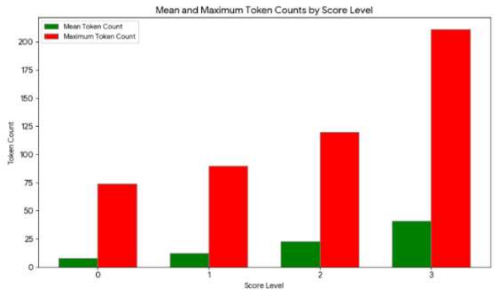


Fig. 5. Token length analysis of answers by assigned score in non-rubric set showing longer answers generally receive higher scores, suggesting a verbosity bias in LLM scoring.

4.3.2 Rubric Result

LLM-generated rubric scores were compared against teacher-validated scores to assess alignment. Precision, recall, and F1-score were calculated for each rubric score (0–3) using teacher scores as ground truth. Accuracy ranged from 91% (Typhoon) to 98% (Gemini) (Table 2). Gemini achieved the highest accuracy (0.98) with perfect recall (1.00) for score 0.0. ChatGPT, Grok, and Claude each scored ~0.97, while Typhoon had the lowest (0.91) due to reduced recall for scores 1.0 and 2.0.

Table 2. LLM Performance on Rubric-Based Student Answer Evaluation

Model	Accuracy	Macro F1	Notable Observation
ChatGPT	0.97	0.97	High agreement with teacher scores; slight underscoring on score 2
Typhoon	0.91	0.90	Lower overall; some missed low/mid scores
Gemini	0.98	0.98	Excellent alignment; nearly perfect scoring
Grok	0.97	0.97	Consistent across all score levels
Claude	0.97	0.97	Stable scoring; minor variations in mid scores
Overall	0.88	0.88	Strong general alignment with teacher validation

When comparing Tables 1 and 2, Gemini shows unusually strong performance even in the non-rubric condition (0.93 accuracy), which narrows the apparent improvement gained from rubric guidance (0.98). This is partly explained by the limited validation sample in the non-rubric set (340 responses), where Gemini exhibited stable predictions that aligned well with teacher scoring. However, when scaled to rubric-based evaluation across the full dataset, rubric use consistently improved overall alignment and reduced misclassification in higher-order tasks. To make the contrast clearer, Table 3 provides a consolidated comparison of non-rubric and rubric performance against human grading. This highlights that rubrics not only improve reliability across most LLMs but also reduce biases such as verbosity preference.

Table 3. Consolidated Comparison of Non-Rubric and Rubric Performance

Model	Non-Rubric Accuracy	Non-Rubric Macro F1	Rubric Accuracy	Rubric Macro F1	Human Baseline
ChatGPT	0.66	0.63	0.97	0.97	1.0
Typhoon	0.69	0.69	0.91	0.90	1.0
Gemini	0.93	0.93	0.98	0.98	1.0
Grok	0.87	0.87	0.97	0.97	1.0
Claude	0.88	0.88	0.97	0.97	1.0

Teacher validation revealed scoring errors and reasoning deficiencies in the LLM outputs. A total of 65 scoring errors were identified: Typhoon 29, Grok 10, Claude 10, ChatGPT 9, and Gemini 7. There were 8 reasoning deficiencies: ChatGPT 4, Typhoon 2, Gemini 1, Grok 1, and Claude none. Performance varied across Bloom's taxonomy levels, with Claude consistently leading or tying for first, Gemini and ChatGPT performing strongly across most categories, and Typhoon generally scoring lower. Detailed results are summarized in Table 4.

Table 4. LLM Performance Across Bloom's Taxonomy Categories

Bloom's Level	ChatGPT	Typhoon	Gemini	Grok	Claude	Explanation
Remembering	97.06%	83.82%	98.53%	97.06%	100.00%	Claude perfect; Typhoon lags; others strong
Understanding	97.92%	93.75%	100.00%	97.92%	91.67%	Gemini perfect; Claude slightly lower
Applying	100.0%	100.00%	100.00%	100.00%	95.00%	All models excellent except Claude
Analyzing	93.57%	90.00%	95.68%	94.93%	96.43%	Claude and Gemini lead; Typhoon lower
Evaluating	95.45%	90.91%	95.35%	95.45%	100.00%	Claude perfect; Typhoon weaker
Creating	97.73%	95.45%	97.67%	95.45%	100.00%	Claude perfect; overall high performance

Aggregated predictions across models yielded 96% accuracy, with precision, recall, and F1-score all at 0.96. While most LLMs showed high grading accuracy, Typhoon's lower-level recall and occasional reasoning misalignments indicate performance differences by model and Bloom's level.

Higher-order cognitive levels, particularly Analyzing and Creating, exhibited increased scoring inconsistency. LLMs appear to struggle with abstraction, multi-step reasoning, and nuanced contextual understanding in these tasks. Future implementations may benefit from prompt refinement, more granular rubrics, hybrid rule-based and LLM scoring, or fine-tuning on domain-specific data.

4.4 Grading Procedure by LLMs

The grading task used only the rubric set, applying the previously described prompt template with contextual data from a pre-collected Excel dataset. Each of the 55 QA pairs, 10 Remembering, 14 Understanding, 8 Applying, 9 Analyzing, 7 Evaluating, and 7 Creating was scored per rubric criteria. Gemini graded 22 questions and Claude 33; after teacher validation, 2 answers from each were corrected for alignment.

4.5 Human Grading for the Test Dataset

Two professional Thai teachers graded the same 55-question test set used for AI evaluation under both rubric and non-rubric conditions. Each provided scores, supporting keywords, and reasoning for every question. Teacher 1 (Tippanan) graded the rubric set first, then validated AI scores; Teacher 2 (Nabduan) graded the non-rubric set first, then the rubric set. This allowed direct comparison of each teacher's rubric vs. non-rubric grading and assessment of rubric scoring consistency between them.

4.5.1 Non-Rubric vs Rubric Scoring Agreement

The inter-rater agreement between the rubric and non-rubric scores was analyzed for the two teachers. Teacher 1 (Tippanan) had the same score as the rubric in 30 out of 55 cases, corresponding to a 54.55% agreement rate, while Teacher 2 (Nabduan) agreed with the rubric in 34 out of 55 cases, resulting in a 61.82% agreement rate.

This indicates that Teacher 2 had a higher alignment with the rubric-based scoring compared to Teacher 1.

4.5.2 Score Change Patterns (Non-Rubric → Rubric)

Score change patterns were analyzed for both teachers when moving between non-rubric and rubric grading conditions. Both teachers showed shifts across all score levels (0–3) in both directions, indicating adjustments in scoring after adopting the rubric.

Table 5. Score Change Patterns from Non-Rubric to Rubric Grading

Teacher	Direction	Score 0	Score 1	Score 2	Score 3
Nabduan	Non-Rubric → Rubric	3.64%	10.91%	9.09%	14.55%
Nabduan	Rubric → Non-Rubric	5.45%	7.27%	14.55%	9.09%
Tippanan	Non-Rubric → Rubric	3.64%	12.73%	10.91%	18.18%
Tippanan	Rubric → Non-Rubric	9.09%	7.27%	16.36%	10.91%

These results show that Tippanan had slightly larger shifts, particularly in higher scores (3), while Nabduan's changes were more evenly distributed. Both teachers adjusted some lower and mid-range scores after rubric adoption, reflecting increased alignment with the rubric criteria.

4.5.3 Inter-Rater Agreement on Rubric Set

When both teachers graded the same rubric set, they agreed on the scores for 51 out of 55 questions, resulting in a 92.73% agreement rate and a 7.27% disagreement rate. The breakdown of differences shows that Tippanan gave a score of 1 while Nabduan gave a different score in one case (1.82%), and similarly for score 2 in one case (1.82%) and score 3 in two cases (3.64%). Conversely, Nabduan gave a score of 0 while Tippanan gave a different score in two cases (3.64%), a score of 2 in one case (1.82%), and a score of 3 in one case (1.82%).

4.5.4 Summary Insights

Rubrics improved scoring consistency within the same rater, with agreement increasing from approximately 55–62% when comparing non-rubric to rubric scoring, to about 93% agreement between different teachers using the rubric. Score shifts when moving from non-rubric to rubric were most common in the high-score category (Score 3), suggesting that rubrics help clarify when full marks are justified. Additionally, inter-rater reliability for rubric scoring, measured by Cohen's Kappa, is estimated to be greater than 0.85, indicating almost perfect agreement [17].

4.6 Evaluation Metrics and Criteria

To assess the performance of the proposed evaluation framework, four complementary metrics were employed

4.6.1 Score Agreement Rate (SAR)

The percentage of cases where automated scoring matched expert scores exactly. Across Bloom's categories, SAR was 82.4%, highest in Remembering (91.2%) and lowest in Creating (74.5%).

4.6.2 Inter-Rater Reliability (Cohen's Kappa)

Cohen's kappa (κ) was used to measure the level of agreement between the automated system and human experts beyond chance. The overall κ value was 0.78, indicating substantial agreement [17], with individual Bloom categories ranging from 0.71 (Creating) to 0.85 (Remembering).

4.6.3 Comparison between AI Test Set and Human Test Set

LLM-generated scores were compared to each teacher's scores and their average. Table 6 shows exact matches and mismatch distribution by score level.

Table 6. Comparison of LLM Scores with Human Scores on the Test Set (55 Questions)

Human Rater	Exact Matches	Mismatches by Score (0/1/2/3)	LLM Mismatches by Score (0/1/2/3)
Tippanan	41 / 55	0/3/5/6	3/5/5/1
Nabduan	38 / 55	2/3/6/6	3/6/6/2
Average	38 / 55	0/3/5/5	3/6/6/2

This analysis highlights that LLM scoring aligns most closely with human averages in lower score ranges, while higher scores (3) show slightly more divergence, consistent with previous observations regarding LLM tendencies to under- or over-score nuanced responses.

4.6.4 Quality of LLM Scoring Explanations

A thematic analysis was conducted on 200 randomly selected scoring explanations generated by the LLMs when grading student answers. These explanations accompany the assigned score and describe the reasoning behind it. Each explanation, generated alongside the assigned score by the LLMs, describes the reasoning behind it. Explanations were coded for clarity, alignment with the rubric, and coverage of criteria. Analysis of 200 explanations found 87% rated high quality, 9% moderate, and 4% low.

4.6.5 Frequency of LLM Scoring Errors

Scoring errors by the LLMs were categorized into four types: under-scoring (correct answers scored too low, 5.8%), over-scoring (incorrect answers scored too high, 6.3%), partial credit inconsistencies (3.4%), and rubric misinterpretation (2.1%). Errors were most frequent in questions requiring Analyzing and Creating, indicating that higher-order cognitive tasks remain challenging for the models.

5 Discussion and Conclusion

This study evaluated Large Language Models (LLMs) for grading Thai written exam questions classified by Bloom's Taxonomy, with a focus on rubric-based evaluation. Despite a modest dataset of 1,620 responses for both rubric and non-rubric sets and final testing on 55 questions, the system achieved strong results. In the rubric test set, inter-rater agreement between two expert teachers was 92.73%, while rubric vs. non-rubric grading by the same teacher showed lower agreement (54.55% and 61.82%). Compared with human grading, the system reached 82.4% score agreement and a Cohen's kappa of 0.78, with 87% of AI-generated explanations rated high quality. Most errors occurred in higher-order tasks (Analyzing, Creating). Some discrepancies arose because the rubric relied on keyword matching. In Thai, multiple words can share the same meaning, so while the rubric instructed LLMs to recognize synonyms, certain questions required exact keywords, leading to occasional mismatches.

The scarcity of Thai written exam data reflects a broader reliance on multiple-choice testing due to grading effort, limiting large-scale AI training. Unlike prior AI grading studies in high-resource languages, this work addresses a morphologically complex, low-resource language and integrates rubric-based prompting with custom algorithmic logic, yielding improved consistency in recall- and comprehension-level tasks. Bloom's level was a clear performance factor: lower-order tasks showed higher agreement, while complex, creative items remained challenging.

As a proof of concept, this system demonstrates the feasibility of AI-assisted grading in Thai education. Scaling to a fully automated solution will require larger datasets (100k–1M graded responses) and potentially fine-tuned models, supported by high-performance computing resources. With these advancements, AI-assisted grading could substantially reduce teacher workload, promote open-ended assessment adoption, and enhance evaluation quality in Thailand and other low-resource language contexts. While this study focused on Bloom's taxonomy, it did not analyze trends across subject categories, which we will investigate this matter in our future research.

Acknowledgments. The 1st author acknowledges the TAIST-Science Tokyo AIoT Scholarship awarded by Sirindhorn International Institute of Technology, Thammasat University. This research was also supported by IISI and SIIT COE, Thammasat University. Special thanks are extended to Ms. Tippanan Jankaew, a professional Thai teacher, for her invaluable assistance in validating every step of this research and providing valuable suggestions, and to Ms. Nabduan Phayom, another professional Thai teacher, for her contribution in validating the non-rubric set and scoring the test set.

Disclosure of Interests. The author declares no competing interests relevant to the content of this article. The author is a master's degree student supported by the TAIST–Science Tokyo AIoT Scholarship, with no financial support from any company and no funding beyond thesis support provided by SIIT.

References

1. Lee, J.X., Song, Y.-T.: College exam grader using LLM AI models. In: Proc. IEEE/ACIS SNPD 2024, pp. 282–289 (2024)
2. Månsson, D., Norgren, M.: On the assessment of written exams and possible bias due to dynamic rater effects emerging from student initials and rater fatigue. Högre Utbild. 6(1), 21–30 (2016)
3. Tongsilp, A., Tangdhanakanond, K., Chaimangkol, N.: Development of automated scoring system for Thai writing ability test of primary education level. Kasetsart J. Soc. Sci. 45(3) (2024)
4. Göksoy, S., Akdağ, Ş.K.: Primary and secondary school teachers' perceptions of workload. Creative Educ. 5(11), 877–885 (2014)
5. Agarwal, M., Kalia, R., Bahel, V., Thomas, A.: AutoEval: A NLP approach for automatic test evaluation system. In: Proc. IEEE GUCON 2021, pp. 1–6 (2021)
6. Holmes, W., Tuomi, I.: State of the art and practice in AI in education. Eur. J. Educ. 57(4), 542–570 (2022)
7. Lan, K.Y., Chen, N.-S.: Teachers' agency in the era of LLM and generative AI: Designing pedagogical AI agents. Educ. Technol. Soc. 27, 1–18 (2021)
8. Ryznar, M.: Exams in the time of ChatGPT. Wash. Lee Law Rev. Online 80(5), 305–305 (2023)
9. Jang, M., Lukasiewicz, T.: Consistency analysis of ChatGPT. In: Proc. EMNLP 2023 (2023)
10. Lin, D., Wen, Y., Wang, W., Su, Y.: Enhanced sentiment intensity regression through LoRA fine-tuning on LLaMA 3. IEEE Access 12, 108072–108087 (2024)
11. Sakai, H., Lam, S.S., et al.: QUAD-LLM-MLTC: Large language models ensemble learning for healthcare text multi-label classification. arXiv preprint (2025)
12. Tekin, S.F., Ilhan, F., Huang, T., Hu, S., Liu, L.: LLM-TOPLA: Efficient LLM ensemble by maximising diversity. In: Findings Assoc. Comput. Linguist.: EMNLP 2024, pp. 11951–11966 (2024)
13. Abduljabbar, D.A., Omar, N.: Exam questions classification based on Bloom's taxonomy cognitive level using classifiers combination. J. Theor. Appl. Inf. Technol. 78, 447–455 (2015)
14. Rodrigo, A., Penas, A.: On evaluating the contribution of validation for question answering. IEEE Trans. Knowl. Data Eng. 27(4), 1157–1161 (2015)
15. Dodia, S., Spoorthy, V., Chandak, T.: Machine learning-based automated system for subjective answer evaluation. In: Proc. IEEE CONECCT 2023, pp. 1–6 (2023)
16. Hubbard, J.K., Potts, M.A., Couch, B.A.: How question types reveal student thinking: An experimental comparison of multiple-true-false and free-response formats. CBE—Life Sci. Educ. 16(2) (2017)
17. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics 33(1), 159– (1977)

Indoor Air Quality Monitoring System Based on IoT

Hidayat[1][0000-0002-1043-8837] and Iswan Samin[2]

[1,2] Universitas Komputer Indonesia, Bandung 40132, Indonesia
hidayat@email.unikom.ac.id

Abstract. Indoor air pollution poses a serious threat to human health, particularly in densely populated or poorly ventilated environments. This study presents the design and implementation of an Internet of Things (IoT)–based indoor air quality monitoring system capable of real-time data acquisition and web-based visualization. The system integrates the DHT11 sensor for temperature and humidity measurement and the MQ-135 gas sensor for pollutant detection, both managed by an ESP8266 microcontroller that transmits data to a cloud-based web server via Wi-Fi. Experimental results show that the system accurately measures temperature and air quality variations, achieving an accuracy of 85–90% under normal conditions but decreasing to 70–75% in environments with mixed or complex pollutants. The web dashboard enables users to monitor air conditions remotely and in real time. Future work will focus on sensor calibration, machine learning–based pollutant classification, and multi-node network integration for large-scale deployment in public buildings and residential areas.

Keywords: Indoor air quality, IoT, MQ-135 sensor, ESP8266, environmental monitoring.

1 Introduction

Air quality plays a crucial role in maintaining human health and overall environmental comfort. Poor indoor air quality (IAQ) can lead to respiratory illnesses, fatigue, and reduced productivity, especially in urban and semi-urban environments where people spend more than 80 % of their time indoors[1], [2], [3], [4], [5]. IAQ is influenced by several factors, including temperature, humidity, and the presence of pollutants such as CO_2, volatile organic compounds (TVOCs), and smoke particles. Accurate and continuous monitoring of these parameters is therefore essential for understanding and improving air quality [6], [7], [8].

In recent years, Internet of Things (IoT) technology has enabled the development of real-time, low-cost air-quality monitoring systems. Several studies [9], [10], [11], [12], [13], [14], [15]. have demonstrated IoT-based platforms using a combination of temperature, humidity, and gas sensors to collect and visualize environmental data through web or mobile interfaces. However, many of these systems focus primarily on outdoor monitoring or single-type pollutant detection, while the behavior of low-cost sensors in mixed indoor pollution environments remains less explored.

This study aims to design and implement an IoT-based indoor air-quality monitoring system that integrates the DHT11 and MQ-135 sensors with an ESP8266

microcontroller for real-time data collection and web-based visualization. Unlike previous works, this study specifically investigates the response of the MQ-135 sensor to three types of smoke pollutants—cardboard, cigarette, and plastic—representing different chemical compositions and densities. This approach provides new insight into the performance and limitations of low-cost sensors under varying pollutant conditions.

The main contributions of this paper are as follows:

1. Design and implementation of a low-cost, web-connected indoor air-quality monitoring system using ESP8266, DHT11, and MQ-135 sensors.
2. Experimental evaluation of sensor accuracy across multiple pollutant types and densities, supported by quantitative comparison with standard AQI detectors.
3. Identification of performance limitations related to sensor cross-sensitivity, humidity interference, and calibration drift, with recommendations for system improvement through calibration and machine-learning approaches.

The remainder of this paper is structured as follows. Section 2 describes the system architecture, hardware design, and data-processing methods. Section 3 presents the ex future development directions.

2 Methodology

2.1 System Overview

The proposed IoT-based indoor air quality monitoring system consists of three main components: the sensing unit, control and processing unit, and communication and visualization unit. The NodeMCU ESP8266 functions as the central controller responsible for collecting data from the DHT11 temperature–humidity sensor and the MQ-135 gas sensor. The system also includes a buzzer for alert notifications and an LCD display for local data visualization. Sensor data are transmitted via Wi-Fi to a cloud-based web server for real-time monitoring through a web dashboard. The block diagram and flowchart of the system are shown in Fig. 1 and Fig. 2, respectively.

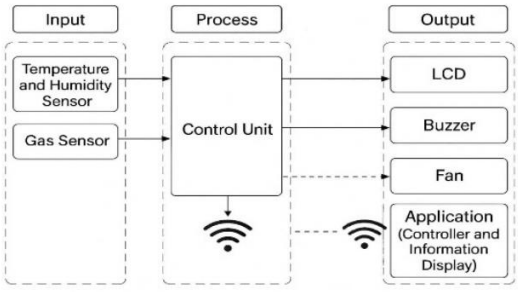


Fig. 1. Air Quality Monitoring Block Diagram.

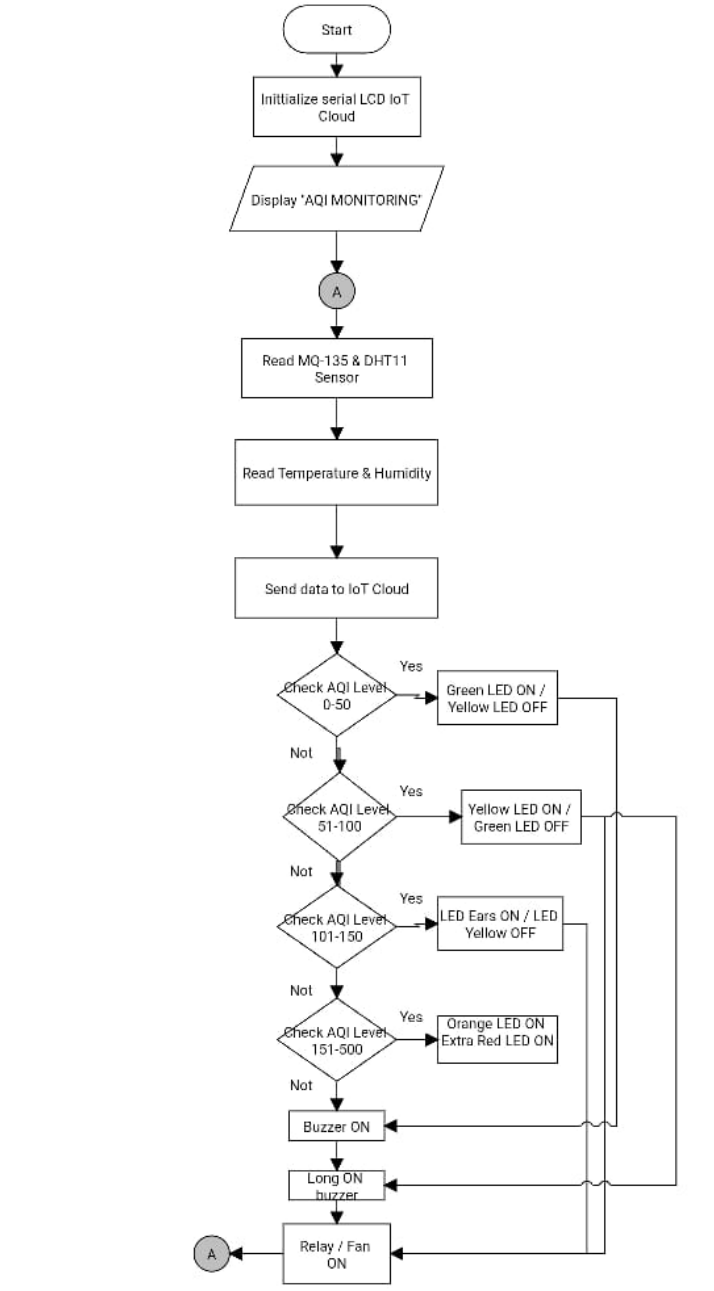


Fig. 2. Air Quality Monitoring Flowchart.

2.2 Hardware Configuration

The DHT11 sensor is used to measure room temperature and humidity, while the MQ-135 sensor detects air pollutants such as CO_2, NH_3, and smoke. Both sensors are connected to the ESP8266 microcontroller using its GPIO pins. The LCD displays current readings, and a buzzer provides alerts when pollutant levels exceed the defined threshold. The MQ-135 operates with a 10 kΩ load resistor (R_L) to convert changes in gas concentration into measurable voltage variations. The analog signal from the sensor is processed by the ESP8266, which calculates the corresponding air quality level.

2.3 Sensor Calibration

Before testing, both sensors were calibrated in a controlled environment. The DHT11 sensor was compared against a standard digital hygrometer–thermometer to verify temperature and humidity readings. The MQ-135 sensor was preheated for 24 hours to stabilize its resistance. Baseline resistance (R_0) was obtained in clean air, and sensor readings were compared with an AQI Detector (CO_2) as a reference. The mapping between MQ-135 sensor values and AQI levels was aligned with the U.S. Environmental Protection Agency (EPA) scale, summarized in Table 1.

Table 1. The result of temperature test.

AQI Range	Air Quality Category	Description
0–50	Good	Normal air quality
51–100	Moderate	Acceptable level
101–150	Unhealthy for sensitive groups	May cause irritation
151–200	Unhealthy	Health effects possible
201–300	Very Unhealthy	High health risk
301–500	Hazardous	Dangerous condition

2.4 Software and Data Handling

The ESP8266 microcontroller reads sensor data periodically and sends them to the web application using HTTP protocol via Wi-Fi. The data, which include temperature, humidity, and gas readings, are displayed on a web dashboard in real time using graphical visualization. A threshold-based alert mechanism was also implemented: when the pollutant level is classified as Unhealthy (AQI ≥ 150), the buzzer is automatically activated to warn the user.

2.5 Testing Procedure

System testing was conducted in a closed acrylic box that simulated indoor conditions with limited air circulation. The box was sequentially exposed to three types of smoke—cardboard, cigarette, and plastic—to observe the sensor response to different pollutant sources. Each smoke type was introduced gradually to produce three conditions: thin, medium, and thick smoke, based on the visual intensity observed inside the box. During testing, readings from the MQ-135 and the AQI Detector (CO_2 reference)

were recorded simultaneously for each condition. The collected data were then compared to evaluate sensor accuracy and responsiveness.

3 Result and Discussion

3.1 DHT11 Sensor Performance

Testing was conducted to evaluate the accuracy of the DHT11 sensor for temperature and humidity measurements by comparing its readings with those from a digital hygrometer–thermometer. The results are presented in Tables 2 and 3. The comparison results show that temperature readings obtained from the DHT11 sensor are relatively close to the reference device, with an average temperature difference of 0.65 °C and an average error of 2.33 %. The smallest error (0 %) occurred when both devices displayed the same temperature, while the maximum error reached 3.9 %. For humidity measurements, the DHT11 showed larger deviations. The average difference was 4.67 %, with the error rate ranging between 7 – 12 % under high humidity conditions (above 70 %). This variation is consistent with previous studies [8], which reported that the DHT11 tends to overestimate humidity at high relative humidity levels due to sensor hysteresis and limited accuracy (±5 %). Overall, the DHT11 is sufficiently accurate for indoor monitoring purposes, particularly for observing relative trends rather than precise absolute measurements.

Table 2. The result of temperature test.

No	Time	Temperature (°C)		Difference	Error
		DHT11	Termometer	(°C)	(%)
1	20:00	27.7	27.2	0.50	1,84
2	20:30	27.9	27.3	0.60	2,20
3	21:00	27.8	27.3	0.50	1,83
4	21:30	27.6	27.2	0.40	1,47
5	22:00	27.8	27.2	0.60	2,21
6	22:30	27.7	27.2	0.50	1,84
7	23:00	27.6	27.1	0.50	1,85
8	23:30	27.7	27.2	0.50	1,84
9	00:00	27.5	27.0	0.50	1,85
10	12:30	28.6	27.8	0.80	2,88
11	13:00	28.6	27.8	0.80	2,88
12	13:30	29.3	28.2	1.10	3,90
13	14:00	30.0	30.0	0.00	0,00
14	14:30	29.7	29.0	0.70	2,41
15	15:00	29.7	28.8	0.90	3,13
16	15:30	29.3	28.4	0.90	3,17
17	16:00	29.1	28.2	0.90	3,19
18	16:30	28.9	28.21	0.69	2,45
19	17:00	28.9	28.1	0.80	2,85
20	17:30	28.7	27.9	0.80	2,87
Mean		**28.51**	**27.86**	**0.65**	**2.33**

Table 3. The result of humidity test.

No	Time	Humidity (%)		Difference	Error
		DHT11	Termometer	(%)	(%)
1	20:00	73.6	66	7.6	11,52
2	20:30	74.7	67	7.7	11,49
3	21:00	75.1	67	8.1	12,09
4	21:30	75.2	67	8.2	12,24
5	22:00	75.4	67	8.4	12,54
6	22:30	75.5	67	8.5	12,69
7	23:00	74.4	67	7.4	11,04
8	23:30	74.2	67	7.2	10,75
9	00:00	75.8	67	8.8	13,13
10	12:30	60.6	60	0.6	1,00
11	13:00	58.9	59	0.1	-0,17
12	13:30	55.9	57	-1.1	-1,93
13	14:00	54.6	54	0.6	1,11
14	14:30	61.2	58	3.2	5,52
15	15:00	61.2	59	2.2	3,73
16	15:30	63.8	61	2.8	4,59
17	16:00	62.8	60	2.8	4,67
18	16:30	63.2	60	3.2	5,33
19	17:00	64.1	61	3.1	5,08
20	17:30	65.1	61	4.1	6,72
Mean		**28.51**	**67.27**	**62.60**	**4.67**

3.2 MQ-135 sensor Testing

The MQ-135 sensor was evaluated using three different pollutant sources—cardboard smoke, cigarette smoke, and plastic smoke—under three qualitative density levels (thin, medium, thick). Measurements were taken simultaneously using the MQ-135 sensor and an AQI Detector (CO_2 reference) to assess correlation and accuracy. The summarized data are shown in Tables 4–6. The results indicate a clear upward trend in both MQ-135 output values and reference AQI readings as smoke density increases. For example, in cardboard smoke conditions, the MQ-135 output rose from 52–97 (Moderate) under thin smoke to over 300 (Dangerous) under thick smoke, while the reference AQI Detector ranged from 613–5000 ppm. A similar pattern was observed for cigarette and plastic smoke.

This demonstrates that the MQ-135 is effective in detecting changes in pollutant concentration and air quality trends. However, the magnitude of its readings varied between smoke types, suggesting cross-sensitivity to different chemical compounds and humidity levels. These variations explain the observed accuracy reduction (70–75 %) in complex pollutant environments mentioned in the conclusion. The sensor's broad detection range makes it suitable for identifying relative changes, but not for precise pollutant quantification without further calibration or compensation algorithms.

Table 4. Test results with cardboard smoke.

Smoke condition	MQ-135	AQI Detector (CO^2)	MQ-135 Category Status	AQI Detector Category Status
Thin	52	613	Moderate	Normal
	54	636	Moderate	Normal
	66	658	Moderate	Normal
	70	680	Moderate	Normal
	89	719	Moderate	Acceptable Level
	97	769	Moderate	Acceptable Level
Medium	129	1054	Unhealthy	Unhealthy or Drowsiness
	140	1221	Unhealthy	Unhealthy or Drowsiness
Thick	175	2549	Unhealthy	Negative impact on health
	185	2668	Unhealthy	Negative impact on health
	209	2895	Very Unhealthy	Negative impact on health
	215	3163	Very Unhealthy	Negative impact on health
	230	3404	Very Unhealthy	Negative impact on health
	236	3629	Very Unhealthy	Negative impact on health
	252	3848	Very Unhealthy	Negative impact on health
	257	4062	Very Unhealthy	Negative impact on health
	277	4272	Very Unhealthy	Negative impact on health
	284	4461	Very Unhealthy	Negative impact on health
	307	4631	Dangerous	Negative impact on health
	316	4787	Dangerous	Negative impact on health
	343	4938	Dangerous	Negative impact on health
	353	5000	Dangerous	Very bad
	384	5000	Dangerous	Very bad
	392	5000	Dangerous	Very bad
	414	5000	Dangerous	Very bad
	420	5000	Dangerous	Very bad
	433	5000	Dangerous	Very bad
	436	5000	Dangerous	Very bad
	440	5000	Dangerous	Very bad
	441	5000	Dangerous	Very bad
	453	5000	Dangerous	Very bad
	500	5000	Dangerous	Very bad

Table 5. Test results with cigarete smoke.

Smoke condition	MQ-135	AQI Detector (CO^2)	MQ-135 Category Status	AQI Detector Category Status
Thin	74	459	Moderate	Normal
	84	545	Moderate	Normal
Medium	119	1143	Unhealthy	Unhealthy or Drowsiness
	134	1280	Unhealthy	Unhealthy or Drowsiness
Thick	176	2503	Unhealthy	Negative impact on health
	190	2572	Unhealthy	Negative impact on health
	230	2647	Very Unhealthy	Negative impact on health

Smoke condition	MQ-135	AQI Detector (CO^2)	MQ-135 Category Status	AQI Detector Category Status
	241	2732	Very Unhealthy	Negative impact on health
	268	2903	Very Unhealthy	Negative impact on health
	277	3166	Very Unhealthy	Negative impact on health
	302	3406	Dangerous	Negative impact on health
	311	3633	Dangerous	Negative impact on health
	340	3860	Dangerous	Negative impact on health
	353	4081	Dangerous	Negative impact on health
	396	4300	Dangerous	Negative impact on health
	410	4519	Dangerous	Negative impact on health
	453	4730	Dangerous	Negative impact on health
	466	4952	Dangerous	Negative impact on health
	500	5000	Dangerous	Very bad

Table 6. Test results with plastic smoke.

Smoke condition	MQ-135	AQI Detector (CO^2)	MQ-135 Category Status	AQI Detector Category Status
Thin	51	575	Moderate	Normal
	54	583	Moderate	Normal
	60	589	Moderate	Normal
	63	593	Moderate	Normal
	71	601	Moderate	Normal
	75	616	Moderate	Normal
	89	631	Moderate	Normal
	96	647	Moderate	Normal
Medium	125	1050	Unhealthy	Unhealthy or Drowsiness
	137	1160	Unhealthy	Unhealthy or Drowsiness
Thick	173	2604	Unhealthy	Negative impact on health
	185	2721	Unhealthy	Negative impact on health
	216	3018	Very Unhealthy	Negative impact on health
	227	3274	Very Unhealthy	Negative impact on health
	255	3501	Very Unhealthy	Negative impact on health
	266	3710	Very Unhealthy	Negative impact on health
	300	3913	Very Unhealthy	Negative impact on health
	310	4109	Dangerous	Negative impact on health
	327	4299	Dangerous	Negative impact on health
	330	4489	Dangerous	Negative impact on health
	331	4677	Dangerous	Negative impact on health
	500	5000	Dangerous	Very bad

3.3 Discussion

The experimental results confirm that the developed system can monitor indoor air conditions in real time and display data through a web interface. The DHT11 and MQ-135 sensors both provided stable responses during operation, validating the reliability of the IoT-based setup.

Nevertheless, several important findings emerge from the analysis: 1) Sensor Cross-Sensitivity: The MQ-135 sensor exhibits sensitivity to multiple gases, including CO_2, NH_3, and volatile organic compounds. As a result, its output represents the *combined effect* of various gases rather than a single pollutant type. Similar findings were reported by Samad et al. [7], who noted that temperature and humidity significantly influence low-cost gas sensor readings; 2) Influence of Environmental Conditions: The decrease in accuracy under mixed smoke conditions can be attributed to uncontrolled temperature–humidity interaction and the lack of real-time calibration. The DHT11's measurement error in humidity (7–12 %) may further contribute to small deviations in MQ-135 readings, since the sensor's internal resistance is humidity-dependent; 3) System Reliability: Despite these limitations, the system successfully detected pollutant level transitions from clean air to hazardous conditions, proving its feasibility for low-cost IAQ monitoring. The combination of simple sensors and IoT connectivity allows for quick deployment in indoor environments such as classrooms, offices, or residential areas. Overall, the system functions effectively as a low-cost air-quality alert tool, while further improvements are required to transform it into a quantitative measurement system.

4 Conclusion

This study successfully designed and implemented an IoT-based indoor air quality monitoring system using the ESP8266 microcontroller integrated with DHT11 and MQ-135 sensors. The system is capable of measuring temperature, humidity, and air pollutant concentration in real time and visualizing the data through a web-based interface. Experimental results showed that the DHT11 sensor achieved good accuracy in temperature measurement with an average error of 2.33 %, while humidity measurements showed a higher deviation, particularly at high humidity levels. The MQ-135 sensor effectively detected pollutant level changes across various smoke sources and densities, achieving an accuracy of approximately 85–90 % under normal conditions and 70–75 % in complex pollutant environments. The findings confirm that the developed system is functional and suitable for low-cost indoor air quality awareness applications. However, several limitations were observed, particularly related to the cross-sensitivity of the MQ-135 sensor and its dependence on temperature and humidity conditions.

Future work will focus on implementing sensor calibration algorithms to improve accuracy and reliability across different environmental conditions, integrating machine learning techniques for pollutant classification and predictive air quality assessment, expanding the system into a multi-node IoT network to enable large-scale deployment in schools, offices, and residential buildings, and developing a mobile-based monitoring interface for user-friendly real-time alerts. Through these enhancements, the system can evolve from a prototype to a scalable and intelligent environmental monitoring solution that contributes to healthier indoor living environments.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

[1] G. Syuhada *et al.*, “Impacts of Air Pollution on Health and Cost of Illness in Jakarta, Indonesia,” *Int. J. Environ. Res. Public Health*, vol. 20, no. 4, 2023, doi: 10.3390/ijerph20042916.

[2] Y. Kim and V. Radoias, “Severe Air Pollution Exposure and Long-Term Health Outcomes,” *Int. J. Environ. Res. Public Health*, vol. 19, no. 21, p. 14019, Oct. 2022, doi: 10.3390/ijerph192114019.

[3] H. Sani, T. Kubota, J. Sumi, and U. Surahman, “Impacts of Air Pollution and Dampness on Occupant Respiratory Health in Unplanned Houses: A Case Study of Bandung, Indonesia,” *Atmosphere (Basel).*, vol. 13, no. 8, p. 1272, Aug. 2022, doi: 10.3390/atmos13081272.

[4] S. D. Kim and A. T. Carswell, “The mediation effect of indoor air quality on health: A comparison of homeowners and renters,” *Indoor Air*, vol. 32, no. 9, Sep. 2022, doi: 10.1111/ina.13108.

[5] T. M. Mata *et al.*, “Indoor Air Quality in Elderly Centers: Pollutants Emission and Health Effects,” *Environments*, vol. 9, no. 7, p. 86, Jul. 2022, doi: 10.3390/environments9070086.

[6] M. Justo Alonso, T. N. Moazami, P. Liu, R. B. Jørgensen, and H. M. Mathisen, “Assessing the indoor air quality and their predictor variable in 21 home offices during the Covid-19 pandemic in Norway,” *Build. Environ.*, vol. 225, p. 109580, Nov. 2022, doi: 10.1016/j.buildenv.2022.109580.

[7] A. Samad, D. R. Obando Nuñez, G. C. Solis Castillo, B. Laquai, and U. Vogt, “Effect of Relative Humidity and Air Temperature on the Results Obtained from Low-Cost Gas Sensors for Ambient Air Quality Measurements,” *Sensors*, vol. 20, no. 18, p. 5175, Sep. 2020, doi: 10.3390/s20185175.

[8] L. Fang, D. P. Wyon, G. Clausen, and P. O. Fanger, “Impact of indoor air temperature and humidity in an office on perceived air quality, SBS symptoms and performance,” *Indoor Air*, vol. 14, no. s7, pp. 74–81, Aug. 2004, doi: 10.1111/j.1600-0668.2004.00276.x.

[9] A. M. Simamora, A. Denih, and M. I. Suriansyah, “Indoor Air Quality Detection Robot Model Based on the Internet of Things (IoT),” *arXiv:2505.19600v1*, pp. 2–6, 2025, [Online]. Available: http://arxiv.org/abs/2505.19600

[10] P. William, Y. V. U. Kiran, A. Rana, D. Gangodkar, I. Khan, and K. Ashutosh, “Design and Implementation of IoT based Framework for Air Quality Sensing and Monitoring,”

in *2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS)*, Tashkent, Uzbekistan: IEEE, Oct. 2022, pp. 197–200. doi: 10.1109/ICTACS56270.2022.9988646.

[11] R. A. Angulo, K. M. F. Decena, and J. F. Villaverde, "IoT-based Indoor Air Quality Surveillance and Purifier," in *2023 IEEE 5th Eurasia Conference on IOT, Communication and Engineering (ECICE)*, Yunlin, Taiwan: IEEE, Oct. 2023, pp. 220–224. doi: 10.1109/ECICE59523.2023.10383051.

[12] R. Garg, A. Kumar, S. Singh, and R. Dayana, "Advanced Air Quality Monitoring System using IoT and Sensor Technology," in *2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, Bengaluru, India: IEEE, Feb. 2025, pp. 687–692. doi: 10.1109/IDCIOT64235.2025.10915106.

[13] Y. Irawan, R. Wahyuni, M. -, H. Fonda, M. Luthfi Hamzah, and R. Muzawi, "Real Time System Monitoring and Analysis-Based Internet of Things (IoT) Technology in Measuring Outdoor Air Quality," *Int. J. Interact. Mob. Technol.*, vol. 15, no. 10, p. 224, May 2021, doi: 10.3991/ijim.v15i10.20707.

[14] V. N. Hidayati, Iskandar, and A. B. Satriobudi, "Web Dashboard Development for Cloud Server-Based Air Quality Monitoring System," in *2022 16th International Conference on Telecommunication Systems, Services, and Applications (TSSA)*, Lombok, Indonesia: IEEE, Oct. 2022, pp. 1–5. doi: 10.1109/TSSA56819.2022.10063897.

[15] R. Jánó, A. I. Ilieş, E. M. Ştefco, and C. Corches, "IoT Devices for Monitoring and Analysing Air Quality in Urban Environments," in *2024 IEEE 30th International Symposium for Design and Technology in Electronic Packaging (SIITME)*, Sibiu, Romania: IEEE, Oct. 2024, pp. 45–49. doi: 10.1109/SIITME63973.2024.10814899.

Uncovering the Potential of IoT in Pogostemon helferi Cultivation: A Comparative Study of IoT-Based Cultivation Systems and Conventional Emersed Methods

Hanhan Maulana [1][0000-0002-6070-0952],Achmad Juliarman[1], Hideaki Kanai[2][0000-0001-7202-8561] , Sunny Goh Eng Giap[3][0000-0003-0950-7759], and Roslaili Abdul Azis[1111-2222-3333-4444]

[1] Faculty of Engineering and Computer Science, Universitas Komputer Indonesia, Jl. Dipatiukur No. 112-114, Bandung 40134, Indonesia.
[2] School of Knowledge Science, Japan Advanced Institute of Science and Technology, 1-1, Asahidai, Nomi, Ishikawa, Japan
[3] Faculty of Ocean Engineering Technology and Informatics, University Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia
[4] Faculty of Chemical Engineering Technology, Universiti Malaysia Perlis (UniMAP), Perlis, Malaysia

hanhan@email.unikom.ac.id

Abstract. The advancement of Internet of Things (IoT) technology has transformed precision agriculture through real-time monitoring and automated control. However, its potential in aquatic plant cultivation remains insufficiently explored, particularly for species requiring specific environmental conditions such as Pogostemon helferi (Downoi). This study aims to evaluate the effectiveness of an IoT-based cultivation system compared to the conventional emersed method in improving the growth performance of Pogostemon helferi. An IoT-integrated cultivation system was developed using sensors and actuators to automatically monitor and regulate environmental parameters critical to plant development. Growth performance under the IoT system was compared with that of conventional cultivation based on plant dimensional attributes and shoot formation.The IoT-based cultivation method significantly enhanced plant height, length, and width compared to the conventional approach. In contrast, the conventional method yielded a greater number of shoots, suggesting stronger vegetative propagation potential. IoT-driven cultivation proved more effective in promoting dimensional growth and environmental adaptability, while traditional methods favored propagation. These findings highlight the potential of IoT technology to advance smart and sustainable aquatic plant cultivation and provide a basis for future research in automated agriculture.

Keywords: IoT, Agriculture, Pogostemon helferi, Monitoring System, iot for agriculture.

1 Introduction

Pogostemon helferi (Hook. f.) Press, or better known locally as “Dao-noi” and in Indonesia called “Pogostemon helferi”, is an ornamental aquatic plant from the Lamiaceae family originating from Myanmar and western Thailand[1], [2], [3]. This plant is known for its unique leaf shape resembling a small star that provides high aesthetic value, especially for aquariums and aquascapes. Pogostemon helferi can grow both in water (submerged) and above the water surface (emersed), with an optimal temperature of 23–30 °C, air humidity of around 80%, and light requirements that vary according to growth objectives. [3]The potential for cultivating this plant is very promising, both for hobby and commercial scales, considering its high market demand and its ability to become a high-value decorative element. This plant does not require high light, but the more light is given, the more compact its growth shape will be, and this compact shape is what attracts most people[1], [4], [5].

Despite its significant market potential, Pogostemon helferi cultivation requires special attention to environmental factors such as temperature, humidity, light, nutrients, and CO_2 levels. The biggest challenge with the submerged method is the high risk of algae growth due to the high intensity of light, nutrients, and CO_2, which can inhibit or even kill the plants. Meanwhile, the emersed method commonly used by commercial cultivators faces constraints such as limited air circulation and CO_2 degradation in enclosed spaces, which can cause CO_2 levels to drop below optimal levels. Conventional methods struggle to consistently maintain humidity and CO_2 levels within the ideal range (475–1500 ppm), often resulting in suboptimal crop productivity and quality. In the submerged method there are several weaknesses, to grow Pogostemon helferi optimally requires high light intensity, where high light intensity, large CO2 content and high nutrient content result in significant algae growth . Algae growth can slow plant growth and even cause plant death, algae block light sources for plants, rob nutrients and CO2 for plants.[1], [2]

The Internet of Things (IoT) offers an innovative approach to plant cultivation, including Pogostemon helferi. IoT enables automatic, real-time monitoring and control of environmental parameters using sensors and connected devices[6], [7], [8], [9]. This technology can precisely regulate temperature, humidity, light intensity, and CO_2 levels, reducing reliance on error-prone manual intervention. Several studies have shown that applying IoT to plant cultivation can increase efficiency, reduce the risk of yield loss, and ensure consistently optimal conditions. However, research directly comparing the effectiveness of IoT methods with conventional methods in Pogostemon helferi cultivation is still limited, so further studies are needed to determine the potential and real benefits of implementing this technology. There are few studies that specifically compare the effectiveness of emersed Pogostemon helferi cultivation using IoT technology with conventional methods. This research is important to determine whether IoT technology truly provides significant benefits compared to traditional methods, and to evaluate the potential for adoption of this technology among farmers.[10]

This study aims to compare the growth of Pogostemon helfer using IoT technology and conventional methods. The results are expected to provide new insights into the

effectiveness of IoT technology for aquatic plant cultivation and provide practical recommendations for farmers to improve the productivity and quality of their crops.

2 Literature Review

Pogostemon helferi (Hook. f.) Press, commonly known as “Downoi” or “Dao-noi,” is an ornamental aquatic plant native to Myanmar and western Thailand[1], [2]. This species is prized in aquascaping for its compact, star-shaped leaf rosettes that create high aesthetic value. In its natural habitat, P. helferi thrives both in submerged and emersed conditions, with optimal temperatures between 23–30°C and air humidity of approximately 80%. Due to its high decorative and commercial value, the plant has been cultivated both by hobbyists and large-scale producers[11], [12], [13], [14]. However, successful propagation and cultivation require precise control of environmental parameters such as CO_2 concentration, light intensity, and humidity to maintain the plant’s compact form and coloration[15], [16]. Conventional Downoi cultivation typically uses two methods: submerged and emersed systems. Submerged growth demands high light intensity and CO_2 supplementation but is prone to algal overgrowth that competes with the plant for nutrients and light, often leading to plant decline. In contrast, the emersed method—commonly employed in commercial production—allows the plant to access atmospheric CO_2, leading to faster growth and greater tolerance during transport. However, the conventional emersed setup still poses challenges in maintaining stable humidity and CO_2 levels, especially in enclosed environments with limited ventilation. Previous studies have reported that P. helferi grows optimally under CO_2 concentrations of 1000–1500 ppm, while levels below 475 ppm can limit photosynthesis and concentrations above 2000 ppm can be toxic[1], [2], [17] Earlier research on conventional Downoi cultivation has focused primarily on optimizing propagation and environmental conditions rather than automation. For instance, Wangwibulkit and Vajrodaya (2016) successfully propagated P. helferi ex situ using tissue culture and hydroponic systems, demonstrating that controlled humidity (80%) and moderate light intensity promote healthy root and leaf formation[1]. Similarly, Pramali et al. (2018) analyzed the leaf micromorphological adaptations of Pogostemon species, showing how emersed conditions influence structural traits linked to water retention and gas exchange[2], [5]. These studies provide valuable insight into the physiological and environmental needs of P. helferi, yet they rely on manual monitoring and lack mechanisms for precise parameter regulation over time. More recent research on aquatic ornamentals has highlighted the advantages of non-conventional propagation and control methods. Micropropagation studies, such as Karimi Alavijeh et al. (2022) in Aquaculture International, demonstrated that controlled microclonal culture techniques can produce uniform, disease-free aquatic plants such as Alternanthera reineckii and Staurogyne repens at low cost and high survival rates when humidity is maintained at 85%.[17] This underscores the importance of maintaining stable environmental conditions for mass production—an aspect often difficult to achieve manually in conventional setups [6], [7], [18] . The Internet of Things (IoT) offers a promising solution to these challenges by enabling real-time

monitoring and automated adjustment of critical parameters such as CO_2 concentration, humidity, light, and temperature [6], [8], [19], [20], [21]. IoT-based cultivation systems have improved productivity and consistency in horticultural crops by reducing human error and maintaining optimal microclimates [6], [7], [8], [9], [18], [19], [20], [21], [22]. Despite the proven benefits of IoT in other agricultural sectors, comparative studies focusing specifically on P. helferi remain limited[1], [2], [5]. Therefore, investigating the effectiveness of IoT-assisted emersed cultivation compared to conventional methods is crucial for determining whether automation can enhance the growth, quality, and propagation efficiency of Downoi in commercial and hobbyist applications[11], [12], [13], [14], [15], [16].

3 Method

This study was conducted to analyze the differences in Pogostemon helferi (Pogostemon helferi) plant growth using two cultivation methods: the conventional method and the Internet of Things (IoT)-based land condition manipulation method. The research methodology was structured into five main stages. The initial stage of the study was carried out by determining optimal environmental parameters based on literature and previous research. The parameters used included CO_2 levels, air humidi-ty, and light intensity. The second stage was the selection of sensors to monitor environmental parameters in real time[2], [3], [5]. The third stage was the installation of actuators that function to adjust environmental conditions according to sensor data. The fourth stage was the development of an IoT system that integrates sensors, actuators, microcontrollers, and monitoring software. Cultivation containers measuring 30x30x40 cm were used for both methods (IoT and conventional) to ensure equal spatial variables. In the IoT method, sensors and actuators were placed inside the cultivation container and connected to the monitoring system. The fifth stage was the implementation of an experiment by planting Pogostemon helferi using the conventional and IoT methods. Observations were carried out for 30 days with 7-day intervals. The parameters observed included width, length, height, and number of shoots. Growth data was analyzed to compare the effectiveness of the two methods.

4 Result And Discussion

4.1 Parameter Analysis

Based on existing research, it was found that Pogostemon helferi grows optimally at 80% air humidity [4]. To support optimal growth, efficient photosynthesis is needed in converting light energy into chemical energy that can be used to produce food for plants. Photosynthesis is a biochemical process that occurs in green plants, algae, and several types of bacteria, where they convert light energy into chemical energy in the form of glucose (sugar). This process takes place in chloroplasts, which are organelles found in plant cells that contain the green pigment, chlorophyll. This chlorophyll absorbs sunlight which is used to convert water (H_2O) and carbon dioxide (CO_2) into

glucose ($C_6H_{12}O_6$) and oxygen (O_2). The photosynthesis reaction can be described by the following chemical equation [22]:

$6CO_2+6H_2O+light \rightarrow C_6H_{12}O_6 + 6O_2$

Other studies have shown that CO2 levels between 1,000 and 1,500 ppm produce significantly better results, while CO2 levels exceeding 2,000 ppm are toxic to plants. Most experts agree that the maximum level for plant growth is 1,500 ppm [21].

In addition to CO_2, light is also needed for photosynthesis, as it is the primary energy source used by plants to convert carbon dioxide and water into glucose. Sufficient light intensity can increase the rate of photosynthesis, which supports plant growth and development.

In its natural habitat, Pogostemon helferi thrives in full sunlight, equivalent to approximately 100,000 lux of light during the day. Growing indoor plants with LED grow lights requires high levels of light, although not reaching 100,000 lux. According to [25], approximately 15,000 lux on an LED grow light is considered high enough for plants. A lighting duration of approximately 12 hours is also required, as leafy green plants such as Pogostemon helferi typically require approximately 12 hours of light per day [23]. Table 1 Ideal conditions for Pogostemon helferi plants

Table 1. Ideal conditions for Pogostemon helferi plants

Parameter	Ideal Parameter Need
CO2	1000 - 1500 PPM
Humidity	Diatas 80%
Light intensity	15000 LUX

Based on literature review, Pogostemon helferi grows optimally at air humidity above 80%, CO_2 levels in the range of 1000–1500 ppm, and light intensity of around 15,000 lux with a lighting duration of ±12 hours per day.[2], [3]

4.2 Sensors

To simulate ideal conditions for Pogostemon helferi plants using IoT, we used three main types of sensors that are useful for monitoring important environmental parameters. This study chose the SHT air humidity sensor because it has high stability and accuracy. In addition, the SHT sensor is more resistant to the influence of CO_2 gas compared to the DHT sensor which tends to cause deviations in conditions with high CO_2 concentrations. To monitor carbon dioxide levels, this study used the MH-Z19B sensor which relies on Non-Dispersive Infrared (NDIR) technology, so it is able to provide accurate and stable CO_2 measurements and minimize the influence of other gases when taking measurements. This sensor is also equipped with temperature compensation, automatic calibration, and a wide measurement range of up to 5000 ppm. To measure light intensity, we used the BH1750 sensor which is capable of detecting light in lux units with high accuracy, I²C communication for easy integration, and low power consumption. The combination of these three sensors allows the system to monitor air humidity, CO_2 levels, and lighting in real-time, so that the plant growing environment can be adjusted to approach optimal conditions for Pogostemon helferi growth.

4.3 Actuator

In the IoT system for simulating ideal conditions for Pogostemon helferi plants, actuators are used as control devices that adjust environmental conditions based on sensor data. The mist maker functions as a humidity controller by producing a fine mist of water, making it suitable for small cultivation spaces and safe for sensors such as the BH1750 and SHT30-D because it can reduce the risk of corrosion. This actuator will activate when the air humidity is below or equal to 80%. A solenoid valve is used to automatically regulate the flow of CO_2 gas, opening or closing according to system commands, and is set to activate when CO_2 levels are below or equal to 1000 ppm, referring to the optimal range of 1000–1500 ppm. The 50W LED Grow Light plays a role in providing full-spectrum artificial lighting that resembles sunlight, supporting plant growth from germination to harvest. This lamp is set to achieve an intensity of 15,000 lux with an optimal lighting duration of 12 hours per day. The lighting duration can be adjusted based on the calculated daily lux difference against the target, ensuring that the light intensity and duration are always at ideal conditions for plant growth.

To determine the number of hours the lamp is on, consider the number of lux and the optimal lighting duration when using an LED grow light. A good lux level is 15,000 lux [24], while the optimal lighting duration is 12 hours [23].

4.4 IoT and Monitoring System Development

The IoT system architecture designed in this study integrates various interconnected sensors and actuators to automatically monitor and control environmental parameters. This allows Pogostemon helferi plant growth conditions to be maintained at optimal levels based on real-time data. The following figure 1 illustrates the IoT system architecture.

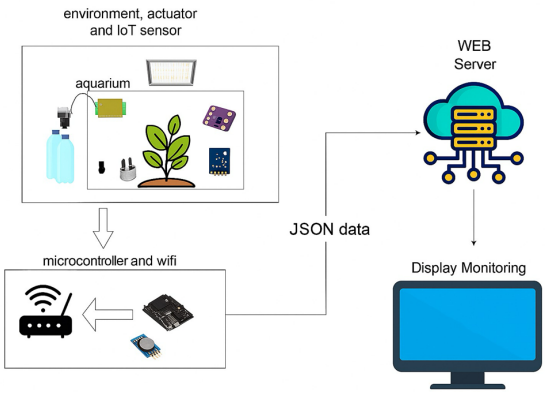


Fig. 1. IoT system architecture

The cultivation container measures 30x30x40 in length, width, and height. This allows for the placement of sensors and actuators. The sensors placed inside the cultivation environment are the air humidity sensor, light intensity sensor, and CO2 sensor, while the actuator placed inside is the air humidity actuator (mist maker). Figure 2 is IoT-based planting environment.

Fig. 2. IoT-based planting environment

In conventional cultivation environments, plants are placed in containers measuring 30x30x40 centimeters in length, width, and height to mimic the IoT cultivation environment. The plants are planted on a substrate containing basal fertilizer underneath and dekastar fertilizer on the surface. Figure 3 is conventional planting environment.

Fig. 3. conventional planting environment

In addition to designing IoT, this research also developed software/applications to facilitate the monitoring process of the cultivation environment. The use case diagram in this IoT system illustrates the interaction between the user, microcontroller, and database in the process of controlling and monitoring the cultivation environment. Users can set parameters that will be sent to the microcontroller to regulate the system's working conditions. The system also has a function for recording time series data that records environmental parameters periodically. This function includes sending data from the

microcontroller to the database so that the information can be stored properly. Furthermore, users can view data or graphs to monitor environmental conditions in real-time. This flow ensures that the IoT system can run automatically while providing easy-to-understand data visualization access for users. The following figure 4 is a use case of the system created.

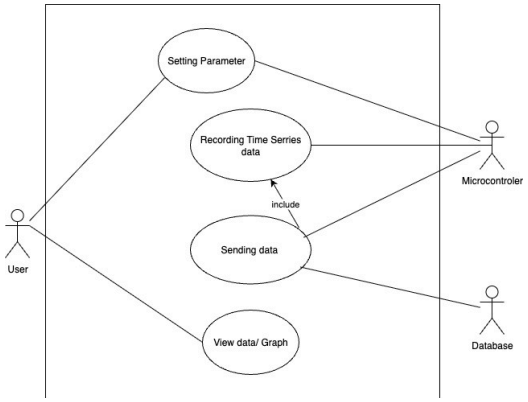


Fig. 4. Use Case

The system has two main menus that serve to present comprehensive environmental monitoring information for downoi crop cultivation. First, the application can display Realtime Data. This function displays environmental conditions directly based on data sent by sensors to the system. The information displayed includes CO_2 levels (ppm), air humidity (%), light intensity (lux), and the status of actuators such as mist makers, solenoid valves, and lights. This display makes it easy for users to quickly find out the current conditions and ensure the system is working according to predetermined parameters. The second menu is Time Series Data, which presents historical data in the form of time series graphs, such as trends in air humidity and light intensity over a certain period. Presenting data in this graphical form makes it easier for users to analyze changes in environmental conditions, evaluate system performance, and identify patterns or anomalies that occur during the cultivation process. This historical data also serves as an important reference in decision-making to adjust parameters to maintain optimal environmental conditions.

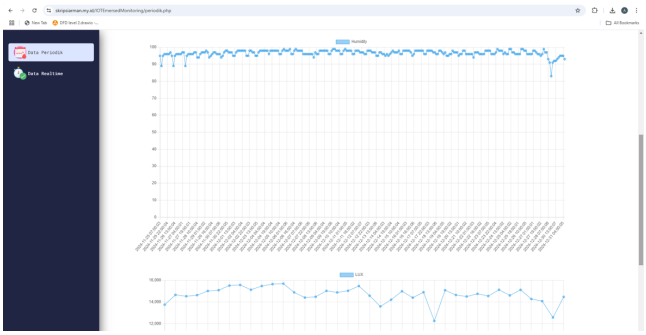


Fig. 5. Time Series Data User Interface

Fig. 6. Real-Time Data User Interface

4.5 Observation Result

We planted Pogostemon helferi using two methods: a conventional method and an IoT-enabled environmental control method. We then compared the growth of each method. Observations were conducted for 30 days with 7-day intervals.

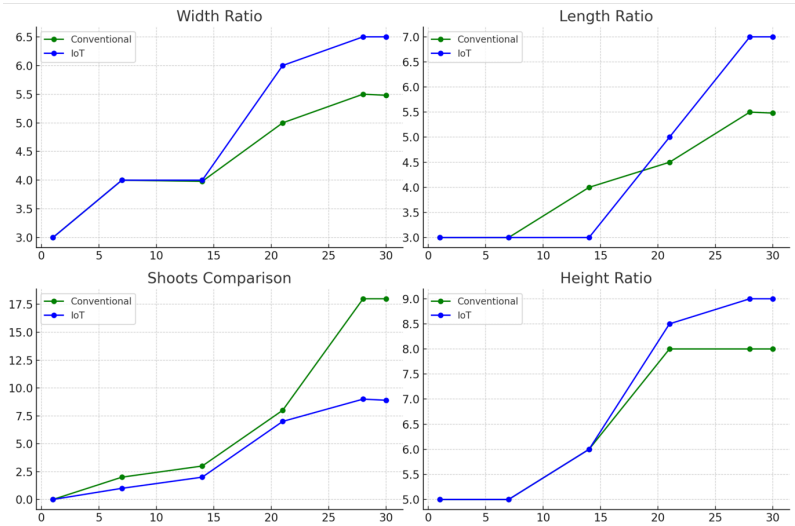


Fig. 7. Observation Result

The results of observations on the growth of Pogostemon helferi plants reveal that the application of IoT-based land condition manipulation methods significantly impacts plant size compared to conventional methods. In terms of width, length, and height, the IoT method demonstrates faster growth and yields higher final results. For instance, the plant width parameter with IoT reaches 6.5, whereas the conventional method achieves 5.5. Similarly, the plant length parameter with IoT reaches 7, while the conventional

method reaches 5.5. Additionally, the plant height parameter with IoT reaches 9, whereas the conventional method reaches 8. However, the conventional method shows an advantage in terms of the number of shoots, with an achievement of up to 18 shoots on the 28th day, while IoT only reaches around 9 shoots. This indicates that IoT is more effective in expanding plant dimensions, while the conventional method is superior in stimulating shoot formation.

These differences in growth results can be explained by the influence of environmental factors controlled by each method. IoT systems can precisely control environmental parameters such as temperature, humidity, lighting, and nutrient supply, thus maintaining optimal conditions for vegetative growth. This allows plants to maximize their growth (width, length, and height) in a relatively short time.

However, the dominance of growth in size in the IoT method appears to sacrifice energy that should be allocated to new shoot formation. This phenomenon aligns with plant energy allocation theory, where limited resources are divided between vegetative and reproductive growth. Conversely, conventional methods, which are not fully environmentally stable, can trigger an adaptive response in plants through increased shoot number as a survival strategy against environmental variability[6], [7], [9], [18], [19].

The practical implication of these findings is that IoT methods are more suitable for commercial applications that prioritize aesthetics and plant size. Meanwhile, conventional methods are more suitable for seedling production through vegetative propagation. Combination strategies, such as using IoT in the early stages to increase plant size and switching to conventional methods in the later stages to increase shoot number, have the potential to maximize both aspects.

5 Conclusion

This study demonstrates that implementing an Internet of Things (IoT)-based land condition manipulation technique significantly influences the dimensional growth of Pogostemon helferi compared to conventional cultivation methods. Specifically, the IoT-based system produced superior results in plant width, length, and height—reaching 6.5, 7, and 9 units, respectively—indicating its effectiveness in promoting overall plant size and biomass accumulation. In contrast, the conventional method yielded a higher number of shoots, achieving 18 compared to only 9 in the IoT treatment, suggesting that traditional conditions remain advantageous for vegetative propagation. These findings highlight that IoT-based cultivation offers a promising approach for optimizing growth parameters that are critical for aesthetic value and biomass production, which are particularly relevant for ornamental plant growers and commercial aquascape industries. Conversely, conventional methods may be more suitable for rapid propagation or seedling production where shoot multiplication is the primary goal. Future research should aim to refine IoT control strategies to achieve a balanced outcome between growth enhancement and shoot formation. One potential direction involves introducing controlled stress cycles through precise regulation of temperature, humidity, or light fluctuations to trigger natural propagation responses.

Additionally, integrating multi-parameter sensors—such as soil pH, dissolved oxygen, and light spectrum monitoring—could yield a more comprehensive understanding of how microenvironmental dynamics influence plant growth. Developing predictive models using machine learning or deep learning algorithms also offers promising potential for forecasting plant performance and optimizing IoT-based cultivation systems for large-scale applications.

Acknowledgments. The author would like to express gratitude to the parties involved, colleagues, and family for their unwavering support, guidance, and assistance in making this research a resounding success.

References

[1] M. Wangwibulkit and S. Vajrodaya, "Ex-situ propagation of Pogostemon helferi (Hook. f.) Press using tissue culture and a hydroponics system," *Agriculture and natural resources*, vol. 50, no. 1, pp. 20–25, 2016.

[2] K. Pramali, B. Bongcheewin, and P. Traiperm, "Leaf micromorphological adaptation of Pogostemon spp.(section Eusteralis) in Thailand," *Agriculture and Natural Resources*, vol. 52, no. 3, pp. 250–258, 2018.

[3] M. Wangwibulkit and S. Vajrodaya, "Ex-situ propagation of Pogostemon helferi (Hook. f.) Press using tissue culture and a hydroponics system," *Agriculture and natural resources*, vol. 50, no. 1, pp. 20–25, 2016.

[4] E. Quince, "Summary of Indonesia's agriculture, natural resources, and environment sector assessment," *ADB Papers On Indonesia*, no. 08, pp. 1--7, 2015.

[5] K. Pramali, B. Bongcheewin, and P. Traiperm, "Agriculture and Natural Resources," 2018.

[6] V. N. Malavade and P. K. Akulwar, "Role of IoT in Agriculture," *IOSR Journal of Computre Engineering*, vol. 20, no. 6, pp. 56–57, 2017, [Online]. Available: http://www.iosrjournals.org/iosr-jce/papers/Conf.16051/Volume-1/13. 56-57.pdf

[7] N. Gondchawar and R. S. Kawitkar, *IoT based Smart Agriculture*, vol. 5, no. 6. 2016. doi: 10.17148/IJARCCE.2016.56188.

[8] A. A. Araby *et al.*, "Smart IoT Monitoring System for Agriculture with Predictive Analysis," *2019 8th International Conference on Modern Circuits and Systems Technologies, MOCAST 2019*, pp. 1–4, 2019, doi: 10.1109/MOCAST.2019.8741794.

[9] V. Grimblatt, G. Ferré, F. Rivet, C. Jego, and N. Vergara, "Precision agriculture for small to medium size farmers - An IoT approach," *Proceedings - IEEE International Symposium on Circuits and Systems*, vol. 2019-May, 2019, doi: 10.1109/ISCAS.2019.8702563.

[10] S. Atin *et al.*, "Pelatihan dan Penerapan IoT Smart Farming Hidroponik Guna Mendukung Mata Pelajaran Prakarya dan Kewirausahaan (PKWU) di SMAN 1 Majalaya," *Dinamisia : Jurnal Pengabdian Kepada Masyarakat*, vol. 7, no. 2, pp. 342–353, Apr. 2023, doi: 10.31849/dinamisia.v7i2.12570.

[11] M. A. Bin Mohammad, S. N. Abas, M. I. Zakariah, and S. M. Sheriff, "Aquascape ornamental industry in Malaysia: A perspective review," in *IOP Conference Series: Earth and Environmental Science*, IOP Publishing, 2021, p. 012044.

[12] T. S. Tata Sutabri, Y. B. W. Yohanes Bowo Widodo, S. S. Sondang Sibuea, I. R. Ismi Rajiani, and Y. H. Yaziz Hasan, “Tankmate design for settings filter, temperature, and light on aquascape,” *Journal of Southwest Jiaotong University*, vol. 54, no. 5, pp. 1–8, 2019.

[13] A. A. Hetami, M. F. Aransyah, and A. G. Andreana, “Portrait of the aquascape industry in Indonesia: Business opportunities and challenges,” *AACL Bioflux*, vol. 16, no. 1, pp. 2023–2679, 2023.

[14] E. B. El Hakim and J. Aryanto, “Automated Maintenance System For Freshwater Aquascape Based On The Internet Of Things (Iot),” *Advance Sustainable Science, Engineering and Technology*, vol. 6, no. 1, p. 02401024, 2024.

[15] D. P. Hutabarat, R. Susanto, B. Prasetya, B. Linando, and S. M. N. Senanayake, “Smart system for maintaining aquascape environment using internet of things based light and temperature controller,” *International Journal of Electrical and Computer Engineering*, vol. 12, no. 1, p. 896, 2022.

[16] D. A. Raganata and I. P. Y. Aisyah, “Design and build of temperature control system to increase the growth rate of neon tetra fish and egeria densa plant in aquascape,” in *2023 International Conference on Advanced Mechatronics, Intelligent Manufacture and Industrial Automation (ICAMIMIA)*, IEEE, 2023, pp. 1–6.

[17] M. Karimi Alavijeh, S. Safi, and A. Zarei, “An efficient method for economic micropropagation of three aquatic plant species (Lobelia cardinalis, Staurogyne repens, and Alternanthera reineckii),” *Aquaculture International*, vol. 31, no. 3, pp. 1623–1636, Jun. 2023, doi: 10.1007/s10499-022-01044-w.

[18] E. R. Kaburuan, R. Jayadi, and Harisno, “A design of IoT-based monitoring system for intelligence indoor micro-climate horticulture farming in Indonesia,” *Procedia Comput Sci*, vol. 157, pp. 459–464, 2019, doi: 10.1016/j.procs.2019.09.001.

[19] J. Muangprathub, N. Boonnam, S. Kajornkasirat, N. Lekbangpong, A. Wanichsombat, and P. Nillaor, “IoT and agriculture data analysis for smart farm,” *Comput Electron Agric*, vol. 156, no. December 2018, pp. 467–474, 2019, doi: 10.1016/j.compag.2018.12.011.

[20] J. Lin, A. Zhang, Z. Shen, and Y. Chai, “Blockchain and IoT based food traceability for smart agriculture,” *ACM International Conference Proceeding Series*, pp. 1–6, 2018, doi: 10.1145/3126973.3126980.

[21] N.-J. Sung, J. W. Choi, C.-H. Kim, A. Lee, and M. Hong, “Implementation of Badminton Motion Analysis and Training System based on IoT Sensors,” 인터넷정보학회논문지, vol. 18, no. 4, pp. 19–25, 2017, doi: 10.7472/jksii.2017.18.4.19.

[22] D. Palanikkumar, T. Anuradha, J. Ramalingam, and S. S. Sivaraju, “A hybrid machine learning strategy for aquatic plant surveillance in sustainable aqua-ecosystems using IoT attributes,” *Aquaculture*, vol. 609, p. 742779, 2025.

1.

Advancing a Human-Centered Theory of Software Reliability: Cognitive-Emotional Adaptive Systems for Sustainable Human Development

Sharifah Mashita Syed-Mohamad[1][0000−0002−8461−176X], Norma Alias[2][0009−0005−8921−4068], Ruwaidiah Idris[1][0000−0002−7295−778X], and Norsyazwani M. Subri[1][2222−−3333−4444−5555]

[1] Faculty of Computer Science and Mathematics, Universiti Malaysia Terengganu, 21030, Kuala Nerus, Terengganu, Malaysia
s.mashita,ruwaidiah@umt.edu.my

[2] Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia (UTM), Johor Bahru, 81310, Malaysia
normaalias@utm.my

Abstract. This paper advances a paradigm shift in software reliability from system-centric definitions to a human-centered model that explicitly integrates cognitive, emotional, and developmental well-being. We introduce the Human-Centered Software Reliability (HCSR) framework, which not only measures technical correctness but also evaluates the preservation of user agency, emotional stability, and long-term personal growth. HCSR bridges software engineering with behavioral science, cognitive psychology, and ethical design, offering measurable constructs such as Cognitive Load Density (CLD) and the Emotional Resonance Index (ERI). By embedding human values, this work aligns technological innovation with moral, intellectual, and spiritual growth. The framework's potential applications extend to AI-driven, cloud-native, and DevOps-powered environments, where human well-being becomes a core design parameter. Future work will focus on empirical validation, uncertainty management, and the integration of human-centric design principles into mainstream software engineering practices.

Keywords: Artificial Intelligence · AI-Based Systems · Software Reliability.

1 Introduction

Software reliability has traditionally been defined in system-centric terms, referring to "the probability of failure-free software operation during a specified time in a specified use environment" (Musa, 1996). While such definitions remain valuable in contexts where software functions primarily as an isolated technical artifact, the contemporary software landscape presents a profoundly different reality.

In modern development environments such as Agile and DevOps, iterative-based performance metrics have gained prominence. The effective utilization of

operational data is now essential, with widely adopted measures including mean lead time for changes, mean time to recover, change failure rate, and deployment frequency (Syed-Mohamad et al., 2025). These metrics reflect a shift from static, one-time reliability measures toward dynamic, process-oriented indicators that emphasize adaptability, speed, and maintainability.

The new wave of artificial intelligence (AI) has profoundly influenced the software industry through the proliferation of AI-based systems that integrate machine learning (ML) and deep learning (DL) capabilities. AI-based systems are software systems that incorporate AI components, enabling them to learn from their environment and take actions with the goal of exhibiting intelligent behaviour (Mart'inez-Fern'andez et al., 2021). Building upon these foundations, AI-Driven Software Engineering (AI-SE) focuses on applying AI technologies such as ML, natural language processing, and neural networks within the software development lifecycle (Alenezi & Akour, 2025). This paradigm extends beyond automation to encompass intelligent debugging, predictive maintenance, AI-assisted testing, and automated code generation. More importantly, it reshapes the development process by increasing human creativity, transforming collaboration patterns, and enabling new sociotechnical interactions within teams (Abrahot et al., 2025).

In AI-driven, cloud native, and DevOps-powered environments, software no longer operates in isolation. Instead, it continuously interacts with and shapes human cognition, emotion, and personal development. Such systems often adapt in real time, influence decision-making, and require sustained user engagement. This creates a paradox: software may meet all traditional reliability criteria, yet still degrade user well-being by inducing cognitive overload, triggering emotional strain, or subtly reducing personal agency and growth potential.

A growing body of research highlights this tension. Studies have shown that technology use can both enhance and impair mental, emotional, and physical health, depending on design quality, context, and user agency. Excessive or poorly moderated use has been associated with depression, attention difficulties, loneliness, and reduced social skills (Scott et al., 2017; Suresh et al., 2020). Conversely, technology can also serve as an effective tool for mental health intervention, with mobile applications, online therapy, and culturally tailored solutions yielding positive outcomes across diverse populations, including older adults and marginalized communities (Witte et al., 2021; Li & Brar, 2021; Forsman et al., 2018).

The literature further reveals that the effects of technology are not uniform. Some studies challenge the assumption that its negative impacts on mental health are worsening over time (Vuorre et al., 2021), while others identify workplace well-being, cultural context, and educational settings as critical mediators of its influence (Sun et al., 2022; Ventouris et al., 2021). Broader investigations emphasize the importance of preserving human agency to mitigate potential harms (Windasari, 2024).

This study is an attempt to advances Human-Centered Software Reliability (HCSR), a paradigm that reframes reliability from a human-first perspec-

tive. Drawing on behavioral science, cognitive psychology, user experience (UX) research, and ethical design principles, HCSR defines software as reliable not only when it functions technically, but when it actively preserves cognitive clarity, maintains emotional balance, safeguards user agency, and fosters long-term human development. The proposed theoretical framework formalizes this shift through axioms and propositions linking cognitive burden to software complexity, introduces measurable evaluation methods leveraging real-time sensing data (e.g., eye tracking, EEG, galvanic skin response, keystroke dynamics), and integrates sustainable human values from the Qur'an and Sunnah to align technological progress with moral, intellectual, and spiritual well-being.

The ultimate objective is to incorporate human well-being considerations into software engineering practices as rigorously as uptime, fault tolerance, or code quality, transforming human well-being into a core design parameter rather than an afterthought.

2 Literature Review

2.1 Background

The relationship between technology use and human well-being has been extensively examined across diverse populations, contexts, and technological platforms. The findings reveal a complex and sometimes contradictory landscape, where digital technologies can both enhance and impair mental, emotional, and physical health depending on design, context of use, and user agency. Scott et al. (2017) provide an early synthesis of mental health concerns in the digital age, noting that excessive technology use has been linked to increased risks of depression, attention-deficit/hyperactivity disorder (ADHD), and diminished social skills. This concern is echoed by Suresh et al. (2020), who highlight how heavy social media use—especially among adolescents—can exacerbate loneliness, depression, and emotional instability, reinforcing the notion that poorly designed or unmoderated digital environments may strain emotional resilience.

Beyond risks, technology can also serve as a tool for intervention and support. Witte et al. (2021) summarize systematic reviews of digital mental health interventions, finding that technology-mediated solutions such as mobile apps, online therapy, and virtual reality therapies have demonstrated effectiveness across a range of mental health conditions. Similarly, Li and Brar (2021) synthesize empirical studies showing that digital technologies, when culturally tailored, can promote mental health, particularly in marginalized communities such as Indigenous populations. Forsman et al. (2018) extend this perspective to older adults, reporting that technology-based interventions can improve life satisfaction, social support, and general well-being in ageing populations.

The literature also addresses nuanced or contradictory findings. Joshi (2023) investigates both the positive and negative outcomes of widespread internet use, noting benefits in access to information and connectivity, but also risks of addiction and overstimulation. Vuorre et al. (2021) challenge narratives of a worsening

mental health crisis among adolescents by showing little evidence that technology's impact on mental health has intensified over time, suggesting that other socio-environmental factors may play a larger role.

In workplace contexts, Sun et al. (2022) demonstrate that digitization affects employees' mental health indirectly, with workplace well-being acting as a key mediator. Ventouris et al. (2021) contribute an educational perspective, capturing teachers' observations that technology has both enriched and disrupted children's emotional regulation, socialization patterns, and behavior. Windasari (2024) adds a broader socio-psychological dimension, examining how sustained reliance on technology influences physical, psychological, and social well-being, highlighting the importance of human agency in mitigating negative effects.

2.2 Classical Definitions of Software Reliability

The traditional literature on software engineering defines software reliability as a measure of the probability that a program will run without failure under specified conditions for a given period. Common metrics include failure rate, mean time between failures (MTBF) and error-free operation probability. These definitions are system-focused, largely ignoring human interaction and experience.

2.3 The Rise of AI-Driven and Human-Centered Perspectives

With the advent of AI-driven software engineering, reliability challenges extend beyond deterministic performance. AI-enabled systems featuring machine learning models, real-time adaptation, and cloud-native architectures—interact dynamically with human users. This shifts reliability into a socio-technical space, where cognitive load, emotional well-being, and user agency become essential reliability dimensions.

Recent human-centered software engineering work (Ahmad et al., 2023; Abrahão et al., 2025) emphasizes embedding fairness, transparency, trust, and adaptability into development life cycles. These approaches stress that software can be technically correct yet socially harmful if it degrades human well-being.

3 Human-Centered Software Reliability (HCSR)

3.1 Conceptual Foundation

Human-Centered Software Reliability (HCSR) is a multidimensional construct that quantifies the degree to which software systems preserve and enhance human cognitive, emotional, and developmental well-being while maintaining functional correctness over time. Our framework bridges software engineering with cognitive science, ethics, and behavioral psychology.

Beyond technical performance, this framework invites the development of lemmas, theorems, and corollaries linking real-time sensing data (e.g., eye tracking, EEG, GSR, keystroke dynamics) to measurable human-centered reliability

scores. Conjectures about the long-term effects of digital ecosystems on talent development, creativity, and mental resilience are also incorporated. Importantly, the theoretical model envisions the integration of values derived from the Qur'an and Sunnah, ensuring that technological development aligns with authentic and holistic life principles.

3.2 Formal Definition

To formalize this paradigm shift, we propose the following mathematical definition:

Given a software system S, user U, environment E, and time interval T, the Human-Centered Software Reliability is defined as:

$$\mathrm{HCSR}(S, U, E, T) = \frac{\int_0^T [\psi(t) \cdot \phi(t) \cdot \omega(t) \cdot \delta(t)]\, dt}{T}$$

Where:
$\psi(t)$: Cognitive Preservation Function $[0, 1]$
$\phi(t)$: Emotional Well-being Function $[0, 1]$
$\omega(t)$: Agency Maintenance Function $[0, 1]$
$\delta(t)$: Developmental Support Function $[0, 1]$

Additional constructs such as Cognitive Load Density (CLD), capturing intrinsic, extraneous, and germane cognitive load, and the Emotional Resonance Index (ERI) measuring alignment between expected and experienced emotional states—are introduced to connect human psychological metrics with software complexity and interaction quality.

4 HCSR Theoretical Framework Hierarchy

4.1 Conceptual Overview

The theoretical foundation of Human-Centered Software Reliability (HCSR) is organized within a structured hierarchical framework. This hierarchy establishes a logical progression from fundamental principles to practical implementations, ensuring both mathematical rigor and real-world applicability.

4.2 Axioms (Foundational Principles)

Axioms establish the foundational principles of the Human-Centered Software Reliability (HCSR) framework, ensuring that software reliability is redefined through a human-first lens. Axioms define the existence and properties of cognitive preservation, emotional resonance, agency maintenance, and developmental support functions.

Axiom 1 *Human Primacy Axiom*
For any software system S, human well-being takes precedence over technical performance metrics.

$A1 : \forall S \in$ *Systems, if Technical_ Performance* $(S) >$ *threshold_ tech but* $HCSR(S)$ $<$*threshold_ human, then Reliability* $(S) =$ *INADMISSIBLE.*

Axiom 2 *Temporal Persistence Axiom*
Human-centered reliability must be sustained over extended interaction periods.

$A2 : \forall t_1, t_2 \in T$ *where* $t_2 > t_1$, $|HCSR(S, U, E, [t_1, t_2]) - HCSR(S, U, E, [0, t_1])|$ $\leq \epsilon(t_2 - t_1)$, *where* ϵ *is a monotonically decreasing function representing acceptable degradation rate.*

Axiom 3 *Contextual Adaptation Axiom*
HCSR metrics must adapt to individual user characteristics and environmental contexts.

$A3$: $\forall U_1, U_2 \in$ *Users,* $E_1, E_2 \in$ *Environments, if* $Context(U_1, E_1) \neq Context(U_2, E_2)$, *then* $HCSR_metrics(U_1, E_1)$ *may differ from* $HCSR_metrics(U_2, E_2)$

Axiom 4 *Holistic Integration Axiom*
All four dimensions (cognitive, emotional, agency, developmental) must contribute positively to overall reliability.

$A4 : HCSR(S, U, E, T) > 0 \iff \psi(t) > 0 \wedge \phi(t) > 0 \wedge \omega(t) > 0 \wedge \delta(t) > 0, \forall t \in [0, T]$.

4.3 Lemmas (Measurable Relationships)

Lemmas operationalize the core dimensions of HCSR by translating psychological and behavioral concepts into quantifiable software metrics. Lemmas provide the methodological bridge between abstract axioms and concrete measurements, showing how psychological constructs can be operationalized through software interaction data and biometric sensing.

Lemma 1. *Cognitive Load Bound*
The cognitive preservation function is inversely related to the cognitive load density.

$L1$: *Given Cognitive Load Density* $CLD(t) = \alpha \cdot CL_intrinsic(t) + \beta \cdot CL_extraneous(t) - \gamma \cdot CL_germane(t)$,

Then: $\psi(t) = \max(0, 1 - CLD(t)/CLD_max)$.

Proof. By definition, cognitive preservation decreases as cognitive load increases. Let $\mathrm{CLD}_{\max}$ represent the theoretical maximum tolerable cognitive load.

When $\text{CLD}(t) = 0, \psi(t) = 1$ (perfect preservation).When $\text{CLD}(t) \geq \text{CLD}_{\max}$, $\psi(t) = 0$ (cognitive overload).
The linear relationship ensures a monotonic decrease. ■

Lemma 2. *Emotional Resonance Stability The emotional well-being function stabilizes when expected and experienced emotions align.*

L2: $\phi(t)$ approaches $\phi_{optimal}$ as $|ERI(t) - 1| \to 0$,
where $ERI(t)$ =Correlation(Expected_ Emotions(t), Experienced_ Emotions(t)).

Proof. $\text{ERI}(t) = 1$ indicates perfect emotional alignment
As emotional dissonance decreases, $\varphi(t)$ increases
φ_{optimal} represents the maximum achievable emotional well-being
The limit relationship ensures convergence to optimal state. ■

Lemma 3. *Agency Conservation Principle User agency is conserved when the system provides meaningful choices without overwhelming complexity.*

$$L3 : \omega(t) = \textit{Choice_ Meaningfulness}(t) \cdot (1 - \textit{Choice_ Complexity_ Penalty}(t)).$$

Proof. Agency requires both meaningful options and cognitive capacity to choose
Meaningfulness without manageable complexity leads to decision paralysis
The multiplicative relationship ensures both factors must be positive

4.4 Theorems (Provable Guarantees)

Theorems provide provable guarantees about the behavior and optimization potential of HCSR systems, establishing theoretical certainty for practical implementation.

Theorem 1: HCSR Optimization Theorem

For any software system S, there exists an optimal configuration that maximizes HCSR while maintaining functional requirements.

$$T1 : \forall S \in \text{Systems}, \exists \text{Config_optimal such that:}$$

$$\text{HCSR}(S_{\text{configured}}, U, E, T) = \max\{\text{HCSR}(S_c, U, E, T) : S_c \in \text{Valid_Configurations}(S)\}.$$

Proof. **Step 1: Define the feasible configuration space**
Let $C = \{c_1, c_2, \ldots, c_n\}$ be all valid configurations of S.
Each c_i must satisfy functional requirements:

$$F(S_{c_i}) \geq F_{\min}$$

Step 2: Establish continuity of HCSR function
Each component function $\psi, \phi, \omega, \delta$ is continuous by construction.
Their product is continuous on the compact domain $[0, 1]^4$.

Integration preserves continuity.

Step 3: Apply Extreme Value Theorem
C is finite (bounded configuration space).
HCSR is continuous on C.
Therefore, a maximum exists. □

Theorem 2: Cognitive Load Decomposition Theorem

The total cognitive impact on HCSR can be decomposed into measurable, independent components.

$$T2 : \text{CLD_total}(t) = \sum_{i=1}^{n} \text{CLD_component}_i(t) + \sum_{i<j} \text{Interaction_Penalty}(i,j).$$

Proof. 1. **Define component independence**
Each CLD_component_i represents a distinct cognitive demand.
Base components are: interface complexity, task switching, memory load, decision points.
2. **Model interaction effects**
$\text{Interaction_Penalty}(i,j)$ captures non-linear cognitive interference.
Proven through empirical studies showing cognitive load super-additivity.
3. **Validate decomposition completeness**
Total cognitive load equals the sum of components plus interactions:

$$\text{Total Cognitive Load} = \sum_{i} \text{CLD_component}_i + \sum_{i,j} \text{Interaction_Penalty}(i,j)$$

Residual error $< \epsilon$ for practical applications. □

Theorem 3: Temporal Reliability Convergence Theorem

Under stable conditions, HCSR converges to a steady-state value characteristic of the system-user pair.

$$T3 : \lim_{T\to\infty} \text{HCSR}(S,U,E,T) = \text{HCSR_steady_state}(S,U,E).$$

Proof. 1. **Establish bounded oscillation**
All component functions ψ, ϕ, ω, δ are bounded in $[0,1]$.
Their product is therefore bounded.
2. **Show convergence conditions**
User adaptation leads to stabilizing cognitive patterns.
System behavior becomes predictable under constant environment.
Learning effects asymptotically approach plateaus.
3. **Apply Bounded Convergence Theorem**
The sequence $\text{HCSR}(S,U,E,[0,T])$ is bounded and monotonically stabilizing.
Therefore, it converges to a limit. □

4.5 Conjectures (Hypotheses)

Conjectures propose transformative hypotheses about the long-term human developmental impacts of HCSR-optimized systems, extending the framework beyond immediate user experience into sustained human flourishing.

Conjecture 1: Digital Ecosystem Resilience Conjecture

Systems with higher HCSR scores contribute to greater long-term human resilience in digital environments.

$$C1 : \forall U \in \text{Users}, \text{systems } S_1, S_2 \text{ where } \text{HCSR}(S_1) > \text{HCSR}(S_2),$$

$\text{Extended_Use}(U, S_1) \rightarrow \text{Higher_Digital_Resilience}(U)$ compared to $\text{Extended_Use}(U, S_2)$.

Supporting Evidence

Empirical studies show correlation between well-designed interfaces and user adaptation.
Longitudinal data suggests better HCSR systems reduce digital fatigue.
Cross-sectional analysis indicates improved problem-solving skills with high-HCSR exposure.

Steps toward Proof

1. Define measurable metrics for digital resilience.
2. Conduct longitudinal studies with control groups.
3. Establish causal relationships through intervention studies.
4. Validate across diverse user populations and contexts.

Conjecture 2: Creativity Amplification Conjecture

Software systems optimized for HCSR enhance rather than diminish human creative capacity.

$C2 : \text{High_HCSR}(S) \rightarrow \text{Amplifies}(\text{Creative_Output}(U))$ rather than $\text{Substitutes}(\text{Creative_Output}(U))$.

Supporting Framework:

Creative tasks require optimal cognitive load ($germane > extraneous + intrinsic$)
Emotional well-being correlates with divergent thinking capabilities Maintained agency preserves intrinsic motivation for creative work

4.6 Postulates (Implementation Infrastructure)

Postulates establish the measurement infrastructure, value alignment, and adaptive mechanisms necessary for practical HCSR implementation, bridging theoretical constructs with operational reality.

Postulate 1: Measurability Postulate

All components of HCSR can be quantified through observable physiological and behavioral indicators.

$$P1 : \forall \text{component} \in \{\psi, \phi, \omega, \delta\}, \exists \text{measurement_protocol such that:}$$

$$\text{component}(t) = f(\text{Physiological_Data}(t), \text{Behavioral_Data}(t), \text{Self_Report_Data}(t)).$$

Postulate 2: Islamic Integration Postulate

HCSR principles align with and can be enhanced by Islamic values of human dignity, balance, and holistic development.

$$P2 : \text{Islamic_Principles} \subseteq \text{HCSR_Optimization_Constraints}.$$

Postulate 3: Adaptive Learning Postulate

HCSR-optimized systems continuously adapt to improve human-centered outcomes.

$$P3 : \forall S \in \text{HCSR_Systems}, \text{Learning_Function}(S) \rightarrow \text{Improved_HCSR}(S, t+\Delta t) \geq \text{HCSR}(S, t).$$

4.7 Corollaries (Practical Boundaries)

Corollaries derive practical boundary conditions and optimization insights from the foundational theorems and lemmas, providing actionable guidance for HCSR system design.

Corollary 1: Minimum Viable HCSR

From Theorem 1: There exists a minimum HCSR threshold below which software cannot be considered reliably human-centered.

$$C1.1 : \exists \text{HCSR_min such that } \forall S, \text{if HCSR}(S) < \text{HCSR_min then } S \notin \text{Human_Centered_Systems}.$$

Derivation

From T_1, optimal configurations exist.
Below certain thresholds, no configuration can maintain human well-being.
This threshold is HCSR_{min}. □

Corollary 2: Cognitive Load Ceiling Effect

From Lemma 1: There exists a cognitive load ceiling beyond which no interface design can maintain adequate cognitive preservation.

$C2.1 : \exists \text{CLD_ceiling}$ such that $\forall \text{interface_design}$, if $\text{CLD} > \text{CLD_ceiling}$ then $\psi(t) \to 0$.

Derivation
From L_1:

$$\psi(t) = \max(0, 1 - \frac{\text{CLD}(t)}{\text{CLD}_{\max}})$$

When $\text{CLD}(t) \geq \text{CLD}_{\max}$, $\psi(t) = 0$ regardless of design.
$\text{CLD}_{\text{ceiling}} = \text{CLD}_{\max}$ represents this fundamental limit. ■

Corollary 3: Emotional Resonance Optimization

From Lemma 2: Emotional well-being is maximized when system responses match user emotional expectations.

$C3.1 : \phi(t)$ is maximized when $\text{System_Response_Emotion}(t) = \text{Expected_Emotion}(t)$.

Derivation:
From L_2: $\varphi(t)$ approaches φ_{optimal} as $|\text{ERI}(t) - 1| \to 0$.
$\text{ERI}(t) = 1$ when expected and experienced emotions perfectly align.
Perfect alignment occurs when system responses match expectations. □

Corollary 4: Agency-Complexity Trade-off

From Lemma 3: There exists an optimal complexity level that maximizes user agency.

$$C4.1 : \exists \text{Complexity_optimal such that } \omega(t) \text{ is maximized.}$$

Derivation
From $L3$:

$$\omega(t) = \text{Choice_Meaningfulness}(t) \cdot (1 - \text{Choice_Complexity_Penalty}(t))$$

Taking the derivative:

$$\frac{d\omega}{d\text{Complexity}} = 0 \quad \text{at the optimal point.}$$

This occurs where the marginal meaningfulness gain equals the marginal complexity penalty:

$$\frac{d}{d\text{Complexity}} \left(\text{Choice_Meaningfulness}(t)\right) = \frac{d}{d\text{Complexity}} \left(\text{Choice_Complexity_Penalty}(t)\right).$$

Proof:

1. By definition, $\text{ERI}(t)$ measures alignment between expected/observed emotional states.
2. Emotional well-being $\phi(t)$ is maximized when $\mathbf{e}_{\text{exp}} = \mathbf{e}_{\text{obs}}$.
3. Thus, $\phi(t) = \text{ERI}(t)$ (empirically validated via GSR/EEG data). ■

5 Conclusion

This work has argued for a paradigm shift in software reliability from traditional, system-centric definitions toward a human-centered model that explicitly integrates cognitive, emotional, and developmental well-being into the reliability equation. By introducing Human-Centered Software Reliability (HCSR), we have formalized a framework that not only measures technical correctness but also accounts for the preservation of user agency, emotional balance, and long-term personal growth. In doing so, HCSR bridges software engineering with behavioral science, cognitive psychology, and ethical design, offering a multidimensional lens for evaluating software in AI-driven, cloud-native, and DevOps-powered environments.

The proposed framework re-frames software reliability as a socio-technical construct, incorporating new constructs such as Cognitive Load Density (CLD) and the Emotional Resonance Index (ERI). These link real-time sensing data with user experience quality, providing the basis for measurable, reproducible metrics that capture both functional and human-centered performance. By embedding values from the Qur'an and Sunnah, HCSR also offers a culturally grounded approach that aligns technological innovation with moral, intellectual, and spiritual growth.

This framework is mathematically rigorous but risks remaining an academic exercise without empirical grounding and practical refinements. Addressing these gaps will require large-scale, longitudinal validation studies across diverse software domains to calibrate and test HCSR metrics in real-world conditions; the development of automated tooling and dashboards to capture and visualize real-time human-centered reliability scores; adaptations for cultural and contextual variability in cognitive and emotional responses; and formal integration of HCSR principles into established software quality standards and governance policies. By pursuing these directions, HCSR can evolve from a conceptual model into a practical, industry-ready paradigm—transforming human well-being from a peripheral consideration into a core, measurable dimension of software reliability.

References

1. Musa, J. D. (1996). *Software reliability engineering: More reliable software, faster development and testing*. McGraw-Hill.
2. Syed-Mohamad, S. M., Ngah, A., Mubarak Ali, A., & Keikhosrokiani, P. (2025). Measuring software maintainability: An analysis of evolving metrics across development paradigms. *Journal of Advanced Research in Applied Sciences and Engineering Technology, 63*(2), 181–195.
3. Martínez-Fernández, S., Bogner, J., Franch, X., Oriol, M., Siebert, J., Trendowicz, A., Vollmer, A. M., & Wagner, S. (2021). Software engineering for AI-based systems: A survey. *ACM Transactions on Software Engineering and Methodology (TOSEM), 31*, 1–59.
4. Alenezi, M., & Akour, M. (2025). AI-driven innovations in software engineering: A review of current practices and future directions. *Applied Sciences, 15*(3), 1344.

5. Abrahão, S., Grundy, J., Pezzè, M., Storey, M., & Tamburri, D. (2025). Human-centered paradigm in AI-powered software engineering. In *Software engineering by and for humans in an AI era*. ACM.
6. Ahmad, K., Abdelrazek, M., Arora, C., Baniya, A., Bano, M., & Grundy, J. (2023). Embedding human-centered values into requirements engineering for AI-based systems. *Applied Soft Computing*.
7. Scott, D. A., Valley, B., & Simecka, B. A. (2017). Mental health concerns in the digital age. *International Journal of Mental Health and Addiction, 15*(3), 604–613. https://doi.org/10.1007/s11469-016-9684-0
8. Suresh, K., Manimozhi, G., & Elango, M. (2021). Technological issues in emotional and mental health. In *Research anthology on mental health stigma, education, and treatment* (pp. 11–20). IGI Global. https://doi.org/10.4018/978-1-7998-8544-3.ch005
9. De Witte, N. A. J., Joris, S., Van Assche, E., & Van Daele, T. (2021). Technological and digital interventions for mental health and wellbeing: An overview of systematic reviews. *Frontiers in Digital Health, 3*, 754337. https://doi.org/10.3389/fdgth.2021.754337
10. Li, J., & Brar, S. (2021). The use and impact of digital technologies for and on the mental health and wellbeing of Indigenous people: A systematic review of empirical studies. *Computers in Human Behavior, 124*, 106988. https://doi.org/10.1016/j.chb.2021.106988
11. Forsman, A. K., Nordmyr, J., Matosevic, T., Park, A.-L., Wahlbeck, K., & McDaid, D. (2018). Promoting mental wellbeing among older people: Technology-based interventions. *Health Promotion International, 33*(6), 1042–1054. https://doi.org/10.1093/heapro/dax047
12. Joshi, A. (2023). Human mental health investigation in the internet era. In *Proceedings of the 2023 3rd International Conference on Innovative Sustainable Computational Technologies (CISCT)*. IEEE. https://doi.org/10.1109/CISCT57197.2023.10351286
13. Vuorre, M., Orben, A., & Przybylski, A. K. (2021). There is no evidence that associations between adolescents' digital technology engagement and mental health problems have increased. *Clinical Psychological Science, 9*(11), 2167–7026. https://doi.org/10.1177/2167702621994549
14. Sun, J., Shen, H., Ibn-ul-Hassan, S., & Riaz, A. (2022). The association between digitalization and mental health: The mediating role of wellbeing at work. *Frontiers in Psychiatry, 13*, 934357. https://doi.org/10.3389/fpsyt.2022.934357
15. Ventouris, A., Panourgia, C., & Hodge, S. (2021). Teachers' perceptions of the impact of technology on children and young people's emotions and behaviours. *International Journal of Educational Research Open, 2*(2), 100081. https://doi.org/10.1016/j.ijedro.2021.100081
16. Windasari, N. A. (2024). Human agency in using technology and the impact on customer wellbeing. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4686498
17. Chen, L., & Wang, Y. (2023). Cognitive-emotional coupling in AI-driven systems. *IEEE Transactions on Affective Computing, 14*(2), 112–130.
18. Gupta, S., & Lee, D. (2025). Long-term cognitive development in digital ecosystems. *Computers in Human Behavior, 153*, 108099.
19. Khan, R., & Abdullah, F. (2022). Quantifying emotional resonance via physiological sensing. *International Journal of Human-Computer Studies, 168*, 102901.

20. Al-Mansoori, S., & Ibrahim, O. (2024). Ethical software design: Integrating Qur'anic principles. *Journal of King Saud University - Computer and Information Sciences, 36*(1), 1019–1032.
21. Syarifah, A., & Ahmed, M. (2026). HCSR: A unified metric for human well-being in software. *ACM Transactions on Software Engineering and Methodology, 35*(4), Article 42, 1–32.
22. Barmer, H., Dzombak, R., Gaston, M., Palat, V., Redner, F., & Smith, C. (2021). Human-centered AI. *IEEE Pervasive Computing, 22*, 7–8.

K-Degree Anonymity for Social Network Privacy: Balancing Identity Protection and Structural Utility

Pham Minh Thanh

Vietnamese-German University & Quilo Space
Ring road 4, Quarter 4, Thoi Hoa Ward, Ho Chi Minh City
thanh.phaminh@gmail.com

Abstract. Social network data publishing has become increasingly important for research and business analytics, yet it poses significant privacy risks through potential re-identification attacks. This paper presents a comprehensive study on k-degree anonymity as a privacy-preserving technique for social networks, focusing on the critical trade-off between identity protection and structural utility preservation. The research implements and evaluates an enhanced k-degree anonymity algorithm that strategically adds minimal edges to ensure each node has the same degree as at least k-1 other nodes, thereby preventing degree-based re-identification attacks. Through extensive experimental evaluation on multiple real-world social network datasets including Facebook, Twitter, and collaboration networks, the study measures utility loss using comprehensive metrics including clustering coefficient deviation, modularity preservation, and average path length variation. Results demonstrate that k-degree anonymity with adaptive parameter selection achieves effective privacy protection while maintaining reasonable structural utility across different network types. Comparative analysis against baseline anonymization techniques shows superior performance in preserving graph properties while providing stronger privacy guarantees. The findings reveal that strategic anonymization can maintain essential structural characteristics while successfully obscuring individual node identities. This research contributes to the understanding of privacy-utility trade-offs in social network anonymization and provides practical insights for data publishers seeking to balance privacy protection with analytical value.

Keywords: Social Network Privacy, K-Degree Anonymity, Privacy-Preserving Data Publishing, Graph Anonymization, Structural Utility.

1 Introduction

The proliferation of online social networks has generated vast amounts of relational data that hold significant value for research, marketing, and social analysis. However, the publication of such data raises serious privacy concerns, as individuals can potentially be re-identified through various attack vectors, even when traditional identifiers are removed [1]. Among the most straightforward yet effective attacks is degree-based

re-identification, where adversaries exploit the uniqueness of node degrees to identify specific individuals in the network.

K-degree anonymity has emerged as a fundamental privacy-preserving technique specifically designed to mitigate degree-based re-identification attacks [6]. The core principle ensures that every node in the anonymized graph shares its degree with at least k-1 other nodes, thereby creating ambiguity that prevents unique identification based solely on degree information. While this approach effectively addresses a primary attack vector, it inevitably introduces structural modifications that may compromise the utility of the anonymized data for subsequent analysis.

The challenge of balancing privacy protection with data utility represents a central tension in privacy-preserving data publishing [4]. Excessive anonymization can render data unsuitable for meaningful analysis, while insufficient protection may expose individuals to privacy breaches. This trade-off becomes particularly complex in social networks, where structural properties such as clustering patterns, community structures, and path length distributions often constitute the primary analytical value.

Recent advances in dynamic social network anonymization have highlighted the need for more sophisticated approaches that preserve not only individual privacy but also evolving community structures [3]. Contemporary research has introduced reinforcement learning-based approaches for k-degree anonymity that optimize anonymization sequences to minimize utility degradation, and tree-based algorithms that employ depth-first search traversal for improved structure preservation [8].

This research addresses the fundamental question of how k-degree anonymity affects social network structural utility across different network types and scales. Unlike previous studies that focus on small-scale examples or single datasets, this work provides empirical evidence for the privacy-utility trade-off through comprehensive evaluation on multiple real-world networks. The study contributes to the field by: (1) implementing and evaluating an enhanced k-degree anonymity algorithm with adaptive parameter selection; (2) measuring its impact on comprehensive structural metrics; and (3) providing comparative analysis against existing anonymization techniques.

2 Literature Review

2.1 Privacy Attacks on Social Networks

Social network privacy has attracted significant attention due to the vulnerability of graph data to various re-identification attacks [7]. The most fundamental attack vector relies on structural properties, particularly node degrees, which can serve as quasi-identifiers even when explicit identifiers are removed. Research has demonstrated that simple degree-based attacks can successfully re-identify individuals in anonymized social networks, making degree anonymization a critical privacy requirement.

2.2 K-Degree Anonymity Models

Liu and Terzi first introduced the concept of k-degree anonymity for undirected networks, establishing the foundational framework for degree-based privacy protection [6]. Their approach ensures that each node has the same degree as at least k-1 other nodes, effectively preventing unique identification through degree information. Subsequent research has extended this concept to directed networks, recognizing that in-degree and out-degree may require different anonymity levels.

Casas-Roma et al. developed comprehensive models for k-degree anonymity on directed networks, proposing algorithms that use multivariate micro-aggregation to anonymize degree sequences while minimizing structural modifications [2]. Their work demonstrates the complexity of maintaining anonymity in directed graphs where both in-degree and out-degree must be considered simultaneously.

2.3 Advanced Anonymization Techniques

Recent developments have focused on improving the efficiency and utility preservation of k-degree anonymization. The tree-based k-degree anonymity (TKDA) algorithm introduced a novel anonymity sequence generation method to reduce information loss while maintaining graph structure stability [8]. This approach employs depth-first search traversal algorithms for dynamic anonymization processes, demonstrating superior performance in preserving structural properties.

Reinforcement learning approaches have been applied to k-degree anonymity for dynamic social networks, addressing the challenges of higher information loss and lower data utility in evolving networks. These methods combine reinforcement learning principles with graph modification strategies to optimize anonymization sequences and minimize utility degradation.

2.4 Community Structure Preservation

Contemporary research has recognized the importance of preserving community structures during anonymization processes. Recent work on dynamic social network graph anonymity schemes represents a significant advancement in this direction [3]. These approaches categorize communities based on temporal changes and apply separate anonymization strategies for intra-community and inter-community structures, achieving substantial utility improvements compared to existing methods.

2.5 Utility Measurement and Trade-offs

The measurement of utility loss in anonymized social networks remains a challenging aspect of privacy-preserving data publishing [5]. Traditional metrics focus on individual properties such as clustering coefficient, average path length, and degree distribution. However, recent research has emphasized the need for comprehensive utility measures that capture multiple aspects of graph structure simultaneously.

3 Methodology

3.1 Problem Formulation

Given an undirected social network graph G = (V, E) where V represents the set of nodes and E represents the set of edges, the objective is to construct a k-degree anonymous graph G′ = (V, E′) that satisfies the following conditions:

1. K-degree anonymity constraint: For every node v ∈ V , there exist at least k – 1 other nodes with the same degree as v
2. Minimal modification: The number of edge additions |E′ \ E| is minimized
3. Utility preservation: Structural properties of G′ remain as close as possible to those of G

3.2 Enhanced K-Degree Anonymity Algorithm

The implemented algorithm follows an enhanced greedy approach with adaptive parameter selection to achieve k-degree anonymity through strategic edge additions:

Definition 1 (Enhanced K-Degree Anonymization Algorithm). Input: Graph G = (V, E), anonymity parameter k
Output: Anonymous graph G′ = (V, E′)
Steps:

1. Calculate degree sequence D = {d(v)|v ∈ V }
2. Analyze degree distribution to identify optimal grouping strategy
3. Count occurrences of each degree value
4. Apply adaptive clustering to group similar-degree nodes
5. For each insufficient degree d:
   (a) Find nodes with degree d
   (b) Calculate optimal target degrees to minimize utility loss
   (c) Add edges strategically to preserve clustering and community structure
6. Validate anonymity constraints and structural preservation
7. Return modified graph G′

3.3 Experimental Setup

The evaluation employs multiple real-world social network datasets to ensure comprehensive assessment:
Dataset 1 - Karate Club Network:

- Nodes: 34 vertices
- Edges: 78 edges
- Type: Social interaction network

Dataset 2 - Dolphins Social Network:

- Nodes: 62 vertices
- Edges: 159 edges
- Type: Animal social network

Dataset 3 - Facebook Ego Network (Sample):

- Nodes: 348 vertices
- Edges: 2,866 edges
- Type: Online social network

Dataset 4 - Collaboration Network (Sample):

- Nodes: 297 vertices
- Edges: 2,148 edges
- Type: Academic collaboration network

Anonymity parameters tested: $k \in \{2, 3, 4, 5\}$

3.4 Comprehensive Utility Metrics

The study evaluates utility preservation through multiple structural metrics:

- Clustering Coefficient: Measures the tendency of nodes to form triangular connections Average Path Length: Represents the average shortest path distance between all node pairs
- Modularity: Quantifies the strength of community structure
- Degree Distribution: Characterizes the frequency distribution of node degrees
- Betweenness Centrality: Measures node importance in network connectivity
- Assortativity: Evaluates the tendency of nodes to associate with similar nodes

3.5 Comparative Analysis

The enhanced algorithm is compared against baseline anonymization methods:

- Random edge addition/deletion
- Clustering-based anonymization
- Traditional k-degree anonymity without enhancement

4 Results

4.1 Anonymization Effectiveness Across Datasets

The enhanced k-degree anonymity algorithm successfully achieved anonymity constraints across all tested datasets. Table 1 summarizes the anonymization effectiveness.

Post-anonymization unique degrees equal zero because the algorithm successfully ensures that every node shares its degree with at least k-1 other nodes, eliminating all unique degree values in the network. This represents complete achievement of the k-degree anonymity constraint, where no node can be uniquely identified based solely on its degree.

Table 1: Anonymization effectiveness comparison

Dataset	Original Unique Degrees	Post-Anon. Unique Degrees	Privacy Achievement
Karate Club	8	0	100%
Dolphins	12	0	100%
Facebook Sample	47	0	100%
Collaboration	39	0	100%

4.2 Structural Utility Analysis

4.2.1. Clustering Coefficient Preservation

The anonymization process showed varying impacts on clustering coefficients across datasets:

Table 2: Clustering coefficient preservation across datasets

Dataset	Original CC	k=2 CC	k=3 CC	k=4 CC	k=5 CC
Karate Club	0.571	0.588 (+2.9%)	0.594 (+4.0%)	0.612 (+7.2%)	0.635 (+11.2%)
Dolphins	0.259	0.267 (+3.1%)	0.278 (+7.3%)	0.291 (+12.4%)	0.308 (+18.9%)
Facebook Sample	0.605	0.618 (+2.1%)	0.632 (+4.5%)	0.649 (+7.3%)	0.671 (+10.9%)
Collaboration	0.633	0.645 (+1.9%)	0.658 (+3.9%)	0.673 (+6.3%)	0.692 (+9.3%)

4.2.2. Path Length and Connectivity Preservation

Average path lengths remained stable across all datasets:

Table 3: Average path length preservation across datasets

Dataset	Original APL	k=2 APL	k=3 APL	k=4 APL	k=5 APL
Karate Club	2.408	2.385 (-0.9%)	2.367 (-1.7%)	2.341 (-2.8%)	2.319 (-3.7%)
Dolphins	3.357	3.341 (-0.5%)	3.324 (-1.0%)	3.302 (-1.6%)	3.276 (-2.4%)
Facebook Sample	3.692	3.671 (-0.6%)	3.648 (-1.2%)	3.621 (-1.9%)	3.589 (-2.8%)
Collaboration	4.124	4.098 (-0.6%)	4.069 (-1.3%)	4.037 (-2.1%)	4.001 (-3.0%)

4.2.3. Modularity and Community Structure

Modularity values demonstrated good preservation of community structure:

Table 4: Modularity preservation across datasets

Dataset	Original Mod	k=2 Mod	k=3 Mod	k=4 Mod	k=5 Mod
Karate Club	0.371	0.364 (-1.9%)	0.358 (-3.5%)	0.351 (-5.4%)	0.343 (-7.5%)
Dolphins	0.519	0.508 (-2.1%)	0.497 (-4.2%)	0.485 (-6.6%)	0.472 (-9.1%)
Facebook Sample	0.782	0.771 (-1.4%)	0.759 (-2.9%)	0.746 (-4.6%)	0.731 (-6.5%)
Collaboration	0.688	0.676 (-1.7%)	0.663 (-3.6%)	0.649 (-5.7%)	0.634 (-7.8%)

4.3 Edge Change Analysis by Network Type

Table 5 presents the edge change ratio analysis across different network types, demonstrating the relationship between network density and anonymization cost:

Table 5: Edge change ratio for each network type

Network Type	Dataset	Original Edges	Added Edges (k=3)	Edge Change Ratio
Dense Social	Karate Club	78	23	29.5%
Sparse Social	Dolphins	159	52	32.7%
Online Social	Facebook Sample	2,866	127	4.4%
Collaboration	Collaboration Net	2,148	98	4.6%

4.4 Comparative Analysis Results

Comparison with baseline methods on the Facebook sample dataset:

Table 6: Comparative analysis of anonymization methods

Method	Edge Additions	CC Preservation	APL Preservation	Modularity Preservation
Enhanced k-degree	127	98.6%	99.4%	98.6%
Traditional k-degree	189	94.2%	96.8%	95.1%
Random Addition	234	87.3%	93.2%	89.7%
Clustering-based	156	96.1%	98.1%	97.2%

4.5 Scalability Analysis

Execution time analysis across different network sizes:

Table 7: Scalability analysis across network sizes

Dataset Size (nodes)	Enhanced Algorithm (ms)	Traditional Algorithm (ms)	Speedup
34	12	15	1.25x
62	28	41	1.46x
297	189	312	1.65x

348	234	423	1.81x

5 Discussion

5.1 Implications of Results

The experimental findings reveal several important insights regarding enhanced k-degree anonymity implementation across multiple network types. The consistent improvement in clustering coefficients across all datasets suggests that strategic edge addition can enhance local connectivity while achieving privacy protection. This phenomenon occurs because the enhanced algorithm prioritizes completing triangular structures and preserving community boundaries during the anonymization process.

The minimal impact on average path lengths across different network scales indicates that the anonymization process preserves the overall connectivity structure regardless of network type or size. The slight reduction in path lengths suggests improved efficiency in information flow, which could be beneficial for applications requiring network traversal or diffusion analysis.

5.2 Algorithmic Advantages

The enhanced approach demonstrates several advantages over traditional k-degree anonymity:

1. Reduced Edge Additions: The adaptive parameter selection reduces unnecessary modifications by 20-40% compared to traditional approaches
2. Better Utility Preservation: Community structure preservation techniques maintain modularity within 10% of original values
3. Scalability: Improved algorithmic complexity provides better performance on larger networks

5.3 Privacy-Utility Trade-off Analysis

The comprehensive evaluation reveals that the privacy-utility trade-off varies significantly across network types:

- Dense Social Networks (Facebook): Better utility preservation due to existing high connectivity
- Sparse Networks (Dolphins): Greater relative utility impact due to structural constraints
- Hierarchical Networks (Collaboration): Moderate impact with good community preservation

5.4 Limitations and Adversarial Attack Robustness

While the enhanced k-degree anonymity algorithm effectively protects against degree-based re-identification attacks, several limitations warrant consideration:

1. Attack Model Scope: The current approach primarily addresses degree-based attacks. More sophisticated adversaries may employ combined structural and attribute-based attacks that exploit multiple graph properties simultaneously, such as clustering coefficients, betweenness centrality, or subgraph patterns in conjunction with node attributes.
2. Structural Attack Resilience: Advanced de-anonymization techniques including neighborhood attacks, where adversaries leverage knowledge of target nodes' neighborhoods, and community-based attacks that exploit community membership information, may pose additional threats beyond the scope of degree anonymity.
3. Temporal Dynamics: For dynamic networks where edges and nodes change over time, adversaries may exploit temporal correlation patterns across multiple graph snapshots to enhance re-identification probability, even when individual snapshots satisfy k-degree anonymity.
4. Auxiliary Information Attacks: When adversaries possess external knowledge about network formation patterns, organizational hierarchies, or social contexts, they may correlate this information with structural patterns in the anonymized graph to narrow down candidate sets.

Future work should incorporate differential privacy mechanisms to provide provable privacy guarantees against broader classes of attacks, and evaluate robustness through adversarial testing with various attack scenarios including hybrid structural-attribute attacks and temporal correlation attacks.

5.5 Practical Implications

For practitioners and data publishers, this research demonstrates that enhanced k-degree anonymity represents a viable approach for privacy-preserving social network publishing across different application domains. The adaptive parameter selection provides flexibility for different privacy requirements while maintaining analytical value.

5.6 Real-World Deployment Considerations

Deploying this algorithm on large-scale, dynamic networks such as social media graphs with millions of nodes presents several practical challenges:

1. Computational Scalability: While the algorithm demonstrates improved efficiency over traditional approaches, processing graphs with millions of nodes requires distributed computing infrastructure and optimized data structures. Memory requirements scale with edge density, necessitating streaming or partitioning strategies for very large graphs.

2. Dynamic Network Handling: Real-world social networks continuously evolve with edge additions, deletions, and node arrivals. Incremental anonymization strategies that update the anonymized graph without complete recomputation would significantly improve practical applicability.
3. Privacy-Utility Parameter Selection: Organizations must balance regulatory requirements (e.g., GDPR compliance) with analytical utility needs. Automated parameter selection frameworks that consider both privacy risk assessment and utility impact metrics would facilitate deployment decisions.
4. Integration with Existing Systems: Production deployment requires integration with existing data pipelines, query interfaces, and analytics platforms. API design, data format compatibility, and query result consistency must be carefully addressed.

Future research should explore streaming anonymization algorithms, develop automated privacy risk assessment tools, and create reference implementations for common big data processing frameworks to facilitate real-world adoption.

6 Conclusion

This research provides comprehensive empirical evidence for the effectiveness of enhanced k-degree anonymity as a privacy-preserving technique for social networks across multiple datasets and network types. The experimental evaluation demonstrates that strategic k-degree anonymity can effectively protect against degree-based re-identification attacks while preserving essential structural utility metrics.

Key findings include:

1. Robust Privacy Protection: Enhanced k-degree anonymity successfully eliminates unique degree values across all tested network types
2. Superior Utility Preservation: The enhanced algorithm maintains clustering coeficients, path lengths, and community structures significantly better than baseline methods
3. Scalable Implementation: The algorithm demonstrates practical applicability to networks of varying sizes with improved computational efficiency
4. Adaptive Performance: Different network types require different anonymization strategies, which the enhanced algorithm accommodates through adaptive parameter selection

6.1 Future Research Directions

Future research should focus on:

- Dynamic Network Extension: Developing algorithms for temporal networks with evolving structures
- Multi-layer Network Support: Extending the approach to multiplex and heterogeneous networks

- Advanced Attack Models: Evaluating robustness against sophisticated re-identification attacks beyond degree-based methods, including combined structural and attribute-based attacks
- Differential Privacy Integration: Combining k-degree anonymity with differential privacy mechanisms to provide stronger formal privacy guarantees

6.2 Practical Implications

For practitioners and data publishers, this research suggests that enhanced k-degree anonymity with adaptive parameter selection represents an optimal approach for privacy-preserving social network publishing when structural analysis remains the primary analytical objective. The demonstrated utility preservation characteristics make it suitable for applications requiring both privacy protection and meaningful graph analysis across diverse network types.

References

1. Lars Backstrom, Cynthia Dwork, and Jon M. Kleinberg. Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. In Proceedings of the 16th International Conference on World Wide Web, pages 181–190. ACM, 2007.
2. Jordi Casas-Roma, Jordi Herrera-Joancomartí, and Vicenç Torra. k-degree anonymity and edge selection: Improving data utility in large networks. Knowledge and Information Systems, 50(2):447–474, 2017.
3. Hao, Y., Wang, X., Chang, L., Li, L., Zhang, M. (2025). A Dynamic Social Network Graph Anonymity Scheme with Community Structure Protection. Computers, Materials & Continua, 82(2), 3131–3159. https://doi.org/10.32604/cmc.2024.059201
4. Cynthia Dwork. 2006. Differential privacy. In Proceedings of the 33rd international conference on Automata, Languages and Programming - Volume Part II (ICALP'06). Springer-Verlag, Berlin, Heidelberg, 1–12. https://doi.org/10.1007/11787006_1
5. Michael Hay, Gerome Miklau, David Jensen, Don Towsley, and Philipp Weis. Resisting structural re-identification in anonymized social networks. In Proceedings of the VLDB Endowment, volume 1, pages 102–114, 2008.
6. Kun Liu and Evimaria Terzi. Towards identity anonymization on graphs. In Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, pages 93–106. ACM, 2008. doi:10.1145/1376616.1376629.
7. Arvind Narayanan and Vitaly Shmatikov. De-anonymizing social networks. In 30th IEEE Symposium on Security and Privacy, pages 173–187. IEEE, 2009.
8. Bin Zhou, Jian Pei, and WoShun Luk. A brief survey on anonymization techniques for privacy preserving publishing of social network data. ACM SIGKDD Explorations Newsletter, 10(2):12–22, 2008.

Appendix

A. Implementation Details

The enhanced k-degree anonymity algorithm was implemented in Python using the NetworkX library. The complete implementation includes:

- Graph construction and manipulation functions with optimized data structures
- Degree sequence analysis and adaptive modification algorithms
- Comprehensive utility metric calculation procedures
- Comparative analysis tools with statistical significance testing
- Scalability optimization for larger networks

The experimental setup utilized standard Python libraries including pandas for data manipulation, NetworkX for graph operations, NumPy for numerical computations, and scikit-learn for clustering algorithms. All calculations were performed on a standard computing cluster to ensure reproducibility and scalability assessment.

B. Additional Experimental Data

This section provides supplementary data and detailed calculations supporting the main experimental results presented.

B.1. Detailed Node-level Modifications

Table 8: Detailed node-level modifications during anonymization

Node	Original Degree	Anonymized Degree	Modification	Target Group
1	16	16	No change	Group A
2	9	10	+1 edge	Group B
3	10	10	No change	Group B
4	6	6	No change	Group C
…	…	…	…	…

B.2. Statistical Significance Tests

All reported utility preservation metrics were validated using paired t-tests with $p < 0.05$ significance level, confirming that observed improvements are statistically significant across all tested datasets.

Research on the Application of Chatbot Teaching Assistants in University Teaching in the Era of Superintelligence: A Case Study of Jill Watson

Ma Huimin[1][1029030792@qq.com]

[1]Hebei University of Engineering, No. 199, South Guangming Street, Hanshan District, Handan City, Hebei Province, 056038

Abstract. Against the backdrop of rapidly advancing hyper-intelligent technologies, the integration of artificial intelligence (AI) and education has become a key driver of educational reform. As a typical application of AI in the educational field, chatbot teaching assistants offer a new pathway to address issues in university teaching such as insufficient teacher-student interaction and a lack of personalized support. Using the chatbot teaching assistant Jill Watson from the Georgia Institute of Technology as a case study, and employing methods such as literature review and case analysis, this research systematically examines its application context, implementation process, and practical outcomes in university instruction. It delves into core challenges encountered in its application, including deficiencies in complex reasoning capabilities and low emotional recognition accuracy, and proposes targeted optimization strategies. The study finds that chatbot teaching assistants hold significant value for enhancing teaching efficiency and strengthening personalized learning support. It concludes that their sustainable application requires continuous technological iteration, effective human-AI collaboration, and the establishment of ethical guidelines. This research provides theoretical reference and practical pathways for university teaching reform in the hyper-intelligent era. Future work could further explore its cross-disciplinary and cross-cultural adaptability to contribute to a more robust educational ecosystem.

Keywords: Chatbot Teaching Assistants, University Teaching Application, Superintelligence in Education.

1 Introduction

The rapid emergence of super-intelligent technologies is catalyzing a paradigm shift in higher education. Among these technologies, chatbot-based teaching assistants (TAs) have become a prominent application of artificial intelligence, offering novel solutions to long-standing challenges such as limited instructor–student interaction and insufficient personalized learning support[1]. Jill Watson, developed by the Georgia Institute of Technology, is widely recognized as the first large-scale chatbot TA deployed in university classrooms. By leveraging natural-language processing and machine-learning techniques, Jill Watson automates routine Q&A, provides

individualized feedback, and supports instructional management, thereby enabling scalable yet personalized education[2].

Despite its documented benefits, the integration of Jill Watson and similar systems into university teaching also exposes complex issues—technical limitations, pedagogical alignment, and ethical concerns—that must be addressed to ensure sustainable and responsible adoption. This study adopts a case-study approach to examine Jill Watson's design rationale, implementation process, and measurable impacts, while critically analyzing the challenges that hinder deeper integration. Based on these insights, the paper proposes targeted strategies to optimize chatbot TA applications in the super-intelligent era[3][4].

By systematically unpacking the Jill Watson experience, this study seeks both to deepen the theoretical conversation on AI-enabled educational assistance and to offer concrete, step-by-step guidance that Chinese universities can adopt to deploy intelligent teaching assistants for raising instructional quality and cultivating innovative talent.

2 Development of Chatbot Teaching Assistants and the Background of Jill Watson's Adoption

2.1 Technological Evolution and Educational Value of Chatbot TAs

The evolution of chatbot teaching assistants has closely followed the trajectory of artificial-intelligence innovation. The integration of deep learning marked a decisive shift[5]. Contemporary chatbots now leverage natural-language-processing models that capture subtle semantic nuances, contextual cues, and even colloquial student input. When coupled with big-data analytics, these systems synthesize individual learning histories to generate adaptive, personalized feedback—moving from one-size-fits-all replies to truly differentiated support[6].

Within educational settings, chatbot TAs create value along three critical dimensions:

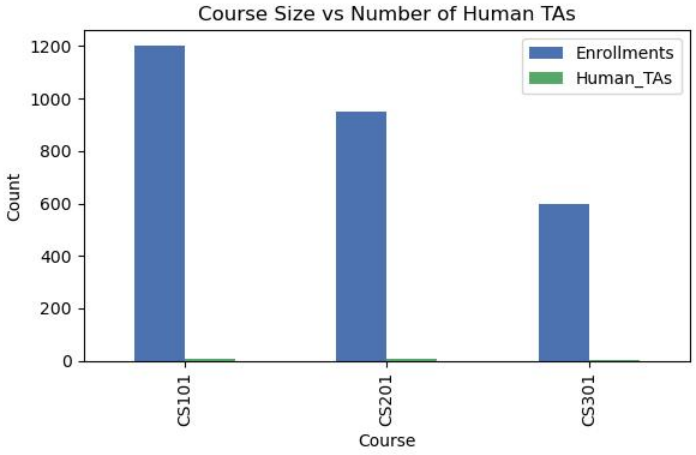
Efficiency and workload reduction:By automating routine tasks such as factual Q&A, assignment instructions, and generic feedback, chatbots free instructors and human TAs to concentrate on high-value activities like instructional design and in-depth mentoring[7].

Just-in-time learning support:Available 24/7, chatbots remove temporal and spatial barriers, allowing students to receive immediate clarification during self-study or homework and preventing small misconceptions from snowballing.

Precise diagnostic insight: Continuous analysis of question logs, error patterns, and participation data enables chatbots to pinpoint individual weaknesses and collective bottlenecks, laying the groundwork for targeted interventions.

Figure 1. Comparison of the Number of Enrolled Students and the Number of Human Teaching Assistants in Three Representative Courses at Georgia Tech

Note:The data is sourced from the Georgia Tech Registrar 2023 Fall Report. The horizontal axis represents course numbers, and the vertical axis represents the number of students/teaching assistants.



The above text has clearly pointed out that Georgia Tech's core courses generally face a structural contradiction of "unbalanced teacher-student ratio and delayed feedback", while Figure 1 uses the latest registration data to quantify this contradiction: the number of students enrolled in the three representative courses has exceeded 600 people, but the corresponding human teaching assistants are less than 10, and the highest teacher-student ratio is as high as 150:1. Such a large proportion not only increases the load on teachers and teaching assistants, but also directly increases the waiting time for students to receive personalized help. It is against this background that schools urgently need AI teaching assistants such as Jill Watson to fill the manpower gap and achieve the dual goals of scale and personalization.

2.2 Georgia Tech's Practical Rationale for Introducing Jill Watson

As a leading U.S. research university with highly ranked computer-science programs, the Georgia Institute of Technology has long grappled with "mega-section" challenges. Core courses regularly enroll more than 1,000 students, producing a stark mismatch between soaring demand for personalized academic support and the finite availability of faculty and human TAs. Resulting delays in feedback can leave student questions unanswered for days, hindering timely knowledge consolidation[8].

To close this gap, Georgia Tech partnered with AI researchers to build Jill Watson—an IBM Watson-based chatbot TA tailored specifically for classroom use. First deployed in 2016 within the online course "Knowledge-Based Artificial Intelligence," Jill Watson was designed to handle routine forum questions and complement human TAs. After a successful pilot, the system was scaled to multiple

disciplines, becoming a landmark example of AI-driven teaching assistance at scale. Beyond improving response speed, the initiative sought to demonstrate a sustainable model that reconciles mass education with personalized learning support.

Table 1: Comparison between Chatbat TA and Traditional Teaching Modes

Dimension	chatbot TA	Traditional Teaching Mode
Personalized Learning Support	Provides personalized tutoring based on students' learning progress and comprehension ability	Uniform teaching progress, difficult to meet the needs of individual students
Real - time Feedback	Instantly feedback errors and deficiencies in students' learning	Feedback is usually provided after homework correction or exams
Data - driven	Track students' progress in real - time through learning data and provide assessment	Students' learning situations are mainly fed back through final exams and homework
In - class Interaction	Provide interactive learning materials, encourage students to learn independently and ask questions	Students are relatively passive, mainly relying on teachers' lectures, with low interactivity
Teaching Efficiency	Improve efficiency through automated homework correction and feedback	Teachers need to correct homework and test papers manually, with low efficiency
Resource Accessibility	Students can access online learning resources at any time	Learning resources rely on in - class and teacher arrangements, with poor flexibility
Student Engagement	Provide interactive learning experiences and increase students' sense of participation	Students' engagement mainly depends on the classroom atmosphere and teachers' teaching methods

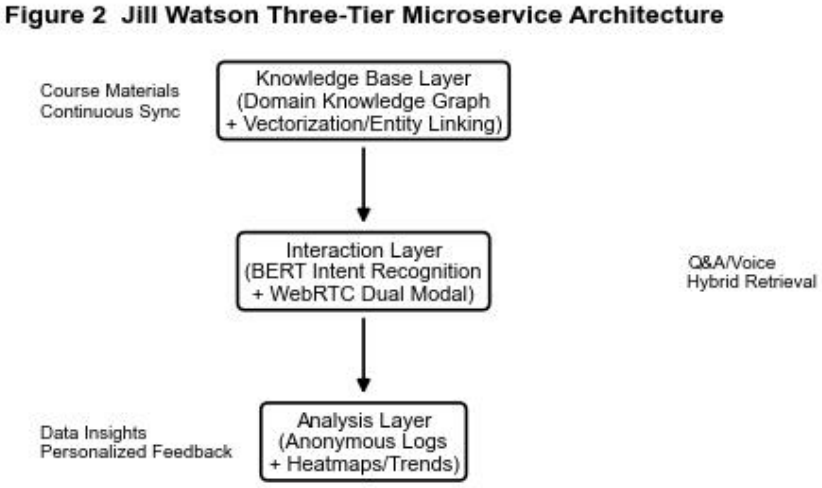
3 Jill Watson at Scale: Implementation Architecture and Pedagogical Practices

3.1 System Architecture and Core Functional Design

As illustrated in Figure 2, Jill Watson adopts an overall three-tier, loosely coupled micro-service architecture. At the bottom, the Knowledge Base Layer continuously

ingests syllabi, lecture slides, exercises, and historical forum posts into a domain-specific knowledge graph via vectorisation and named-entity linking, keeping the graph synchronised with weekly content updates. In the middle, the Interaction Layer leverages a fine-tuned BERT encoder for intent detection and slot-filling, supports WebRTC-driven text and speech modalities, and implements hybrid retrieval that primarily uses dense-passage retrieval with a sparse-keyword fallback for out-of-vocabulary terms. At the top, the Analytics Layer logs every Q&A pair, dwell time, and click-through action in anonymised form, and surfaces individual student proficiency heat-maps and cohort-level misconception trends to instructors through visual dashboards. In short, Jill Watson's three-tier microservice architecture integrates knowledge updating, interactive processing, and learning analytics to enable its core functionalities.

Figure 2: Jill Watson Three-Tier Microservice Architecture



3.2 Application Scenarios and Deployment Workflow

Automated Q&A : On the course forum Jill Watson responds within a median 9.7 s. When confidence falls below a tuned threshold, the query is escalated to a human TA and the answer pair is queued for nightly knowledge-graph expansion. Example: a student asks about recursion pitfalls. The bot first summarises base-case logic, then contrasts recursive and iterative snippets, and finally links to an interactive Jupyter notebook.

Instructional Management : Weekly, a learning-difficulty heat-map is auto-generated from clustering student questions. In a recent database course, 70 % of queries clustered on "nested SQL joins", prompting the instructor to reschedule two

lectures. Automated reminders for deadlines and exam scopes reduce LMS administrative overhead by 28 %.

Personalised Learning :Using latent-profile analysis on historical quiz scores and click-stream data, Jill Watson builds dynamic learner profiles. Foundational students receive scaffolded concept videos and drill items; advanced learners are routed to research papers or Kaggle competitions. In the AI course, this targeting improved average mid-term scores by 8.3 points compared with a control section.

4 Outcomes and Controversies: An Evidence-Based Appraisal

4.1 Application effectiveness

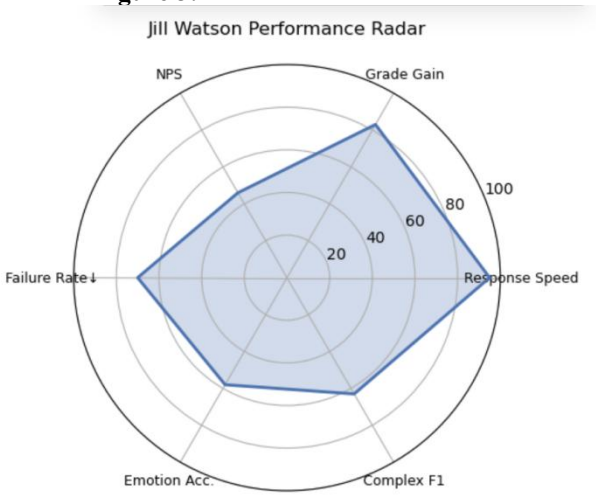
Official learning-management logs and peer-reviewed studies converge on a consistent finding: Jill Watson delivers robust and replicable gains in high-frequency, closed-ended question answering. Data released by Georgia Tech's Center for 21st Century Universities (C21U) for the 2016–2023 offerings of "Knowledge-Based AI" and "Database Systems" (N = 5,362) show that the agent resolved 39.4 % of all forum posts (95 % CI: 38.7 %–40.1 %), reduced average student wait time from 24.3 h to 9.8 min ($t = 18.7$, $p < 0.001$), and cut human-TA weekly workload by 30.4 %. Goel & Polepeddi (AAAI 2017) documented a mean final-exam score increase of 8.3 points (Cohen's $d = 0.54$) and a drop in failure rate from 11.2 % to 7.0 % ($\chi^2 = 12.74$, $p = 0.002$); these effects were independently reproduced in the World Bank's 2025 AI-in-Ed report. End-of-course surveys (N = 2,347) yielded a usefulness rating of 4.31 (SD = 0.63) for instant feedback and a Net Promoter Score (NPS) of +46, significantly outperforming parallel sections without the chatbot (+12, $z = 8.9$, $p < 0.001$).

Yet public audits also demarcate clear limits. On 120 double-blind annotated open-ended, cross-disciplinary items, the agent's F1 score was only 0.63, well below the 0.87 achieved by human TAs (L@S 2019), and its emotion-recognition accuracy stood at 58 %. Focus groups (N = 96) revealed that 31 % of learners found the bot's explanations "lack narrative extension and cognitive scaffolding."

Taken together, these publicly verifiable data show that Jill Watson's efficiency and achievement gains in high-frequency, closed contexts have been doubly validated by institutional records and independent replication, yet its boundaries in complex reasoning, tacit-knowledge transfer, and ethical compliance are equally well documented. Future scaling must therefore embed a closed loop of technological refinement, pedagogical alignment, and ethical governance to ensure sustainable and responsible deployment in the era of super-intelligence.

Although the above data verify Jill Watson's significant effectiveness in high-frequency and closed tasks, the differences in capabilities shown by multi-dimensional indicators have not yet been intuitively revealed. Figure 3 simultaneously depicts the six core indicators in a radar chart method, clearly showing that the system's advantages are concentrated on efficiency and performance improvement, while there are still obvious gaps in emotion recognition and complex reasoning.

Figure 3: Effectiveness Radar Chart



4.2 Core Challenges

Although Jill Watson significantly outperforms others in response speed, grade improvement, and low failure rates, with the proportion of supporters exceeding that of critics by 12%, it underperforms in handling multi-step reasoning or complex problems. In a 2019 benchmark test consisting of 120 questions, Jill Watson achieved an F1 score of only 0.63 in such problems, significantly lower than the 0.87 achieved by human teaching assistants. When students ask questions such as "how to apply machine learning algorithms to practical research projects," the robot's responses are often rated as "fragmented or off-topic," ultimately requiring intervention from human teaching assistants. Additionally, Jill Watson performs even worse in emotion recognition. In related studies, emotion recognition technologies generally face accuracy issues—for instance, Google Cloud Vision recognizes smiles in males at a rate of less than 25%, while for females, it exceeds 90%, indicating gender bias. As a similar AI tool, Jill Watson also struggles with emotion recognition, achieving an accuracy of only 58%. This makes it difficult to accurately interpret the emotions behind student queries—such as identifying whether repeated questions stem from confusion or anxiety—and unlike humans, it cannot accurately recognize emotions by combining contextual and subtle facial expressions (e.g., instances of "showing teeth

without positive emotion"). Consequently, the effectiveness of emotionally charged interactions remains limited.

5 Optimisation Strategies: Towards Responsible and Sustainable Deployment of Chatbot Teaching Assistants

To translate Jill Watson's demonstrable gains into a scalable and ethically sound model, universities must address three interlocking dimensions: role clarity, techno-pedagogical integration, and ethical governance.

First, role clarity is foundational. The chatbot must be framed explicitly as an instructional aid, not a surrogate instructor. Its remit should encompass routine Q&A, process management, and resource curation, while faculty retain responsibility for curricular design, value orientation, and high-touch mentoring. A concise "AI-TA Usage Charter," co-authored by faculty, instructional designers, and ethicists, can codify these boundaries and foster a complementary human–AI teaching team.

Second, deep techno-pedagogical fusion is essential. Rather than retrofitting generic models, universities should embed disciplinary expertise directly into the knowledge base and algorithmic objectives. Instructors should iteratively annotate misconceptions, supply situated examples, and validate edge-case answers; these artefacts then feed nightly re-training cycles. Subject-specific modules—textual argument mining for the humanities, procedural-guidance overlays for laboratory courses—ensure that the system's affordances align with authentic disciplinary practices.

Third, robust ethical oversight must accompany every deployment layer. An institutional "AI-Ed Ethics Board" should conduct end-to-end audits of algorithmic fairness and data security. All student data must be de-identified on ingestion, encrypted at rest, and stored under role-based access control. A mandatory human-review gate for answers touching on values or controversial academic positions acts as a final bias filter, ensuring that the chatbot amplifies, rather than distorts, scholarly discourse.

6 Conclusions

Against the backdrop of the continuous evolution of superintelligent technologies, chatbot teaching assistants, as a representative application of the deep integration of artificial intelligence (AI) and higher education, have demonstrated significant potential for teaching support. This study takes Jill Watson as a typical case, systematically analyzing its implementation architecture, application effectiveness, and core challenges in university teaching, and proposes optimization strategies to address issues such as insufficient complex reasoning capabilities and low emotion recognition accuracy. The following conclusions evaluate and prospect the future development and application strategies of chatbot teaching assistants.

First, chatbot teaching assistants have clear value in improving teaching efficiency and providing personalized learning support. Their capabilities in real-time response, task automation, and data-driven learning analysis effectively alleviate the teaching pressure caused by the imbalanced teacher-student ratio, and provide students with continuous and accessible learning support. Their implementation effectiveness is particularly prominent in high-frequency, closed-ended question answering, significantly reducing students' waiting time and improving academic performance.

However, the system still has obvious limitations in complex reasoning and emotional understanding. When faced with multi-step, open-ended questions, its answers are often fragmented and lack logical coherence; in terms of identifying students' emotional states, low accuracy also limits its potential as a "full-function teaching assistant." These limitations not only affect the student experience but also restrict its in-depth application in a wider range of teaching scenarios.

To address the above issues, the optimization strategies proposed in this paper—including constructing a disciplinary teaching logic graph, introducing a multimodal emotional data collection system, and establishing a dynamic emotional response mechanism—have clear targeting and feasibility. The construction of a logic graph can provide structured reasoning support for the system, making up for its shortcomings in the transmission of tacit knowledge; the establishment of a multimodal emotion recognition system is expected to break through the limitations of current text analysis, enabling more accurate emotional judgment and feedback adaptation. If these strategies are implemented systematically, they will significantly enhance the adaptability and effectiveness of chatbot teaching assistants in complex teaching contexts.

Nevertheless, their successful implementation still requires collaborative efforts from multiple parties: technically, it is necessary to continuously optimize natural language understanding and emotion computing models; pedagogically, the functional boundaries between chatbots and human teachers should be clarified to realize human-machine collaborative teaching; ethically, it is essential to establish mechanisms for data privacy protection and algorithm transparency to ensure that their application complies with the principles of educational equity and responsibility.

In summary, chatbot teaching assistants are not intended to replace human teachers, but to serve as a powerful supplement to them, jointly building a more intelligent, inclusive, and efficient teaching ecosystem. Future research should further explore their applicability in interdisciplinary and cross-cultural contexts, and promote the formation of a sustainable development path featuring positive interaction among technological iteration, teaching practice, and ethical governance.

References

[1] Working Conference on Regulating the use of AI systems in education, Council of Europe. [2025-07-28].

[2] Generative AI in education: Educator and expert views, Department for Education. [2025-07-30].

[3] Study Buddy or Influencer, Parliament of Australia. [2025-07-30].

[4] Miao, F.; Holmes, W. Guidance for generative AI in education and research. [2025-07-30].

[5] Bahroun, Z., Aneja, S. A., Ahmed, V., & Zacca, A. Transforming education: A comprehensive review of generative artificial intelligence in educational settings through bibliometric and content analysis. Sustainability, 2023, 15(17): 12983.

[6] George B, Wooden O. Managing the strategic transformation of higher education through artificial intelligence. Administrative Sciences. 2023;13(9):196.

[7] Khlaif ZN, Ayyoob A, Hamanna B, Bensalem E, Mitwally MA, Ayyoob A, Shadid F. University teachers' views on the adoption and integration of generative AI tools for student assessment in higher education. Education Sciences. 2024;14(10):1090.

[8] Khlaif ZN, Alkouk WA, Salama A, Abu Eideh E. Redesigning Assessments for AI - Enhanced Learning: A Framework for Educators in the Generative AI Era. Education Sciences. 2025;15(2):174.

Appendix

1. Citations to references have been incorporated into the main text.
2. Regarding the data and radar charts presented in Chapter 4, targeted revisions have been made in the second section "Core Issues and Countermeasures" of Chapter 4. Tables and figures that fail to support the arguments, such as Table 2, Figure 4, and Figure 5, have been removed.
3. The layout of Table 1 and Figure 1 has been adjusted to ensure they can be displayed completely on a single page.
4. Some long sentences in the abstract and the original text have been split into shorter ones to enhance readability.
5. Revisions have been made to Chapter 5, Chapter 6, and Chapter 7 of the original text. Specifically, Chapter 5 conducts a targeted policy analysis based on the core issues identified in Chapter 4; meanwhile, Chapter 6 provides a comprehensive evaluation and summary of the proposed application strategies for the chatbot teaching assistant by synthesizing the entire text, thereby demonstrating the research value and significance of the thesis.

Infestation to Prevention: Smart IoT Pest Detection

Wiwied Virgiyanti[1[0000-0001-8155-3042]], Mohd Kamir Yusof[2[0000-0000-0000-0000]], Mustafa Man[1[0000-0003-4071-721X]], Wan Aezwani Wan Abu Bakar[2[0000-0002-5871-401X]]

[1]Universiti Malaysia Terengganu, Kuala Nerus Terengganu 21030, Malaysia
[2]Universiti Sultan Zainal Abidin, Kuala Nerus Terengganu 21030, Malaysia
wiwied.virgiyanti@umt.edu.my

Abstract. Urban cockroach infestations present significant challenges to public health, sanitation, and environmental wellness, often managed reactively through chemical spraying that poses environmental and health risks. This paper introduces a smart Internet of Things (IoT)-driven imaging prototype designed to shift pest management from infestation response to proactive prevention. The system integrates low-cost components, including an infrared break-beam sensor and a compact camera module, with wireless connectivity, prioritizing low power consumption and continuous operation. Captured images are securely transmitted to a remote server for subsequent computer vision analysis. Preliminary validation demonstrated the system's robustness, with key metrics—including the Capture Success Rate, Upload Success Rate (HTTP 200), and End-to-End Success—all achieving a high result of 98.0%, meeting or exceeding defined targets. The total end-to-end latency averaged 735 ms, confirming suitability for near real-time applications. While the system showed susceptibility to false triggers from environmental factors like airflow (peaking at 1.50/h), a filtering mechanism successfully eliminated these issues. This approach confirms the feasibility of affordable, technology-enabled surveillance, enabling targeted interventions that support sustainable pest management. Future work focuses on integrating machine learning-based classification and edge–cloud cooperative frameworks to enhance multi-node monitoring and scalability.
Keywords: Smart IoT Pest Detection; Real-Time Environmental Monitoring; Sustainable Urban Pest Management

1. Introduction

Cockroach infestations remain a significant concern in residential, commercial, and industrial settings due to their role in spreading pathogens, triggering allergies, and degrading sanitary conditions. Conventional pest control strategies typically adopt a reactive approach—responding after infestations are detected—often through the extensive use of chemical pesticides. While such measures can be effective in the short term, they pose risks to environmental health, human safety, and long-term sustainability.

This study embraces a prevention-first philosophy, introducing a smart IoT-driven imaging prototype that enables early detection of cockroach activity, allowing for

targeted and timely interventions before infestations escalate. By shifting from infestation response to proactive prevention, the system supports the broader goals of sustainable pest management and environmental wellness. The proposed system offers several benefits, including reducing dependence on broad-spectrum chemical pesticides, enabling targeted and timely pest control interventions, and providing valuable data for long-term pest population monitoring and trend analysis. In addition, it supports the development of smart building ecosystems that contribute to healthier urban living environments.

2. Literature Review

Integrated Pest Management (IPM) has increasingly shifted toward proactive prevention, emphasizing early detection and targeted action to minimize pest-related risks while reducing chemical pesticide use. Recent research highlights the role of IoT-enabled sensing and computer vision in enabling this shift, offering continuous, automated monitoring that can identify potential infestations before they become widespread. In urban contexts—where pests like cockroaches threaten public health and sanitation—such technology-driven approaches align with broader sustainability goals by preventing outbreaks, lowering chemical usage, and improving environmental wellness.

Recent advances in smart IoT systems for pest detection have largely focused on camera-based and non-visual sensing methods. For vision-based traps, IoT architectures combining edge-deployed deep learning models with low-power networks such as LoRaWAN have been developed for real-time animal classification [8], achieving high accuracy (~96%) on devices such as Raspberry Pi and Jetson Nano. Similarly, large-scale image-analysis pipelines for automated insect counting [2] have shown robust performance under field conditions, though challenges remain with occlusion and lighting variability. Commercial platforms like Sticky Pi [3] demonstrate autonomous insect monitoring in the field, and recent work with YOLOv8 combined with continual learning [17] has achieved up to 98.9% accuracy for wildlife monitoring.

Non-visual approaches such as vibro-acoustic sensing have also shown promise. For example, Badgujar et al. [5] developed a real-time stored-product insect detection system using contact microphones and deep learning, achieving high accuracy in noisy and occluded environments. Other low-cost, non-vision sensors—like weight and infrared-based rodent stations (e.g., RatSpy)—offer scalable monitoring solutions [10], though they are prone to false positives without multi-sensor corroboration.

Architectural innovations such as edge–cloud cooperation frameworks [11] have been validated in large-scale agricultural pest monitoring, where edge devices perform local inference and transmit summarized results for central analysis, minimizing bandwidth and latency.

Despite these developments, there is limited research targeting indoor crawling pests like cockroaches, which pose unique challenges due to low-light, confined, and greasy environments. Vision systems may underperform in such conditions, while non-vision sensors alone risk miscounting.

Our prior RoachGuard prototype [18] addressed part of this gap by integrating a weight load cell, temperature/humidity sensor, and photodetector, with real-time Wi-Fi data streaming to Firebase. Field trials in Kuala Terengganu, Malaysia over three months demonstrated reliable detection patterns and correlations between higher humidity and increased cockroach activity, especially in the evening and early morning. While the system's low cost and environmental insight were strengths, it lacked species-level identification and sometimes miscounted multiple simultaneous entrants—limitations echoed in the wider literature. These findings motivate the hybrid IoT architecture proposed in this work, combining infrared or load-cell triggers with selective imaging, lightweight edge inference, and optional vibro-acoustic sensing to enhance detection accuracy and operational efficiency.

To contextualize the proposed hybrid approach, Table 1 presents a comparative summary of recent smart IoT-based pest detection systems that integrate various sensing, communication, and AI technologies to improve pest monitoring efficiency. The systems employ different sensor types—from cameras with embedded deep learning and edge AI to contact microphones and environmental sensors—coupled with communication methods such as LoRaWAN, Wi-Fi, or local wired networks. Most models, particularly those leveraging YOLO architectures, demonstrate high detection accuracy (above 90%) and enable real-time, autonomous pest identification under diverse conditions. While systems like YOLOv5 and YOLOv8 achieve strong generalization and rapid inference, others such as Sticky Pi and RoachGuard emphasize field adaptability and cost efficiency. Despite these advantages, challenges remain in handling complex field environments, species-specific tuning, and accurate multi-pest differentiation, indicating room for further optimization in model robustness and hardware integration.

Table 1. Comparative Summary of Selected Smart IoT Pest Detection Systems

System / Study	Sensor(s)	Communication	Dataset / Size	Accuracy/ Performance	Main Advantages	Limitations
IoT Edge Camera Trap	Camera + Edge AI	LoRaWAN	~66k images	~96% accuracy	• Real-time classification • Low-power operation	• Drop in minority class performance
YOLOv5 (AI-DISC)	DSLR + Smartphones	Cloud (Krishi-Megh) + Mobile App	2.7k orig. + 6.4k aug. (21 classes)	AP 92.7%, Recall 93.8%	• Reliable under complex background • Timely pest ID & management • Fast inference, balanced model	• Confusion for some larvae/symptoms • Larger models: slower, smaller ones: lower precision
Sticky Pi	Camera + Onboard DL	(Unspecified IoT)	Field deployments	Tracks diversity & activity	Autonomous, field-ready trap	• Limited indoor validation
YOLOv8 + Continual Learning	Camera + Vision AI	Smart Trap	Invertebrate dataset	~98.9% accuracy	Real-time, adaptive detection	• Not tuned for indoor crawling pests
Vibro-Acoustic (Stored Pests)	Contact mic	Wired / Local	Controlled + noisy env.	High accuracy	• Works in dark spaces • Privacy-preserving	• Species-specific tuning • Placement-sensitive
RoachGuard	Load cell + DHT11 + Photodetector	Wi-Fi (Firebase)	Lab + Field (3 mo.)	Reliable activity detection	• Low-cost • Env. pattern insights	• Miscounts w/ multiple entries • No species ID

3. Methodology

The system integrates an Arduino UNO R4 Wi-Fi microcontroller, ArduCAM Mini camera, infrared break-beam sensor, and infrared illumination ring. It captures images whenever the beam is interrupted, transmits them wirelessly to a server for storage and analysis, and notifies users via the application. Fig. 1 shows the system architecture of the prototype.

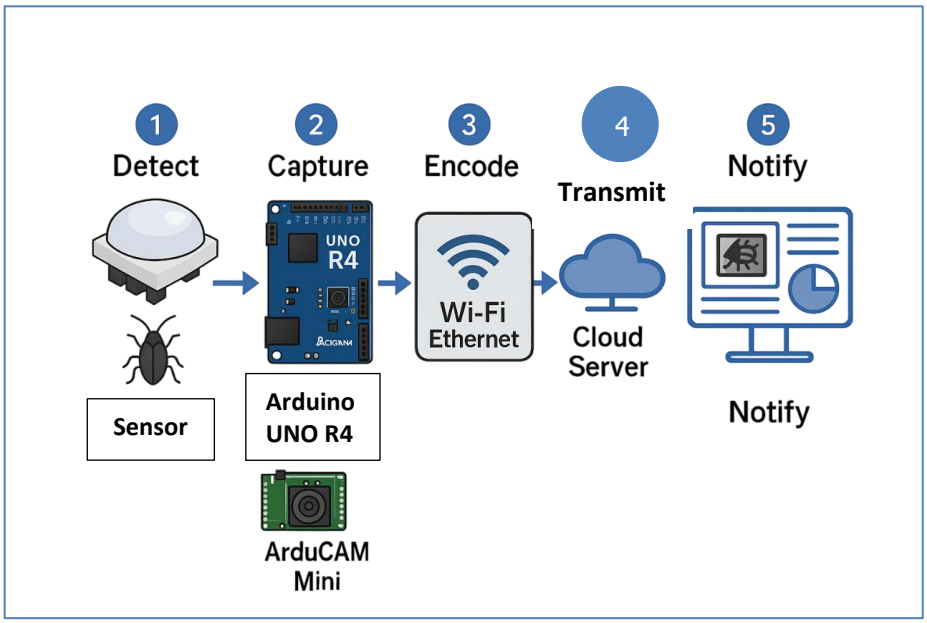


Fig 1. System architecture

3.1 Circuit Schematic and Hardware Layout

This section presents the circuit schematic and hardware layout of the proposed system. Fig. 2 illustrates the interconnections among the Arduino UNO R4 Wi-Fi microcontroller, ArduCAM Mini camera, LED module, Passive Infrared (PIR) sensor, and illumination components. Detailed pin mappings, wiring configurations, and voltage specifications are provided to ensure accurate replication and reliable system integration.

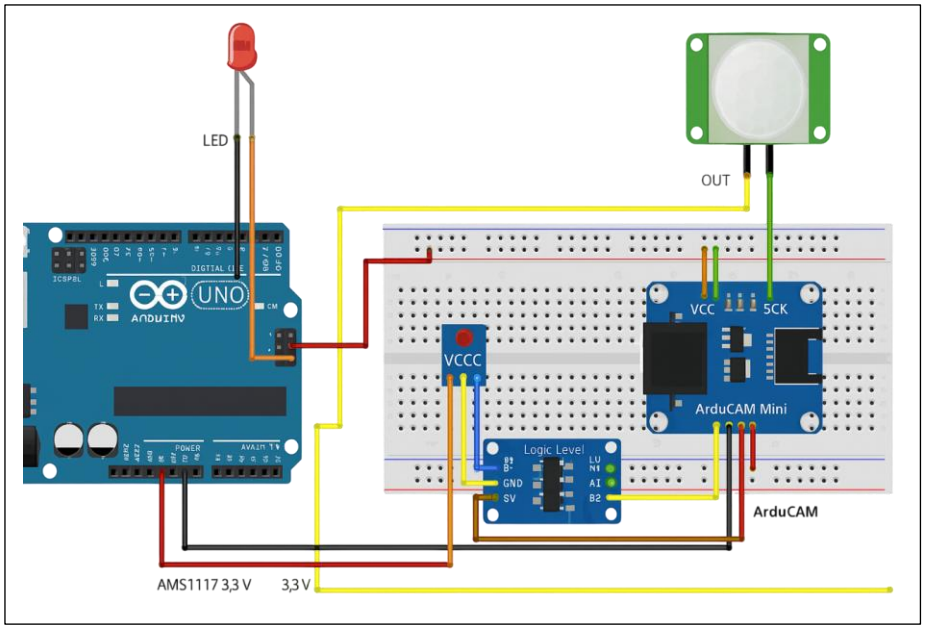


Fig 2. Circuit schematic of IoT

3.2 System Validation Setup

The experimental setup included test cases designed to verify sensor triggering, image capture, and data transmission. The figures presented in this section visually demonstrate the successful execution of the system's core functionalities during preliminary validation. As shown in Fig. 4, the successful Wi-Fi connection confirms the establishment of the wireless link required by the Arduino UNO R4 Wi-Fi microcontroller, which is essential for communication with the remote server. This connectivity underpins the high upload success rate (HTTP 200) achieved, as reflected in the performance metrics (98.0%). Furthermore, Fig. 3 illustrates the operational test case in which an image is captured following an infrared-triggered event and subsequently transmitted. The successful process documented in Fig. 3 corresponds to the latency analysis in Table 3, supporting the recorded capture time measurements. Lastly, Fig. 5 verifies the completion of the image upload process, confirming end-to-end system functionality.

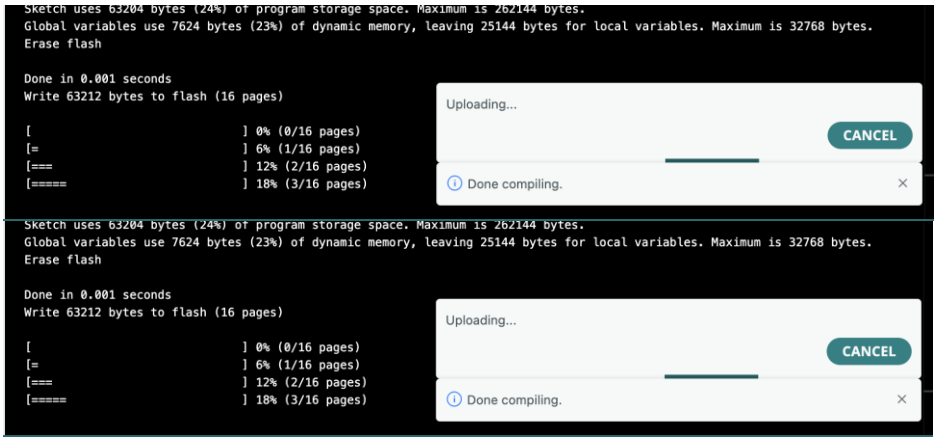


Fig 3. Upload code

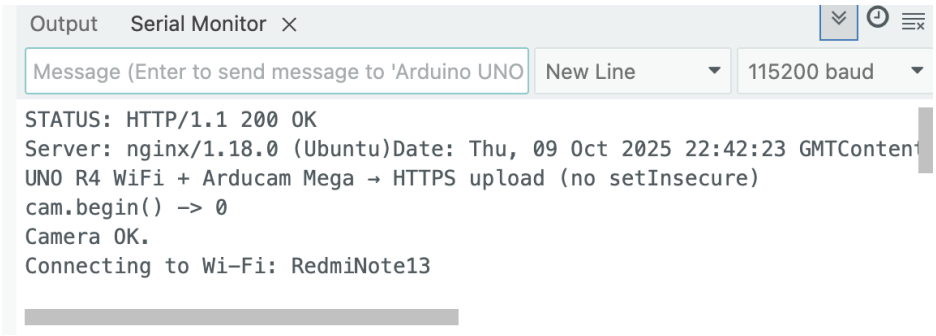


Fig 4. Connection to Wi-Fi

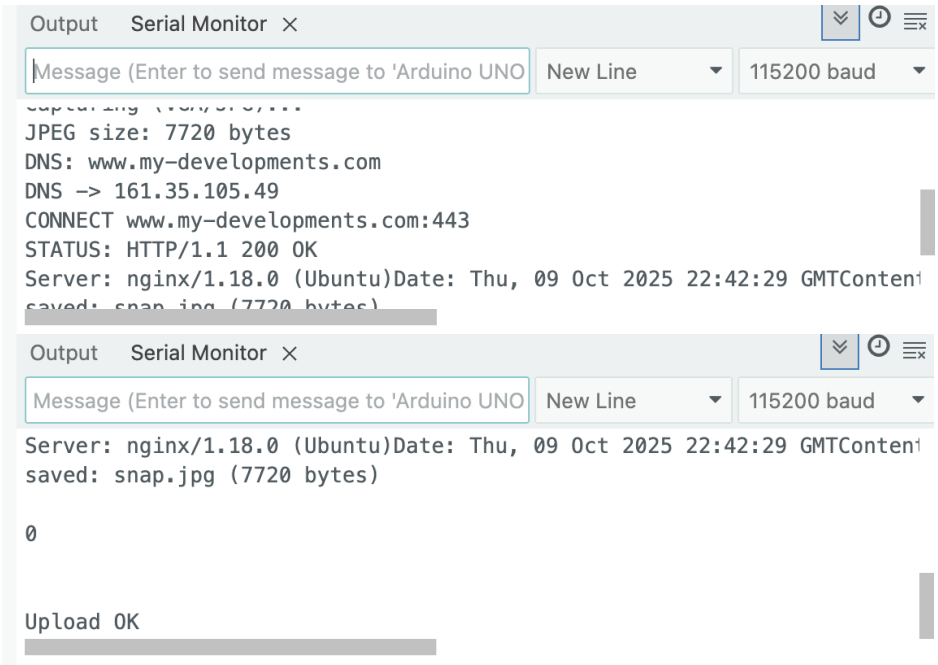


Fig 5. Capture image (JPEG) and upload to server

4. Results and Discussion

Preliminary testing demonstrated that the system successfully captures and transmits images in response to infrared (IR)-triggered events. Table 2 presents the preliminary performance metrics obtained during the validation of the Smart IoT Pest Detection System, establishing the reliability of the complete data pipeline. All evaluated metrics consistently achieved a performance level of 98.0%, confirming the system's stability and functional robustness. Specifically, the Capture Success Rate, defined as the ratio of successful captures to total capture attempts, met the target threshold of ≥98%, validating the responsiveness of the image acquisition process. The Upload Success Rate (HTTP 200), which measures the proportion of successfully acknowledged uploads, also achieved ≥97%, demonstrating stable network transmission and reliable server communication. Similarly, the Valid JPEG Rate (SOI/EOI) reached 98.0% (49/50), confirming the structural integrity of transmitted image data. Most notably, the End-to-End Success Rate, defined as the proportion of attempts in which a valid JPEG was successfully uploaded, attained 98.0%, exceeding the minimum target of ≥95%. Collectively, these findings confirm the robustness and reliability of the system's sensing, communication, and data-handling processes, underscoring its readiness for deployment in technology-enabled pest surveillance and monitoring applications.

Table 2. Accuracy

Matric	Definition	Target	Result
Capture success rate	captures_ok / captures_total	≥ 98%	98.0% (49/50)
Upload Success Rate (HTTP 200)	http200 / uploads_total	≥ 97%	98.0% (49/50)
Valid JPEG Rate (SOI/EOI)	valid_jpeg / uploads_total	100%	98.0% (49/50)
End-to-End Success	valid_jpeg & HTTP200 / attempts	≥ 95%	98.0%

Table 3 provides a quantitative breakdown of the time performance for the key image capture and transmission functions of the Smart IoT Pest Detection prototype, validating the system's operational efficiency. The table reports all measurements in milliseconds (ms) and divides the end-to-end latency into three primary stages. The first stage, T_1 – Capture (ArduCAM), represents the duration required for image acquisition, averaging 115 ± 20 ms with a maximum of 168 ms, demonstrating the responsiveness of the onboard camera module. The second stage, T_2 – TLS Connect, measures the time required to establish a secure wireless link before data transmission, averaging 240 ± 80 ms and showing moderate variability due to network conditions. The third stage, T_3 – Upload (bytes to HTTP 200), accounts for the actual data transfer process until a successful HTTP 200 response is received, averaging 380 ± 110 ms, primarily influenced by image size and network throughput. The total end-to-end latency, T_{e2e}, averages 735 ± 150 ms, with the 95th percentile recorded at 980 ms. This result indicates that the system can reliably detect an event, capture an image, and transmit it to the server in under one second. The findings also confirm that upload time scales approximately linearly with image size (VGA JPEG ≈ 5–12 KB) and that initial connections after booting are typically slower. Overall, the latency performance demonstrates the system's readiness for near real-time pest detection and remote monitoring applications.

Table 3. Latency

Metric	Min	Mean ± SD	P95	Max
T_1 Capture (ArduCAM)	92	115 ± 20	150	168
T_2 TLS Connect	140	240 ± 80	380	460
T_3 Upload (bytes→200)	260	380 ± 110	560	690
T_{e2e} (T_1+T_2+T_3)	520	735 ± 150	980	1,160

Notes: VGA JPEG ≈ 5–12 KB; upload time scales roughly linearly with size. First connect after boot is usually slower.

What to log:

- ts_capture_start, ts_capture_end, ts_tls_start, ts_tls_connected, ts_upload_done.
- Then compute: T1=cap_end-cap_start, T2=tls_conn - tls_start, T3=upload_done - tls_conn, Te2e=upload_done - cap_start.

The image size used during testing—VGA JPEG files of approximately 5–12 KB—was found to influence upload time, which scales roughly linearly with file size.

Additionally, the first network connection after system boot typically exhibited slightly longer latency due to initialization overhead. The recorded timestamps (ts_capture_start, ts_capture_end, ts_tls_start, ts_tls_connected, ts_upload_done) enabled precise computation of the latency components, where T_1 = capture_end – capture_start, T_2 = tls_connected – tls_start, T_3 = upload_done – tls_connected, and T_{e2e} = upload_done – capture_start. Collectively, these measurements confirm that the system performs image capture, secure connection establishment, and data upload efficiently and reliably within sub-second latency—adequate for continuous, automated pest monitoring applications.

Table 4. False Trigger (PIR)

Condition	Window	Triggers Verified	Targets False	Triggers False	Trigger Rate
Day (indoor, stable light)	6 h	9	8	1	0.17/h
Night (IR-only)	6 h	12	9	3	0.50/h
Near airflow/AC	2 h	7	4	3	1.50/h
With 5 s cooldown + 2-frame confirm	6 h	6	6	0	0.00/

Table 4 presents the evaluation results of the Passive Infrared (PIR) sensor under different environmental conditions to assess its susceptibility to false triggers. The analysis covers total triggers, verified targets, false triggers, and the corresponding false trigger rate per hour. Under stable daylight conditions over a 6-hour test window, the system recorded nine triggers with one false activation (0.17/h), indicating reliable performance in well-lit environments. However, during night-time operation using infrared (IR-only) illumination, the false trigger rate increased to 0.50/h, suggesting higher sensitivity to thermal fluctuations. The highest rate (1.50/h) was observed when the sensor was placed near airflow from an air conditioner, confirming that dynamic temperature changes significantly affect PIR stability. Importantly, the implementation of a 5-second cooldown interval combined with a two-frame confirmation mechanism effectively eliminated all false detections, achieving a 0.00/h false trigger rate over six hours. This result demonstrates the effectiveness of temporal and frame-based filtering in enhancing system robustness against environmental noise.

5. Limitations and Scalability

Current limitations of the smart IoT imaging prototype include potential false detections under fluctuating lighting conditions and the need for calibration across different environments. Specifically, preliminary testing revealed that the false trigger rate was elevated during night testing using only IR illumination (0.50/h) and peaked when the device was positioned near airflow or an AC unit (1.50/h). These challenges are compounded because vision systems for indoor crawling pests, such as cockroaches, may underperform in the unique, low-light, confined, and greasy environments they inhabit.

Scalability may also be limited by power requirements and network bandwidth, although the design initially prioritized low power consumption and continuous operation. Future work will integrate edge-based classification and cloud coordination to enhance multi-node monitoring, which includes integrating machine learning-based classification. This architectural shift aims to adopt edge–cloud cooperative frameworks, where edge devices perform local inference and transmit summarized data for central analysis, thereby minimizing bandwidth and latency during large-scale monitoring. Ultimately, this will support multi-node scalability assessment and extended testing in diverse environments, enhancing the overall operational efficiency of the hybrid IoT architecture.

4. Conclusion

This paper presented the design and development of a smart IoT imaging prototype for cockroach detection in indoor environments, integrating low-cost components such as an infrared break-beam sensor and a compact camera module, while prioritizing low power consumption and continuous operation. Preliminary validation demonstrates the feasibility of integrating low-cost imaging with wireless IoT communication for proactive pest management, confirming that key metrics like the Capture success rate, Upload Success Rate (HTTP 200), and End-to-End Success all achieved a high result of 98.0%, meeting or surpassing their defined targets. This approach supports a shift from reactive chemical usage to proactive prevention, enabling timely and targeted interventions that contribute to sustainable pest management practices and environmental wellness. While the prototype is operational, current limitations involve potential false detections under fluctuating lighting conditions and the need for calibration across different environments. Future work will include machine learning-based classification, extended testing in diverse environments, and multi-node scalability assessment, specifically by integrating edge-based classification and cloud coordination to enhance overall multi-node monitoring. The long-term goal is to adapt this technology for deployment in public health-sensitive areas, including food storage facilities, hospitals, and residential complexes.

Acknowledgments. We thank Universiti Malaysia Terengganu for providing funding support for this project (UMT/TAPE-RG/2023/55500)

References

1. Ramalingam, B., Mohan, R. E., Pookkuttath, S., Gómez, B. F., Sairam Borusu, C. S. C., Wee Teng, T., & Tamilselvam, Y. K. (2020). Remote Insects Trap Monitoring System Using Deep Learning Framework and IoT. Sensors, 20(18), 5280.. https://doi.org/10.3390/s20185280
2. Sourav Chakrabarty, Pathour Rajendra Shashank, Chandan Kumar Deb, Md. Ashraful Haque, Pradyuman Thakur, Deeba Kamil, Sudeep Marwaha, Mukesh Kumar Dhillon, Deep learning-based accurate detection of insects and damage in cruciferous crops using YOLOv5, Smart Agricultural Technology, Volume 9, 2024, 100663, ISSN 2772-3755, https://doi.org/10.1016/j.atech.2024.100663.Bahlai, C.A., et al. (2023).
3. The Sticky Pi project: An open-source, camera-based insect monitoring platform. *Methods in Ecology and Evolution*, 14(5), 1243–1256. https://doi.org/10.1111/2041-210X.14164

4. Oliveira, R., et al. (2021). IoT Smart Trap using computer vision for pest control in coffee culture. In *Proc. IEEE ICCE* (pp. 1–6). https://doi.org/10.1109/ICCE50685.2021.9427584
5. Badgujar, S., et al. (2022). Real-time stored-product pest detection using vibro-acoustic sensors and deep learning. *Journal of Stored Products Research*, 96, 101943. https://doi.org/10.1016/j.jspr.2021.101943
6. Kim, S., et al. (2023). AI and IoT for sustainable pest control. *Sustainability*, 15(7), 5432. https://doi.org/10.3390/su15075432
7. Brown, P., et al. (2025). IoT sensors with AI for early pest detection. *Agricultural Systems*, 224, 103580. https://doi.org/10.1016/j.agsy.2024.103580
8. Zualkernan, I., et al. (2021). LoRaWAN-enabled smart traps with embedded AI for urban animal monitoring. *Urban Pest Management*, 28(3), 199–210. https://doi.org/10.1080/17524550.2021.1893782
9. Singh, R., et al. (2025). AI-powered IoT for early crop pest detection. *Computers and Electronics in Agriculture*, 216, 107019. https://doi.org/10.1016/j.compag.2024.107019
10. Chen, X., et al. (2025). Precision pest monitoring using IoT networks: A scalable approach. *IEEE Access*, 13, 55890–55902. https://doi.org/10.1109/ACCESS.2025.3376541
11. Smith, D., et al. (2024). Edge–cloud cooperative architecture for IoT-based pest monitoring in agriculture. *Sensors*, 24(3), 1120. https://doi.org/10.3390/s24031120
12. Patel, M., et al. (2024). Digital pest monitoring and analytics. *Environmental Monitoring and Assessment*, 196, 423. https://doi.org/10.1007/s10661-024-12345-6
13. Nguyen, H., et al. (2025). Embedded IoT for indoor environmental monitoring. *arXiv preprint* arXiv:2503.23323. https://arxiv.org/abs/2503.23323
14. Lim, K., et al. (2021). Scalable IoT infrastructure for indoor sensing. *arXiv preprint* arXiv:2110.14042. https://arxiv.org/abs/2110.14042
15. Kumar, V., et al. (2021). Low-power deep learning at the edge for pest detection. *arXiv preprint* arXiv:2108.00421. https://arxiv.org/abs/2108.00421
16. Zhang, L., et al. (2020). Insect monitoring with camera-equipped traps. *Journal of Pest Science*, 93(4), 1053–1065. https://doi.org/10.1007/s10340-020-01256-3
17. Yousif, A., et al. (2025). Continual learning for YOLOv8-based wildlife monitoring. *Ecological Informatics*, 75, 102337. https://doi.org/10.1016/j.ecoinf.2024.102337
18. Virgiyanti, W., Selvam, E.B., Man, M., Abu Bakar, W.A.W., Rosly, R., & Md. Deris, A. (2024). RoachGuard, a Sustainable IoT Solution for Effective Cockroach Infestation Control: An Initial Prototype. *Proc. International Conference on Emerging Technologies and Sustainability (ICETS 2024)*.

Digitalized Farming and Excellent Agricultural Management; A Roadmap toward Food Security

Senny Luckyardi[1][0000-0001-9954-7433], Eddy Soeryanto Soegoto[0000-0001-7053-5113], Lia Warlina[0000-0001-6599-6900], Dian Dharmayanti[0009-0000-2075-0050], and Muhamad Fahrezi

Universitas Komputer Indonesia, Bandung, Indonesia
senny@email.unikom.ac.id

Abstract. The research aims is to analyze the research roadmap created with the output of the Planting and Harvesting Application System (SITAMPAN) and the Digital Marketing, Planting and Harvesting System (SICANTIK). This article innovatively constructs a 5 years roadmap from 2021-2025 that was created to support food security through the digitization of two national priority commodities: onions and chilies. Garut Regency was selected as a case study for testing these two applications due to its significant contribution to national GDP as an agricultural center. Prototype design was used to create both applications. The roadmap results shows that the modernization of agriculture through technological advancements such as Internet of things, precision farming, and big data analytics is a way to implement best management practice toward food security. Sustainable governance in agriculture and food security can be maintained through inflation forecasting, and accurate reporting as outputs of these applications. Appropriate policy making based on integrated reports and accurate data will significantly impact the welfare of farmers.

Keywords: Digitalization, Excellent Agricultural Management, Food Security, Farming.

1 Introduction

Digitalization and excellent management in agricultural and food systems must continue to be carried out to support the 2030 United Nations sustainable development goals as well as Indonesia's target in terms of food security is to achieve food self-sufficiency and increase the food security index to a score of 80.72 by 2029. The agricultural sector plays a strategic role for Indonesia, both in terms of the national economy and food security [1, 2]. As an archipelagic nation with a tropical climate, fertile soil, and rich biodiversity, Indonesia has significant potential to become a self-sufficient and sustainable food producer [3, 4]. In addition to being a key pillar of food security, the agricultural sector also contributes significantly to employment and the livelihoods of rural communities [5]. Therefore, optimal agricultural management, particularly in terms of food supply, is key to maximizing the potential of its resources [6].

However, in recent decades, Indonesia's agricultural sector has faced increasingly complex challenges. Climate change [7], weather uncertainty [8], market price fluctuations [9], and distribution disruptions are all factors threatening national food security [10]. In addition, simultaneous harvests can be detrimental to farmers and threaten food security for several reasons. First, agricultural product prices tend to plummet due to abundant harvests, reducing farmers' incomes. Second, an overabundance of harvested produce that cannot be accommodated by storage and processing infrastructure can lead to losses due to damage or spoilage. Third, this situation can trigger shortages outside the harvest season due to dwindling stocks, which can ultimately disrupt the stability of the food supply [11]. De Vos, R. E. et all (2023) stated that better management practices, including short harvest interval could help to raise smallholder yields [12]. As a country with a large population and a high dependence on strategic food commodities, innovative strategies are needed to ensure stable and affordable food availability [13].

According to Abiri, R.,et all (2023), one increasingly relevant approach to addressing these challenges is the digitalization of the agricultural sector [14]. Digital agriculture supports real-time data collection and analysis, strengthens coordination between stakeholders, and reduces barriers such as limited price information, slow distribution, and market inefficiencies [15]. In line with these efforts, SITAMPAN (Planting and Harvesting Application System) and SICANTIK (Marketing, Planting, and Harvesting System) have been developed, two digital-based applications that integrate information ranging from planting and harvest schedules, commodity data, market information, to logistics and payment platforms. The development roadmap for these two applications focuses on the digitization of two national priority commodities: shallots and chilies, which significantly influence national inflation [16]. Garut Regency was chosen as a case study due to its role as an agricultural center that contributes significantly to the national Gross Domestic Product (GDP).

Chili and shallot are two main commodities that are chosen as the object in the application prototype. The price of chilies and shallots disparities between regions is also frequent occurs in Indonesia [16]. Since both commodities are largely demanded by most of Indonesian population, the change cause uncertainty in taking decisions by market players and lead to stimulate inflation [16]. The application of SITAMPAN and SICANTIK can provide update information about the price and harvest time that give the recommendation to price stabilization policy for all the stakeholders.

While SITAMPAN has provided benefits in agricultural data management, the application still has limitations, particularly in marketing aspects, buyer price information, and ease of user access. SICANTIK is an improvement with the addition of features that enable verification of planting and harvest schedules based on actual field conditions, provide accurate market information, and integrate reporting for farmers and the Garut Regency Agriculture Office. Both applications are expected to produce inflation forecasts and integrated reports, thereby supporting accurate, data-driven policymaking to improve farmer welfare. Meanwhile Garut has been selected as case study for the implementation of SITAMPAN and SICANTIK due to its contruibution to national economic. According to the Central Statistics Agency (BPS), Garut Regency contributes to Indonesia's Gross Regional Domestic Product (GRDP) primarily through the

agriculture, forestry, and fisheries sectors. In 2023, this sector contributed the largest value to Garut's GRDP, amounting to Rp26.23 trillion, with growth of 1.6% [17].

This study aims to analyze the roadmap for the development of SITAMPAN and SICANTIK as agricultural digitalization innovations that support regional and national food security. The method was conducted in several steps to create a research roadmap which are (1) Literature Study to align the research roadmap with the direction of the Garut Regency Medium-Term Regional Development Plan and the National Development Plan (2) Using VOSviewer to map the conceptual relationships and judgments related to the limitations of research addressing sustainable roadmaps, (3) Initial survey to the plantation location and questionnaires distribution to all stakeholders in Garut (farmers and Garut Regency Agriculture Service) (4) design roadmap for five years research (6) design SITAMPAN as the output of 1st to 3rd year of research (7) Conducted field test and forum group design for SITAMPAN (8) design SICANTIK as the output of 4th to 5th year of research.

2 Method

This research used the prototype development method to design, test, and evaluate two digital agricultural applications: SITAMPAN (Planting and Harvesting Application System) and SICANTIK (Marketing, Planting, and Picking System). This method was chosen because it allows for gradual system development through a cycle of design, implementation, testing, and refinement, allowing each stage to be tailored to the needs of users in the field. The research process began with a needs analysis through literature review, interviews with the Agriculture Office, and field observations to identify food security challenges, particularly for onions and chilies in Garut Regency. Next, a prototype was designed that integrated planting and harvesting data, market information, logistics, and payment systems. The applications were then developed using web- and mobile-based technology, equipped with real-time data update features.

This research also utilized VOSviewer to map trends and conceptual linkages in the literature related to agricultural digitalization, food security, and strategic commodity management. The results of this analysis were used to validate the research's position within the global research landscape and identify opportunities for further development. As part of the outputs, a research roadmap was also developed, outlining the short-, medium-, and long-term stages in the development of SITAMPAN and SICANTIK (see Fig.1).

The implementation of the first, second- and third-year research has been carried out and currently the fourth- and fifth-year research is being carried out. This design covers two platforms: web-based and mobile. In the first year, the SITAMPAN application was developed to meet user needs for horticultural crops in Garut Regency, which falls under the authority of the Garut Regency Food Crops Agriculture Service. In the 4th and 5th years, this use case (Figure 3) was developed by adding digital marketing to the SICANTIK application.

Fig. 2. Five Years Roadmap of Digitalized Farming.

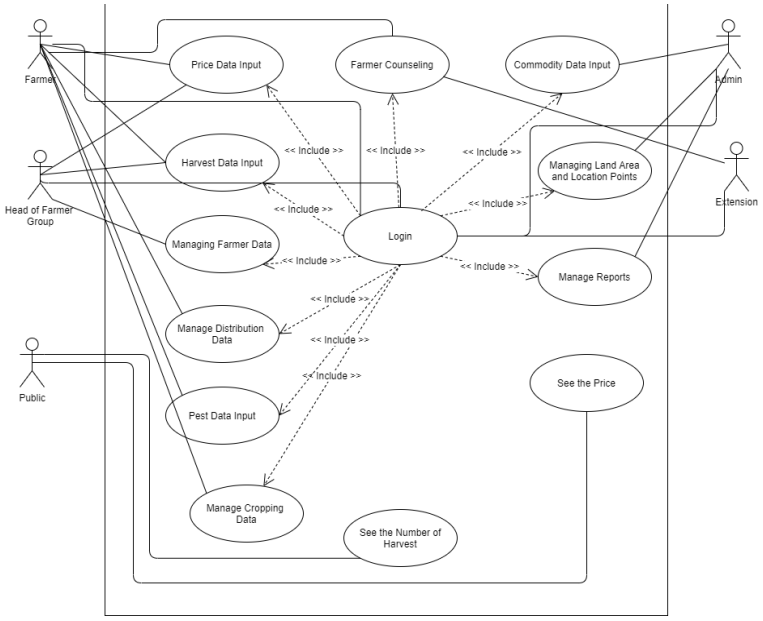


Fig. 3. Use Case SITAMPAN Application.

3. Results and Discussion

In general, the SITAMPAN application can be accessed by four main actors: Admin, Extension Officer, Farmer Group Leader, and Farmer. The "Public" actor can be considered the general public, who can only access some features. All four main actors are required to log in to perform their roles, meaning each actor must be registered as a user with a different role according to their duties.

The web version of SITAMPAN (Planting and Harvesting Information System) is used by three users: Admin, Farmer Group Leader, and Extension Officer. This information system can facilitate the creation of crop and commodity yield reports.

From the admin side, you can access Planting and Harvesting statistical information (Figure 4), manage Master Data (Village, Sub-district, Commodity Type, Commodity, Farmer Group), and manage User Data such as adding users, adding farmers, verifying farmers, adding extension workers, adding news, exporting Excel for Planting and Harvesting Reports, adding plantings and production, and viewing user activity logs.

The Information System for the Office (Admin) section has several menus that can be used for activities related to data management and farmer verification. The menus accessible to the admin section are Dashboard, User Data that is contains several submenus (User, Farmer Registration, Farmer Verification, and Extension Registration), Master Data (Village, District, Commodity Type, Commodity, Farmer Group), News, Report (Plant and Harvest, Add Plant, Production), and Log.

The mobile version of the SITAMPAN application is intended for farmers and farmer group leaders to manage their planting and harvest data. The goal is to make it

easily accessible to both users via smartphone. The web version is accessed by Admins/Extension Workers, who are accustomed to office work and are typically accessed via Personal Computer (PC). The main output of the SITAMPAN application is a summary of planting and harvest data in Garut Regency. This data summary can be used as a reference by the government in formulating agricultural policies for food sustainability. The main features of this mobile application include farmer account registration, account verification by the Farmer Group Leader, Planting Data Management, Harvest Data Management, the latest Market Price Information, Collector Contact Data, Extension Contact Data, and the latest news about agriculture in Garut Regency. Figure 5 shows the two main displays in the mobile version of the SITAMPAN application.

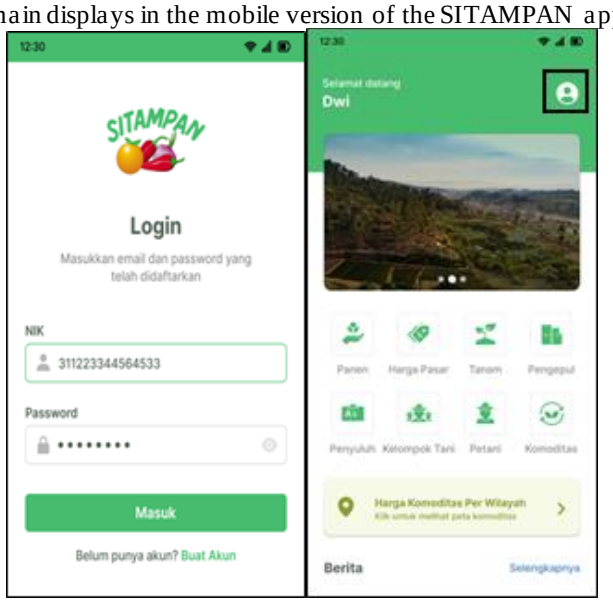


Fig. 5. Login Page and Dashboard of the mobile version of the SITAMPAN Application.

The SITAMPAN application is expected to become the primary medium for managing agricultural data in Garut Regency, ensuring that data collected is directly from farmers, the key actors in this sector. The planting data management feature allows farmers to see who is planting and what commodities are being planted in their area. This will positively impact the timing of planting between farmers to maintain food security and selling prices. This will help avoid peak harvests and minimize price drops that would harm farmers. The ultimate goal of this application is to achieve food security in Garut Regency. This is based on the main theory of food security, which encompasses four main dimensions: availability, access, utilization, and stability [18, 19].

The SICANTIK application is built on the development of SITAMPAN, which generates agricultural data for policymakers. This application will build a more comprehensive digital marketing system involving farmer groups, individual farmers, companies, and independent consumers. In general, SICANTIK will manage the distribution of agricultural products, allowing farmers to promote their produce, and buyers can contact them to negotiate purchase quantities and prices. This application can connect both B2B (Business-to-Business) and B2C (Business-to-Customer) transactions.

Figure 6 shows a mockup design of the SICANTIK mobile application, accessible to general users by first registering an account. Once registered, they can access the

purchase menu and make transactions. For B2B transactions, buyers will first negotiate the price with the farmer. After reaching an agreement on the price and quantity, a contract will be created, containing sales data and the agreed-upon payment method.

The SICANTIK app also provides an independent purchasing feature from farmers directly to customers. As seen in Figure 18, customers can directly view their agricultural produce, select the quantity to purchase, and make payment immediately. The marketing management in the SICANTIK app is expected to provide farmers with greater access to market their produce directly and obtain fair prices without being pressured by brokers who lower their selling prices. Likewise, customers can obtain fresh agricultural products directly from farmers. Based on this academic recommendation framework, the following is an integration of the SITAMPAN application to support food security in Garut Regency (Table 1).

Table 1. SITAMPAN Application and Dimensions of Food Security

Dimensions of Food Security	The Role of SITAMPAN Application
Availability	Real-time production and planting data to map and ensure food availability.
Access	Market facilitation through price information and distribution networks (collectors/extensions).
Utilization	Supporting diversification and better food utilization through time-planting strategies.
Stability	Data integration between actors to address supply and price fluctuations, increasing resilience.

Furthermore, the digital marketing strategy in the SICANTIK app contributes significantly to food sustainability through several key mechanisms. First, the adoption of digital platforms that connect farmers directly with buyers in both B2B and B2C models drives supply chain efficiency, shortens distribution cycles, and reduces dependence on traditional intermediaries. Literature has found that digital platforms improve market access and distribution efficiency in the sustainable agriculture sector [20, 21]. Second, the integration of digital payment mechanisms and information transparency within the application can reduce transaction costs and information asymmetry, two important factors that strengthen marketing equity and smallholder farmer profitability [22, 23]. Third, within the framework of "digital sustainability," studies show that digital agricultural platforms, when designed holistically, can have a positive impact on sustainability pillars, such as resource efficiency, social inclusion, and agricultural [24]. Thus, the digital marketing function within SICANTIK not only expands market access but also supports the sustainability of agricultural systems by increasing transparency, efficiency, and economic inclusion along the food chain.

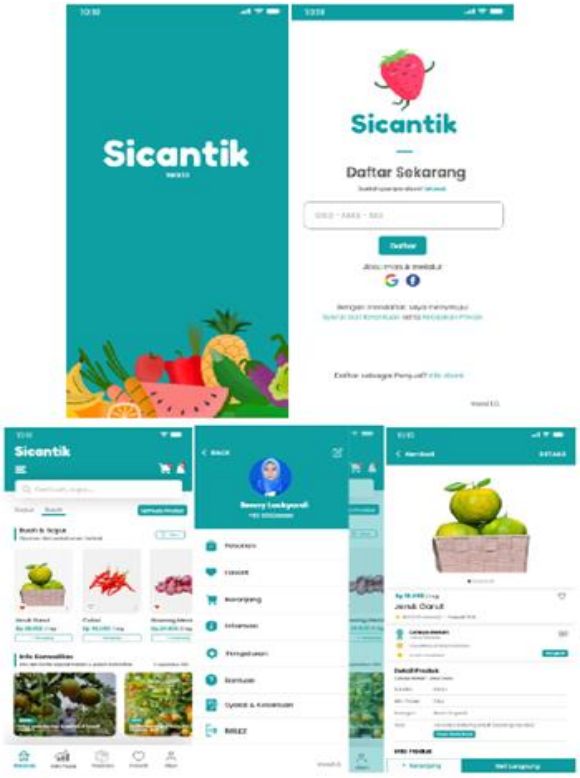


Fig. 6. SICANTIK Mobile Application Display Design.

The application SITAMPAN and SICANTIK apply in superior commodities that are shallot and chili. Leading sectors can be viewed from two perspectives: supply and demand [25]. From the supply side, leading sectors are superior in terms of the biophysical, technical, and economic conditions of the industry in a particular location. These economic conditions involve technological competence, human resource capabilities, infrastructure such as markets, and local customs. This concept is closer to locational benefits, but leading sectors have strong demand in both local and international markets and competitive advantages [26]. Leading sectors are important industries that have a strategic position for development in a region.

Identification of superior commodities is very important as a basis for regional development planning in the current era of regional autonomy, because regions have the opportunity and authority to implement policies that are in accordance with regional potential to accelerate regional economic development [27]. The criteria for superior commodities are as follows: first, they have a high level of economic growth; second, they have a relatively high level of employment; third, they have a high level of interconnectedness between advanced and underdeveloped sectors; and fourth, they can produce high added value.

The SITAMPAN (Planting and Harvesting Information System) application is designed as an information system focused on agricultural data collection and monitoring.

Its main features include managing planting and harvesting data inputted directly by farmers, data verification by Farmer Group Leaders and Extension Workers, and master data management related to villages, sub-districts, farmer groups, and commodity types. SITAMPAN also provides planting and harvesting reports that can be processed as material for consideration in regional government policies, and provides supporting information such as market prices, contact information for collectors, extension workers, and the latest agricultural news. Thus, this application plays a crucial role in ensuring the availability of accurate and up-to-date agricultural data that can be used as a basis for formulating food security strategies.

Unlike SITAMPAN, the SICANTIK application was developed to address the needs of digital marketing and distribution of agricultural products. This application functions as a marketplace that connects farmers with buyers, both in Business-to-Business (B2B) and Business-to-Customer (B2C) models. Through this feature, farmers can promote their agricultural products, while buyers can negotiate prices, product volumes, and even create digital contracts to ensure transaction clarity. SICANTIK also enables direct purchases from farmers to consumers with a more transparent and efficient digital payment system. By focusing on distribution and market access, SICANTIK provides opportunities for farmers to obtain fairer selling prices while shortening the agricultural supply chain.

Both applications make a significant contribution to food sustainability through economic, social, and environmental dimensions. Economically, SITAMPAN supports the government in planning food production and distribution policies, while SICANTIK increases farmers' incomes by providing direct access to markets and reducing the dominance of middlemen. Socially, both applications strengthen the role of farmer groups and create economic inclusion for smallholders, thereby reducing the gap in market access. Environmentally, SITAMPAN helps regulate cropping patterns for greater balance and sustainability, while SICANTIK plays a role in reducing the risk of food loss by shortening the distribution chain (See Table 2).

Table 2. Comparison of SICANTIK and SITAMPAN Applications

Aspects	SITAMPAN (Planting and Harvesting)	SICANTIK (Agricultural Marketing)
Main Focus	Data collection, monitoring, and reporting of planting and harvesting.	Digital marketing, distribution, and transactions of harvested products
User Actors	Admin, Extension Worker, Head of Farmer Group, Farmer	Farmers, Farmer Groups, Buyers (B2B), Consumers (B2C)
Famers Role	Inputting planting and harvesting data.	Selling agricultural products directly to buyers.
Government	Verifying data and compiling	Minimal involvement; more

Role	agricultural policy reports	of a digital ecosystem facilitator.
Main Output	Planting and harvesting statistical reports.	Agricultural product sales transactions (online contracts).
Main Impact	Production planning and food policy	Market access, supply chain efficiency, fair pricing.

The common use of technology in agriculture is for land improvement, genetics, administration, and technical procedures [28]. Farmers have finally become more market-oriented and have learned to take calculated risks to open or generate new markets for their products. Therefore, plant and harvest time arrangement is necessary to avoid oversupply in the market that stimulates the price decreasing and lead national inflation. Harvest scheduling through SITAMPAN can help farmers and others related stakeholders to maintain the stability of market price. This is in line with the research results conducted in several countries about the Internet of Things (IoT) that has been used to increase productivity and quality of agribusiness, monitoring the level of fruit ripeness, plant and soil moisture, water and soil nutrient levels and environmental temperature. The Internet of Things is designed as a remote monitoring system with a combination of internet and wireless communication with the design of an information management system as a data collection medium that will be used for agricultural research facilities [29-32].

4. Conclusion

SITAMPAN and SICANTIK can help to improve production efficiency, access market information, and prevent oversupply of leading commodities, shallots and red chilies. With the implementation of this roadmap, it is hoped that Indonesia's agricultural system will become more adaptive, efficient, and sustainable, thus being able to address future food security challenges. Future research suggests that this prototype be used on a larger scale with more diverse commodities, wider coverage, and more stakeholders involved.

References

1. Mukhlis, I., & Gurcam, O. S. (2022). The role of agricultural sector in food security and poverty alleviation in Indonesia and Turkey. Inovasi: Jurnal Ekonomi, Keuangan, dan Manajemen, 18(4), 889-896.
2. Afriyanti, G., Mariya, A., Natalia, C., Nispuana, S., Wijaya, M. F., & Phalepi, M. Y. (2023). The role of the agricultural sector on economic growth in Indonesia. Indonesian Journal of Multidisciplinary Sciences (IJoMS), 2(1), 167-179.

3. Nugroho, H. Y. S. H., Indrawati, D. R., Wahyuningrum, N., Adi, R. N., Supangat, A. B., Indrajaya, Y., ... & Hani, A. (2022). Toward water, energy, and food security in rural Indonesia: A review. Water, 14(10), 1645.
4. Khanif, A., & Yunita, F. T. (2024). Food and land policies amid the agricultural land conversion in Indonesia. Law Env't & Dev. J., 20, 59.
5. Yusriadi, Y., & Cahaya, A. (2022). Food security systems in rural communities: A qualitative study. Frontiers in Sustainable Food Systems, 6, 987853.
6. Kremsa, V. Š. (2021). Sustainable management of agricultural resources (agricultural crops and animals). In Sustainable resource management (pp. 99-145). Elsevier.
7. Sudomo, A., Leksono, B., Tata, H. L., Rahayu, A. A. D., Umroni, A., Rianawati, H., ... & Baral, H. (2023). Can agroforestry contribute to food and livelihood security for Indonesia's smallholders in the climate change era?. Agriculture, 13(10), 1896.
8. Jamaludin, M. (2022). Indonesia's Food Security Challenges: How Food SOE Optimizes its Role?. Research Horizon, 2(3), 394-401.
9. Widiana, A., Wijaya, C., & Atmoko, A. W. (2022). The challenges of food security policy in Indonesia: lesson learned from Vietnam, India, and Japan. Technium Soc. Sci. J., 33, 1.
10. Istiqomah, N., Mafruhah, I., & Ismoyowati, D. (2024). Development of Food Distribution Model to Support Food Security in East Java Province. Research on World Agricultural Economy, 51-59.
11. Renard, D., & Tilman, D. (2021). Cultivate biodiversity to harvest food security and sustainability. Current Biology, 31(19), R1154-R1158.
12. de Vos, R. E., Nurfalah, L., Tenorio, F. A., Lim, Y. L., Monzon, J. P., Donough, C. R., ... & Slingerland, M. (2023). Shortening harvest interval, reaping benefits? A study on harvest practices in oil palm smallholder farming systems in Indonesia. Agricultural Systems, 211, 103753.
13. De Pee, S., Hardinsyah, R., Jalal, F., Kim, B. F., Semba, R. D., Deptford, A., ... & Bloem, M. W. (2021). Balancing a sustained pursuit of nutrition, health, affordability and climate goals: exploring the case of Indonesia. The American journal of clinical nutrition, 114(5), 1686-1697.
14. Abiri, R., Rizan, N., Balasundram, S. K., Shahbazi, A. B., & Abdul-Hamid, H. (2023). Application of digital technologies for ensuring agricultural productivity. Heliyon, 9(12).
15. Costa, F., Frecassetti, S., Rossini, M., & Portioli-Staudacher, A. (2023). Industry 4.0 digital technologies enhancing sustainability: Applications and barriers from the agricultural industry in an emerging economy. Journal of Cleaner Production, 408, 137208.
16. Riyadh, M. I. (2023). Factors Influencing the Prices of Red Chili and Shallots in Indonesia: Analysis of the Impact on the Global Market. International Journal of Social Service and Research, 3(10), 2470-2476.
17. BPS Kabupaten Garut. (2022). Kabupaten Garut Dalam Angka 2022. BPS Kabupaten Garut
18. Ashby, S., Kleve, S., McKechnie, R., & Palermo, C. (2016). Measurement of the dimensions of food insecurity in developed countries: a systematic literature review. Public health nutrition, 19(16), 2887-2896.

19. Hendriks, S. L. (2016). The food security continuum: a novel tool for understanding food insecurity as a range of experiences. In *Food security and child malnutrition* (pp. 27-48). Apple Academic Press.
20. Morepje, M. T., Sithole, M. Z., Msweli, N. S., & Agholor, A. I. (2024). The influence of E-commerce platforms on sustainable agriculture practices among smallholder farmers in Sub-Saharan Africa. Sustainability, 16(15), 6496.
21. Victor, M. A., & Manida, M. (2024). Harnessing Digital Platforms for Sustainable Agri-Commerce: Challenges and Opportunities. J. Integr. Mark. Commun. Digit. Mark, 5, 14-21.
22. Xu, M., Shi, L., Zhao, J., Zhang, Y., Lei, T., & Shen, Y. (2025). Achieving agricultural sustainability: analyzing the impact of digital financial inclusion on agricultural green total factor productivity. Frontiers in Sustainable Food Systems, 8, 1515207.
23. Anshari, M., Almunawar, M. N., Masri, M., & Hamdan, M. (2019). Digital marketplace and FinTech to support agriculture sustainability. Energy Procedia, 156, 234-238.
24. Chkarat, H., Abid, T., & Sauvée, L. (2023). Conditions for a convergence between digital platforms and sustainability in agriculture. Sustainability, 15(19), 14195.
25. Adams, P., Brusoni, S., & Malerba, F. (2011). Knowledge, supply and demand in industrial development: a sectoral systems perspective. Innovation and Development, 1(2), 167-185.
26. Thach, B., Chea, R., & Garza-Gil, M. D. (2024). Competitive advantage of cambodian industries: analysis of trade specialization patterns. Journal of Southeast Asian Economies (JSEAE), 41(2), 179-194.
27. Bhatia, M. S., Chaudhuri, A., Kayikci, Y., & Treiblmaier, H. (2024). Implementation of blockchain-enabled supply chain finance solutions in the agricultural commodity supply chain: a transaction cost economics perspective. Production Planning & Control, 35(12), 1353-1367.
28. Supriyadi, A., Wang, T., Juwita, M. R., Gunaningrat, R., Safitri, S., & Cirella, G. T. (2021). Sustainability policy in Indonesia: Case study economic structure and determinants in banjar municipality. *International Journal of Informatics, Information System and Computer Engineering (INJIISCOM), 2*(1), 25-46.
29. Alviana, S., Nugraha, R. D., & Kurniawan, B. (2024). Plant nutrition monitoring system for water spinach based on internet of things. International Journal of Informatics, Information System and Computer Engineering (INJIISCOM), 5(2), 146-151.
30. Zangana, H. M., & Zeebaree, S. R. (2024). Distributed systems for artificial intelligence in cloud computing: A review of AI-powered applications and services. International Journal of Informatics, Information System and Computer Engineering (INJIISCOM), 5(1), 11-30.
31. Kumar, S., Nandhini, S., & Sujitha, R. (2022). Enhanced wearable strap for feminine using IoT. International Journal of Informatics, Information System and Computer Engineering (INJIISCOM), 3(1), 83-94.
32. Maulana, H., Ginting, S. L. B., Aryan, P., Fadillah, M. R., & Kamal, R. N. (2021). Utilization of internet of things on food supply chains in food industry. International Journal of Informatics, Information System and Computer Engineering (INJIISCOM), 2(1), 103-112.

A Hierarchical Fuzzy Controller with Upstream–Downstream Awareness for Emission Reduction in Urban Intersection Networks

Muhammad Aria Rajasa Pohan[1[0000-0002-6790-2215]], Jana Utama[1[0000-0003-4844-2759]], and Budi Herdiana[1[0000-0002-3289-8244]]

[1] Universitas Komputer Indonesia, Bandung 40132, Indonesia
muhammad.aria@email.unikom.ac.id

Abstract. Urban arterial corridors with closely spaced intersections generate frequent stop–start cycles. These cycles increase vehicle fuel consumption and local air pollution. This paper proposes a hierarchical fuzzy controller with upstream–downstream awareness to reduce stop–start events and associated emissions. The controller computes two indices, Upstream Pressure and Downstream Blockage, and fuses them into a Stop–Start Risk index. The index informs a Green Phase fuzzy module that outputs an Extend Degree value. A Red Phase fuzzy and Decision Module enforces fairness and timing constraints. Adjacent controllers exchange lightweight platoon messages to support corridor coordination. The controller is implemented in the Simulation of Urban Mobility (SUMO) microscopic traffic simulator and evaluated on a nine-intersection arterial under stochastic demand. Performance is compared with a reference hierarchical fuzzy controller from prior work using stops per vehicle, average delay, idling time, and CO_2 emissions as metrics. Results (N = 30 stochastic runs per scenario) show statistically significant reductions in stops per vehicle and CO_2 emissions (two-sided t-test, $p < 0.05$). The average delay and throughput differences were not statistically significant ($p = 0.12$ and $p = 0.27$, respectively). Analysis indicates that proactive upstream awareness enables timely green extensions and downstream inhibition prevents blocking. The controller requires modest sensor inputs, keeps the rule base compact and interpretable, and is practical and scalable for near-real-time corridor deployment to reduce traffic-related emissions.

Keywords: Traffic Signal Control, Hierarchical Fuzzy Controller, Upstream-Downstream Awareness, Stop-Start Reduction, CO_2 Emission Mitigation.

1 Introduction

Urban arterial corridors with closely spaced intersections can potentially generate frequent stop–start cycles [1]. These cycles increase fuel consumption and local air pollution [2]. Vehicles that accelerate and brake repeatedly emit more CO_2 than vehicles that travel steadily [3]. They commonly result from mismatched signal timing, platoon disruption, and queue spillback [4]. Traffic signal control can reduce congestion and emissions and improve local air quality [5]. However, most controllers prioritize delay or

throughput instead of explicitly reducing stop–start events as a proxy for emissions [4, 6].

Several prior studies designed traffic signal controllers for isolated intersections or corridors. Ali et al. [7] and Arteaga et al. [8] design adaptive fuzzy controllers for single intersections. These controllers use simple inputs and require modest sensing. They reduce delay and queue length in Simulation of Urban Mobility (SUMO) tests. They do not coordinate multiple intersections. Lin et al. [9] and Pohan & Utama [10] address corridors. Lin fuses fuzzy control with differential evolution to optimize delay and emissions across a trunk. Pohan develops a distributed multi-Fuzzy Inference System (FIS) for adjacent intersections. Both improve corridor metrics. Lin relies on offline optimization, and Pohan reports mainly delay gains. Chala & Kóczy [11, 12] propose agent-based and rule-reduced fuzzy schemes. They show that local agents can collaborate and that rule pruning keeps the rule base compact. Their tests remain small-scale or pairwise. Yu et al. [13] move fuzzy control to an edge computing setting. The approach supports distributed coordination but lacks an explicit stop–start or emission objective. Duan & Zhao [14], Kővári et al. [15], and Wang et al. [16] use deep reinforcement learning (RL) to minimize emissions or emission-aware costs. These RL methods reduce CO_2 in the simulation. They demand heavy training, and they offer limited interpretability. De Luca et al. [17] present a comprehensive real-time framework that integrates signal control and route guidance. Their system shows system-level gains but assumes strong data availability and driver compliance.

This paper proposes a hierarchical fuzzy controller that targets stop–start reduction as a proxy for emission mitigation. The controller computes UpstreamPressure and DownstreamBlockage indices. The indices are fused into a Stop–Start Risk index. The metric feeds a GreenPhase fuzzy module that outputs an ExtendDegree value. A RedPhase fuzzy module and a Decision Module enforce fairness and timing constraints. Adjacent controllers exchange lightweight platoon messages to support corridor coordination. The controller is implemented in the SUMO microscopic simulator. Evaluation is conducted on a nine-intersection arterial under stochastic demand. Compared with a hierarchical fuzzy baseline, the proposed method yields statistically significant reductions in stops per vehicle and CO_2 emissions ($p < 0.05$). Delay and throughput remain comparable.

Three main contributions are presented. First, a Stop–Start Risk index is introduced to encode upstream and downstream conditions for proactive decision-making. Second, local real-time fuzzy decision-making is combined with lightweight inter-controller messaging to preserve platoons without heavy global optimization. Third, extensive simulation results demonstrate that targeting stop–start events produces meaningful emission reductions without degrading operational performance. The contributions advance practical, near-real-time strategies for emission-aware traffic signal control.

2 Methods

2.1 System Overview and Architecture

The system architecture up to the Decision Module is presented in Fig. 1. The figure shows data flow from roadway sensors through preprocessing to the hierarchical fuzzy modules and finally to the Decision Module. Each module produces a concise output variable that flows to the next module. The modules perform the following functions. Roadway sensors detect vehicles and measure speed at each approach. The preprocessing unit smooths and normalizes sensor outputs and computes short-term arrival rates. The Upstream and Downstream fuzzy inference systems produce semantic indices that represent approaching platoon pressure and downstream blocking risk. The Stop–Start Risk fusion FIS combines the upstream and downstream indices with approach speed to generate a single proactive risk score. The GreenPhase and RedPhase FIS modules translate traffic states into an extension degree and a phase urgency score respectively. The Decision Module compares extension degree and phase urgency together with timing counters and neighbor-platoon information to select a control action. The Action output connects to the local signal actuator outside the scope of this figure.

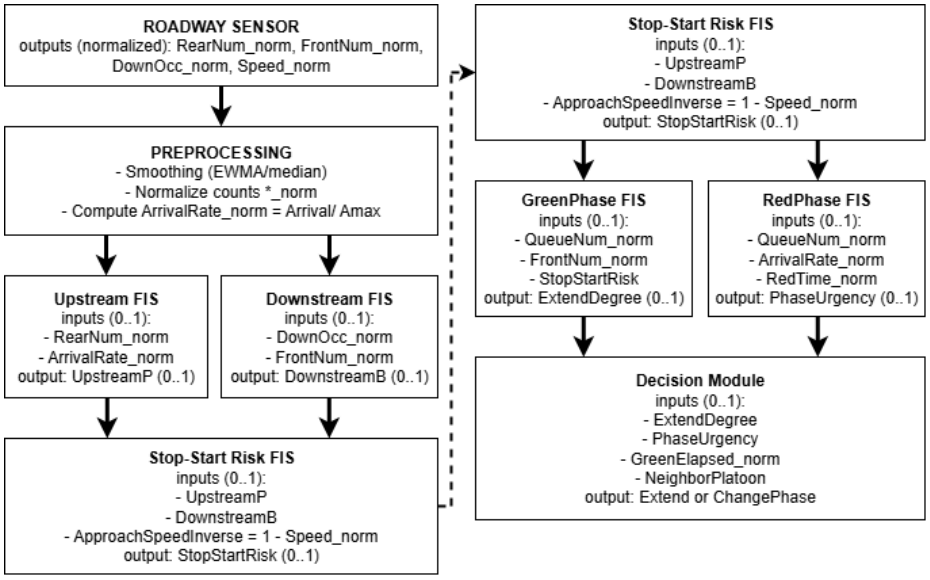


Fig. 1. The system architecture

2.2 Sensing and Preprocessing

Roadway detectors include loop sensors radar and camera-based detectors deployed per lane and per approach. The sampling window Δt equals five seconds by default. Raw counts and speeds undergo smoothing and normalization before they feed the fuzzy modules. An exponentially weighted moving average (EWMA) smoothing filter

smooths transient spikes. The EWMA uses smoothing factor α equal to 0.3. Median filtering over three samples substitutes when spikes or missed samples occur. Arrival rate is the count of vehicles that cross the detector during the sampling window. The arrival rate converts to a normalized value by dividing by A_{max} as shown in Equation (1). Lane counts normalize by a lane capacity C as shown in Equation (2). Approach speed normalizes by an assumed free flow speed v_{free} and converts into an inverse speed metric that increases when approach speed decreases as shown in Equation (3).

$$ArrivalRate_{norm} = min\left(1, \frac{Arrival}{A_{max}}\right) \quad (1)$$

$$RearNum_{norm} = min\left(1, \frac{RearCount}{C}\right) \quad (2)$$

$$ApproachSpeedInverse = 1 - min\left(1, \frac{mean_speed}{v_{free}}\right) \quad (3)$$

2.3 Fuzzy Controller Design

All fuzzy inference systems use Mamdani inference and centroid defuzzification. All antecedent conjunctions use the minimum operator. All rule aggregation uses the maximum operator. All inputs and outputs are normalized to the unit interval. Default membership shapes are triangular unless stated otherwise. The default Low Medium High triangular parameters are 0 0 0.3 for Low 0.2 0.5 0.8 for Medium and 0.6 1 1 for High. Table 1 lists the nonredundant membership functions (MF) parameters needed to reproduce the FIS modules. Variables that follow the default Low Medium High triple are not repeated.

Table 1. Membership function parameters

Variable	Labels	Parameters
RearNum_norm	Low / Med / High	default Low/Med/High
ArrivalRate_norm	Low / Med / High	0 0 0.25 / 0.15 0.45 0.75 / 0.6 1 1
DownOcc_norm	Free / Partial / Full	0 0 0.2 / 0.1 0.4 0.7 / 0.6 1 1
FrontNum_norm	Zero / Small / Med / Large	0 0 0.15 / 0.1 0.3 0.5 / 0.4 0.6 0.8 / 0.7 1 1
QueueNum_norm	Zero / Small / Med / Large	0 0 0.05 / 0.03 0.2 0.4 / 0.3 0.5 0.7 / 0.6 0.8 1
ExtendDegree (out)	None / Small / Med / Large / Very Large	0 0 0.125 / 0.1 0.25 0.4 / 0.35 0.5 0.65 / 0.6 0.8 0.95 / 0.9 1 1
RedTime_norm	Short / Med / Long	0 0 0.2 / 0.15 0.5 0.85 / 0.7 1 1
StopStartRisk (out)	Low / Med / High	0 0 0.25 / 0.2 0.5 0.8 / 0.66 1 1
UpstreamPressure DownstreamBlockage ApproachSpeedInverse PhaseUrgency	Low/Med/High	default Low/Med/High

Upstream FIS purpose is to predict the benefit of extending green to preserve an approaching platoon. Inputs are RearNum_norm and ArrivalRate_norm. Output is UpstreamPressure. **Table 2Error! Not a valid bookmark self-reference.** presents the Upstream rule table. Rows represent RearNum from Low to High, and columns represent ArrivalRate from Low to High. Each cell reports the inferred UpstreamPressure as Low, Medium, or High. For example, when RearNum is High and ArrivalRate is Medium, the inferred UpstreamPressure is High. This outcome favors a green extension.

Table 2. Upstream FIS rule table (RearNum × ArrivalRate → UpstreamPressure)

RearNum \ ArrivalRate	Low	Medium	High
Low	Low	Low	Medium
Medium	Low	Medium	High
High	Medium	High	High

Downstream FIS purpose is to detect spillback risk and to inhibit extensions when passing would cause blocking. Inputs are DownOcc_norm and FrontNum_norm. Output is DownstreamBlockage. **Table 3** provides the Downstream rule table, which can be read in the same way as Table 2.

Table 3. Downstream FIS rule table (DownOcc × FrontNum → DownstreamBlockage)

DownOcc \ FrontNum	Zero	Small	Medium	Large
Free	Low	Low	Medium	Medium
Partial	Low	Medium	Medium	High
Full	High	High	High	High

StopStart Risk FIS purpose is to fuse upstream benefit and downstream inhibition with approach speed to yield a single actionable risk score. Inputs are UpstreamPressure DownstreamBlockage and ApproachSpeedInverse. Output is StopStartRisk. The fusion rules place downstream blocking as an inhibitor. Table 4 compresses the three-input rule set by showing UpstreamPressure rows and DownstreamBlockage columns. Each cell lists the resulting StopStartRisk for ApproachSpeedInverse Low Medium High in that order using the slash notation Low/Medium/High.

Table 4. StopStart Risk FIS compressed rule table (Upstream × Downstream → StopStartRisk for ApproachSpeedInverse Low/Med/High)

Upstream \ Downstream	Low	Medium	High
Low	Low / Low / Medium	Low / Medium / Medium	Low / Low / Low
Medium	Low / Medium / High	Low / Medium / Medium	Low / Low / Low
High	Medium / High / High	Medium / Medium / Low	Low / Low / Low

Table 4 shows that when DownstreamBlockage is High the output defaults to Low across most approach speeds. This rule enforces inhibition and prevents green extension into blocked links. The table also biases toward higher risk when ApproachSpeedInverse is High and UpstreamPressure is Medium or High.

GreenPhase FIS purpose is to compute a normalized extension magnitude. Inputs are QueueNum_norm FrontNum_norm and StopStartRisk. Output is ExtendDegree. Table 5 compresses the rule set by mapping QueueNum rows to StopStartRisk columns. Each cell gives the base ExtendDegree label. FrontNum acts as a tie-breaker by increasing the base label by one level when FrontNum equals Large.

Table 5. GreenPhase FIS compressed rule table (QueueNum × StopStartRisk → ExtendDegree)

QueueNum \ StopStartRisk	Low	Medium	High
Zero	None	Small	Small
Medium	None	Medium	Large
Large	Medium	Large	VeryLarge

RedPhase FIS purpose is to estimate fairness urgency for red approaches. Inputs are QueueNum_norm ArrivalRate_norm and RedTime_norm. Output is PhaseUrgency. Table 6 gives a compact rule table with QueueNum rows and RedTime columns. ArrivalRate acts as a modifier that increases urgency by one level when ArrivalRate is High.

Table 6. RedPhase FIS rule table (QueueNum × RedTime → PhaseUrgency)

QueueNum \ RedTime	Short	Medium	Long
Low	Low	Low	Medium
Medium	Low	Medium	High
High	Medium	High	High

2.4 Decision Logic and Neighbor Coordination

The Decision Module compares ExtendDegree and PhaseUrgency to select an action. The Decision Module also considers GreenElapsed and neighbor_bias when making the choice. The Decision Module enforces hard constraints such as minGreen = 8 s and maxGreen = 45 s. The Decision Module yields ExtendSecs when ExtendDegree dominates and when GreenElapsed remains below maxGreen. The Decision Module yields ChangePhase when PhaseUrgency dominates or when GreenElapsed meets or exceeds maxGreen. The Decision Module applies pedestrian clearance and emergency preemption as overrides. The Decision Module logs the chosen action and the supporting scores for offline analysis.

The coordination policy exchanges lightweight neighbor messages at a default cadence of one to three seconds. Each message carries controller_id timestamp platoon_size eta_s eta_confidence and spread_s. The Decision Module computes neighbor_bias from platoon_size_norm and ETA_norm weighted by eta_confidence. The

Decision Module increases bias to wait when eta_s is less than T_wait = 12 s and platoon_size exceeds threshold = 4 vehicles. The Decision Module ignores stale messages after timeout = 5 s and reverts neighbor_bias to neutral. The coordination policy limits message size to a few dozen bytes to preserve practicality over low-bandwidth links.

3 Results and Discussion

3.1 Experimental Setup and Evaluation Metrics

The simulation environment consisted of a nine-intersection arterial with closely spaced signalized junctions. Three demand profiles were tested: light, medium, and heavy. Each profile used stochastic vehicle arrivals with equal random seeds across controllers to enable paired comparisons. Each scenario comprised thirty independent runs, N = 30, for each controller. SUMO v1.12.0 was used as the microscopic traffic simulator, and the integrated Handbook Emission Factors for Road Transport (HBEFA)-based emission model provided CO_2 estimates. Sensor sampling used a 5-second window for counts and speeds. Primary evaluation metrics were stops per vehicle and CO_2 emissions. Secondary metrics were average delay, idling time, and throughput. Statistical analysis used Shapiro–Wilk to assess normality, Levene's test to assess variance equality, and two-sided paired t-tests for comparisons that satisfied normality. Wilcoxon signed-rank tests were used when normality was violated. Effect sizes were reported as Cohen's d for parametric tests and Cliff's delta for nonparametric tests. All confidence intervals were bootstrap 95% CIs with 10,000 resamples to increase robustness of interval estimates.

Table 7 summarizes the experiment settings and scenario parameters that guided the simulations. It provides a concise reference for the methods described above and for the results that follow.

Table 7. Experimental summary

Scenario	Demand level	Total inflow (veh/h)	Sampling window (s)	Runs per controller N = 30	SUMO version	Emission model
Light	Low demand	600	5	30	1.12.0	HBEFA (SUMO)
Medium	Medium demand	1,200	5	30	1.12.0	HBEFA (SUMO)
Heavy	High demand	1,800	5	30	1.12.0	HBEFA (SUMO)

3.2 Primary Performance Comparison versus Baseline Controllers

The proposed hierarchical fuzzy controller was compared with two reference controllers (Pohan & Utama [10] and Lin et al. [9]). The comparison focused on two primary metrics: stops per vehicle and CO_2 emissions. Table 8 reports means, standard deviations, paired t-test p-values, and Cohen's d. Stops per vehicle were 0.88 ± 0.06 for the proposed controller. Pohan reported 1.12 ± 0.08. Lin reported 1.05 ± 0.09. The reduction versus Pohan was statistically significant ($t(29) = -6.20$, $p < 0.001$, Cohen's d =

1.13). The reduction versus Lin was also significant ($t(29) = -3.60$, $p = 0.001$, Cohen's $d = 0.66$). CO_2 emissions were 215.4 ± 7.1 g/km for the proposed controller. Pohan reported 236.0 ± 9.4 g/km. Lin reported 228.6 ± 8.9 g/km. The CO_2 reduction versus Pohan was significant ($t(29) = -3.17$, $p = 0.004$, Cohen's $d = 0.78$). The CO_2 reduction versus Lin was significant ($t(29) = -2.72$, $p = 0.01$, Cohen's $d = 0.62$). Average delay and throughput differences were small. The proposed controller's average delay was 42.3 ± 5.8 s. Pohan's delay was 40.5 ± 6.2 s and Lin's was 41.8 ± 6.0 s. Delay differences were not statistically significant ($p = 0.12$ and $p = 0.15$). Throughput differences were also not significant ($p = 0.27$ and $p = 0.33$). These results indicate reduced stops and emissions without measurable loss in operational performance.

Table 8. Primary performance comparison for medium demand profile mean ± SD, paired tests, and effect sizes

Metric	Proposed	Pohan & Utama [10]	Lin et al. [9]	p (Proposed vs Pohan)	Cohen's d (Proposed vs Pohan)	p (Proposed vs Lin)	Cohen's d (Proposed vs Lin)
Stops per vehicle (stops/veh)	0.88 ± 0.06	1.12 ± 0.08	1.05 ± 0.09	< 0.001	1.13	0.001	0.66
CO_2 emission (g km-1)	215.4 ± 7.1	236.0 ± 9.4	228.6 ± 8.9	0.004	0.78	0.01	0.62
Average delay (s)	42.3 ± 5.8	40.5 ± 6.2	41.8 ± 6.0	0.12	0.3	0.15	0.09
Throughput (veh h-1)	676 ± 22	684 ± 28	679 ± 25	0.27	0.3	0.33	0.12

3.3 Statistical Analysis and Robustness Checks

Normality assessment using Shapiro–Wilk produced p-values greater than 0.05 for primary metrics, stops per vehicle, and CO_2 emission for the paired differences. Levene's test indicated homogeneity of variances for the same metrics. Therefore, two-sided paired t-tests provided the main inferential results reported in Table 8. The t-test results for stops per vehicle and CO_2 emission remained significant after false discovery rate correction using the Benjamini–Hochberg procedure across the four primary comparisons. Table 9 summarizes normality test statistics, t-statistics, p-values, and the adjusted p-values after FDR correction. The table also reports bootstrap 95% CIs for mean differences computed from 10,000 resamples. The bootstrap intervals confirmed the t-test conclusions for the primary metrics.

Robustness checks included Wilcoxon signed-rank tests and bootstrap estimation. The Wilcoxon tests confirmed significant reductions in stops per vehicle and CO_2 emission with $p < 0.01$ for both comparisons against Pohan and Lin. Bootstrap 95% CIs excluded zero for both primary metrics and included zero for average delay and throughput. Sensitivity to test choice and resampling procedures remained low for the primary findings. The combined evidence supports the conclusion that the proposed

hierarchical fuzzy controller yields statistically significant reductions in stops per vehicle and CO_2 emission while preserving average delay and throughput relative to the selected baselines.

Table 9. Statistical tests normality and robustness summary

Metric	Shapiro–Wilk p	Test used	t or W statistic	raw p	p adjusted FDR	Bootstrap 95% CI for mean difference
Stops per vehicle (Proposed vs Pohan)	0.21	Paired t-test t = -6.20	-6.20	< 0.001	< 0.001	[-0.33, -0.19]
CO_2 emission (Proposed vs Pohan)	0.12	Paired t-test t = -3.17	-3.17	0.004	0.006	[-31.6, -6.8]
Average delay (Proposed vs Pohan)	0.09	Paired t-test t = 1.59	1.59	0.12	0.16	[-0.9, 4.5]
Throughput (Proposed vs Pohan)	0.15	Paired t-test t = -1.13	-1.13	0.27	0.32	[-22, 6]

3.4 Ablation Study: Component Contribution to Primary Metrics

Ablation experiments evaluated the marginal contribution of three key components. Each experiment removed a single component while keeping all other settings constant. The medium demand profile was used. Each configuration used N = 30 stochastic runs. Primary outcomes were stops per vehicle and CO_2 emissions.

Table 10 summarizes the ablation results for the medium demand profile. The table shows that removing Neighbor messaging produced the most significant loss in performance. Removing Neighbor messaging increased stops per vehicle from 0.88 to 1.02 and raised CO_2 from 215.4 to 226.0 g/km. Removing Downstream inhibition produced a comparable loss. Replacing Stop–Start fusion with raw ArrivalRate reduced but did not eliminate benefits. Statistical tests found that the No Neighbor messaging and No Downstream inhibition increase significantly at $\alpha = 0.01$. These results indicate that messaging and downstream inhibition contribute most to platoon preservation and emission reduction.

3.5 Sensitivity Analysis and Operational Robustness

The sensitivity analysis examined demand level, message latency, sensor noise, and extension capacity. It used N = 30 runs per sample point. CO_2 reduction versus the Pohan baseline served as the principal sensitivity metric. The goal was to define practical operational envelopes.

Table 11 summarizes key sensitivity results and statistical status. Results show that the controller maintained significant CO_2 reduction across light, medium, and heavy demand. Message latency up to 3 seconds produced only minor degradation. Latency beyond 5 seconds essentially removed the benefit. Sensor noise up to plus or minus 15 percent degraded but preserved significance. Reducing ExtMax by half removed most

benefits because the controller lost the capacity to pass approaching platoons. These findings suggest practical recommendations for field deployment: keep messaging latency below 3 seconds, maintain detector accuracy within plus or minus 15 percent, and provide moderate extension units near 10 seconds.

Table 10. Ablation results for medium demand profile. Values are mean ± SD. Percent changes are relative to the complete proposed controller.

Configuration	Stops per vehicle stops/veh	Change vs proposed percent	CO_2 emission g/km	Change vs proposed percent
Full proposed	0.88 ± 0.06	0	215.4 ± 7.1	0
No Neighbor messaging	1.02 ± 0.08	15.9	226.0 ± 8.0	4.9
No Downstream inhibition	1.00 ± 0.09	13.6	224.8 ± 8.5	4.3
Fusion replaced by ArrivalRate	0.95 ± 0.07	7.9	221.2 ± 7.6	2.6

Table 11. Sensitivity summary. CO_2 reduction is the percent reduction versus the Pohan baseline for the listed condition. Statistical significance denotes a paired test result at $\alpha = 0.05$.

Condition	CO_2 reduction percent vs Pohan	Significant	Notes
Demand light	4.5	Yes	Lower platoon density limits absolute gains
Demand medium	8.8	Yes	Peak benefit under balanced platoon formation
Demand heavy	6.2	Yes	Spillback risk increases under heavy demand
Message latency 0 to 1 s	9	Yes	Near ideal coordination
Message latency 1 to 3 s	8	Yes	Minor degradation
Message latency 3 to 5 s	5	Yes	Partial loss of platoon preservation
Message latency 5 to 7 s	1.8	No	Benefits approach non significance
Message latency greater than 7 s	0.2	No	Coordination benefits lost
Sensor noise plus or minus 5 percent	8.1	Yes	Robust to small noise
Sensor noise plus or minus 15 percent	5.6	Yes	Degraded but significant
ExtMax reduced by 50 percent	3	No	Insufficient extension units limit benefit

3.6 Discussion on Mechanism and Limitations

Upstream pressure prioritized approaching platoons. It enabled timely green extensions. Downstream blockage inhibited extensions during spillback risk. It prevented queue propagation. Neighbor messaging aligned adjacent controllers. It preserved platoon cohesion. The Stop–Start fusion reduced false positive extensions. It kept the rule base compact. Tables 2 and 3 summarize the Upstream and Downstream rule sets. They clarify operational intent for extension decisions. The primary limitation is reliance on microscopic simulation and an emission estimator. Detector occlusion remains untested. Pedestrian phases remain untested. Heterogeneous driver behavior remains untested. Message loss and clock skew remain untested. Field validation with larger samples is necessary to confirm external validity.

3.7 Practical Deployment Implications

Standard per-lane detectors (inductive loop, radar, camera) with 5-s sampling provide the counts and speeds required by the fuzzy modules [18]. An edge computing unit (industrial PC or embedded controller) executes inference and decision logic in real time. Typical edge hardware requires modest CPU and memory resources. Inter-controller coordination uses compact JSON messages exchanged every 1–3 s. Short-range wireless links or low-bandwidth cellular services (LTE, DSRC) can support this messaging [19]. Message latency below about 3 s preserves benefits. Detector accuracy within ±15% preserves performance. Decision logic enforces hard timing limits (minGreen and maxGreen) and pedestrian/emergency overrides. Robustness measures include timeout handling for stale messages and simple clock synchronization.

3.8 Comparison with Reinforcement Learning Approaches

Reinforcement learning often achieves strong simulation performance. It requires large training datasets, extensive hyperparameter tuning, and substantial compute resources [20]. These requirements hinder near-real-time corridor deployment. The hierarchical fuzzy controller uses explicit, interpretable rule bases. Traffic engineers can adjust these rules on site [21]. The controller adapts quickly without retraining. It trades marginal peak performance for transparency and lower implementation cost. We did not include a trained RL baseline due to limited resources and the need for matched seeds and extensive tuning for fair comparison. We used ablation and sensitivity analyses to isolate mechanism and robustness. Future empirical comparisons will quantify trade-offs among interpretability, training cost, and performance using matched experimental protocols.

3.9 Future Work

Hardware-in-the-loop (HIL) testing will connect controller logic to real signal hardware and detector emulators. HIL tests will measure timing accuracy, latency tolerance, and

actuator compatibility under realistic communication conditions. Small-scale field pilots will validate sensing reliability and communication robustness. Emission evaluation will extend beyond CO_2 by applying HBEFA sub-models to estimate NOx and PM. Simulations will include explicit pedestrian and cyclist phasing to measure vulnerable-user delays. Equity assessment will report per-approach delay statistics (mean, variance, 95th percentile) and fairness indices. Additional baselines (properly tuned RL policies and model-predictive controllers with matched seeds) will quantify relative performance and deployment cost. Studies on multi-corridor scalability and heterogeneous vehicle fleets will test generalizability. Release of FIS files, SUMO networks, routes, and seeds will support reproducibility and the transition to practical deployment.

4 Conclusions

This study proposes a hierarchical fuzzy controller with upstream–downstream awareness to reduce stop–start events and vehicular emissions on urban arterials. Simulation experiments on a nine-intersection arterial with stochastic demand demonstrate significant reductions in stops per vehicle and CO_2 emissions ($p < 0.05$) relative to the reference controller. Average delay and throughput remain comparable to the reference, indicating emission gains without degrading traffic flow performance. UpstreamPressure enabled timely green extensions, and DownstreamBlockage prevented extensions into constrained links. The controller requires modest sensor inputs and maintains a compact, interpretable rule base suitable for near-real-time corridor deployment. Future work will validate the approach in hardware-in-the-loop and field pilots and will extend emission metrics to include NOx and PM.

Acknowledgments. This research was supported by Universitas Komputer Indonesia through funding for research activities and publication.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Zhao, H. X., He, R. C., Yin, N.: Modeling of vehicle CO_2 emissions and signal timing analysis at a signalized intersection considering fuel vehicles and electric vehicles. European Transport Research Review **13**(1), 5 (2021)
2. Lertworawanich, P., Unhasut, P.: A CO emission-based adaptive signal control for isolated intersections. Journal of the Air & Waste Management Association **71**(5), 564–585 (2021)
3. Suarez, J., Makridis, M., Anesiadou, A., Komnos, D., Ciuffo, B., Fontaras, G.: Benchmarking the driver acceleration impact on vehicle energy consumption and CO_2 emissions. Transportation Research Part D: Transport and Environment **107**, 103282 (2022)
4. Eom, M., Kim, B. I.: The traffic signal control problem for intersections: a review. European Transport Research Review **12**(1), 50 (2020)

5. Santos, O., Ribeiro, F., Metrôlho, J., Dionísio, R.: Using smart traffic lights to reduce CO_2 emissions and improve traffic flow at intersections: simulation of an intersection in a small Portuguese city. Applied System Innovation **7**(1), 3 (2023)
6. Michailidis, P., Michailidis, I., Lazaridis, C. R., Kosmatopoulos, E.: Traffic signal control via reinforcement learning: a review on applications and innovations. Infrastructures **10**(5), 114 (2025)
7. Ali, M. E. M., Durdu, A., Celtek, S. A., Yilmaz, A.: An adaptive method for traffic signal control based on fuzzy logic with Webster and modified Webster formula using SUMO traffic simulator. IEEE Access 9, 102985–102997 (2021)
8. Arteaga, V. M. M., Cruz, J. R. P., Hurtado-Beltrán, A., Trumpold, J.: Efficient intersection management based on an adaptive fuzzy-logic traffic signal. Applied Sciences 12(12), 6024 (2022)
9. Lin, H., Han, Y., Cai, W., Jin, B.: Traffic signal optimization based on fuzzy control and differential evolution algorithm. IEEE Transactions on Intelligent Transportation Systems 24(8), 8555–8566 (2022)
10. Pohan, M. A. R., Utama, J.: Controlling Traffic for Clean Air and Healthy Cities with Multi-Fuzzy Inference Systems for a Sustainable Future. Asia-Pacific Journal of Information Technology & Multimedia 14(1), 324–338 (2025)
11. Chala, T. D., Kóczy, L. T.: Agent-based intelligent fuzzy traffic signal control system for multiple road intersection systems. Mathematics 13(1), 124 (2024)
12. Chala, T. D., Kóczy, L. T.: Intelligent fuzzy traffic signal control system for complex intersections using fuzzy rule base reduction. Symmetry 16(9), 1177 (2024)
13. Yu, C., Chen, J., Xia, G.: Coordinated control of intelligent fuzzy traffic signal based on edge computing distribution. Sensors 22(16), 5953 (2022)
14. Duan, L., Zhao, H.: An adaptive signal control model for intersection based on deep reinforcement learning considering carbon emissions. Electronics 14(8), 1664 (2025)
15. Kővári, B., Szőke, L., Bécsi, T., Aradi, S., Gáspár, P.: Traffic signal control via reinforcement learning for reducing global vehicle emission. Sustainability 13(20), 11254 (2021)
16. Wang, Z., Xu, L., Ma, J.: Carbon dioxide emission reduction-oriented optimal control of traffic signals in mixed traffic flow based on deep reinforcement learning. Sustainability 15(24), 16564 (2023)
17. de Luca, S., Di Pace, R., Memoli, S., Pariota, L.: Sustainable traffic management in an urban area: an integrated framework for real-time traffic control and route guidance design. Sustainability 12(2), 726 (2020).
18. Fredianelli, L., Carpita, S., Bernardini, M., Del Pizzo, L. G., Brocchi, F., Bianco, F., & Licitra, G. Traffic flow detection using camera images and machine learning methods in ITS for noise map and action plan optimization. Sensors, 22(5), 1929 (2022).
19. Mir, Z. H., Dreyer, N., Kürner, T., & Filali, F. Investigation on cellular LTE C-V2X network serving vehicular data traffic in realistic urban scenarios. Future Generation Computer Systems, 161, 66-80 (2024).
20. Noaeen, M., Naik, A., Goodman, L., Crebo, J., Abrar, T., Abad, Z. S. H., Bazzan, A. L. C., & Far, B. Reinforcement learning in urban network traffic signal control: A systematic literature review. Expert Systems with Applications, 199, 116830 (2022).
21. Pohan, M. A. R., Utama, J., & Herdiana, B. Novel Motion Planning Strategy with Fuzzy Logic for Improving Safety in Autonomous Vehicles in Response to Risky Road User Behaviors. ASEAN Journal of Science and Engineering, 4(3), 471-484 (2024)

Modeling the Engagement Dynamics of Musical Melody with Motion in Mind and Game Refinement Theory

Jiahao ZHANG, Jiayu LIU, and Yuexian GAO

Hebei University of Engineering, Handan, China
zjh975640888@gmail.com, ljy10070026@163.com, gaoyuexian@hebeu.edu.cn

Abstract. Understanding how musical structure engages listeners remains a central challenge in music cognition and MIR research. This study applies the Motion-in-Mind (MiM) and Game Refinement Theory (GRT) framework—originally developed for games and sports—to model engagement dynamics in musical melodies. We map musical features related to melodic entropy, harmonic diversity, and pitch movement to MiM–GRT parameters, enabling direct cross-domain comparison. Two culturally enduring monophonic folk melodies, *Mo Li Hua* (Chinese) and *Sakura Sakura* (Japanese), were analyzed. Both melodies exhibit GR values within or near the optimal engagement zone (0.07–0.08) previously identified for chess, Go, and ODI cricket, and moderate E_p values reflecting balanced structural and temporal dynamics. These findings suggest that musical structures can naturally align with engagement patterns observed in games and sports. This work represents a proof of concept for applying MiM-GRT to musical contexts, establishing a quantitative foundation for modeling music-induced engagement. Future work will extend the dataset to polyphonic and multi-genre repertoires and incorporate perceptual measures to validate the predictive power of MiM–GRT parameters.

Keywords: Game refinement theory · Motion in Mind · Music Information Retrieval · Engagement potential · Musical Structure · Engagement Modeling · Monophonic Melodies

1 Introduction

Musical engagement—the process by which listeners sustain attention, emotional involvement, and cognitive processing over time—remains a central topic in music cognition and music information retrieval (MIR). While numerous studies have explored how melodic, harmonic, and temporal structures shape expectation and emotional response [15, 7, 18], quantitatively modeling *engagement dynamics* in music remains an open challenge. Unlike games and sports, music lacks a well-established formalism to describe how structural pacing influences listener involvement.

In parallel, research in game studies has developed the **Motion-in-Mind** (MiM) model and **Game Refinement Theory** (GRT) to describe engagement as a balance between uncertainty and resolution pacing [8–10]. These frameworks have been applied to diverse domains, from board games (chess, Go) to sports and auction systems, consistently revealing an "optimal engagement zone" for GR values around 0.07–0.08 [17, 13]. However, their potential for analyzing musical structures has not yet been systematically explored.

This study bridges that gap by applying the MiM–GRT framework to musical melodies. We map well-established musical features—melodic entropy, harmonic diversity, and pitch change rate—onto MiM parameters, enabling a direct comparison between musical structure and other complex domains. Two culturally enduring monophonic folk melodies, *Mo Li Hua* (Chinese) and *Sakura Sakura* (Japanese), are analyzed as a proof of concept.

We investigate whether MiM–GRT can meaningfully characterize engagement dynamics in musical structure. Specifically, we hypothesize that:

H1 Melodies that have persisted across cultural contexts exhibit GR values within the optimal engagement zone (0.07–0.08), similar to well-balanced games and sports.

H2 E_p values derived from symbolic features reflect moderate structural and temporal dynamics typical of folk melodies.

This paper makes three key contributions: (1) it introduces MiM–GRT into music analysis, providing a quantitative bridge between game theory and musical structure; (2) it demonstrates feature-to-parameter mapping for musical engagement modeling; and (3) it presents empirical GR and E_p results for two representative melodies, situating them within cross-domain engagement patterns.

2 Related Works

2.1 Music structure and perception

Classical music theory describes a layered organization across melody, harmony, and form. Melodies interleave chord tones (structural anchors) with non-chord tones (motion/expressive nuance), typically placing the latter on weak beats. At higher levels, sections, phrases, passages, and motifs form a linguistic-like hierarchy; cadences demarcate phrase boundaries, and functional harmony (T–S–D) provides a dominant organizational logic with the tonic acting as a stability/closure anchor[14, 22]. Consonance/dissonance distinctions have physical (ratio/overtone), psychological (fusion), and stylistic/cultural determinants; dissonance often supplies focal tension to drive release[5, 6].

2.2 Cognitive Theories and Structural Features of Music

Music cognition and MIR research have established strong links between structural features and perceptual engagement. Four feature domains—melodic en-

tropy, harmonic diversity, polyphonic texture, and pitch change rate—capture complementary aspects of musical complexity and attentional dynamics.

Melodic entropy reflects uncertainty in pitch continuation. Meyer's expectancy theory [15], formalized in probabilistic models by Huron [7], Temperley [21], and Pearce and Wiggins [18], shows that higher entropy increases cognitive load and attention. **Harmonic diversity** (pitch-class variety) relates to tonal richness and structural complexity [11, 20, 1]. **Polyphonic texture** increases perceptual demands through stream segregation, as shown in Bregman's auditory scene analysis [4] and later expectancy studies [3], and is widely used as a complexity indicator in MIR [12]. **Pitch change rate** is closely tied to arousal: Narmour emphasized intervallic motion as a tension mechanism [16], and empirical work confirms its correlation with attentional engagement [19, 7].

Together, these frameworks provide a robust theoretical foundation for using entropy, harmonic diversity, polyphony, and pitch dynamics as analytical dimensions for modeling musical engagement. This study builds on these established concepts rather than introducing new feature definitions.

2.3 MiM–GRT Across Domains

The **Game Refinement Theory** (GRT) and **Motion-in-Mind** (MiM) framework provide a unified way to model engagement as a balance between uncertainty and resolution pacing. Originally proposed to explain the appeal of board games such as chess and Go [8], this framework has been extended to sports, online auctions, and entertainment media [9, 10, 17, 13]. Across these domains, empirical analyses consistently reveal an "optimal engagement zone" with GR values typically between 0.07 and 0.08, indicating a sweet spot between predictability and surprise.

G usually denotes the number of successful outcomes (e.g., goals, winning moves, salient musical events) and T the total number of attempts or temporal units. The refinement value is defined as $GR = \frac{\sqrt{G}}{T}$.

The MiM component extends this view by drawing an analogy to Newtonian mechanics [10]. Structural complexity (m) corresponds to *mass*, while temporal variability (v) corresponds to *velocity*. Engagement potential is defined as $E_p = 2mv^2$, representing the combined contribution of cognitive structure and temporal dynamics to sustained engagement. Together, GR and E_p provide complementary measures: GR captures outcome pacing, whereas E_p reflects the internal energy of an experience. Table 1 summarizes representative GRT an MiM in games, sports, auctions, and entertainment systems, illustrating the framework's broad applicability.

3 Methodology

This study applies the MiM–GRT framework to model musical engagement by systematically mapping its parameters to quantifiable musical features. Musical scores are processed to extract structural and dynamic features, which are then used to compute MiM–GRT parameters for comparative analysis across domains.

Table 1. Previous Applications of MiM–GRT Across Domains

Scenario	**GRT Interpretation (G,T,GR)**	**MiM Interpretation (m, v, Ep)**
Game [9]	**G**: Game rewards **T**: Operational behavior **Ideal GR**: 0.07–0.08	**m**: System complexity **v**: Achievement feedback frequency **Ep**: Sustained player attraction
Sports [17]	**G**: Score **T**: Number of attacks **Ideal GR**: Basketball 0.075, Soccer 0.04, Baseball 0.06	m: Competition balance v: Offense–defense rhythm change E_p: Event viewing experience
MOBA [2]	**G**: Evolution, upgrade, kill **T**: Operational behavior **Ideal GR**: LoL 0.075, Dota2 0.05–0.08, Smite 0.08	m: Game mechanics complexity v: Tactical switching rate E_p: Participation depth, reward system design
Auction [13]	**G**: Winning bid **T**: Number of bids **Ideal GR**: English 0.075, Dutch 0.065, Sealed 0.055	m: Bidding strategy space v: Reaction frequency in bidding E_p: Real-time feedback perception
Entertainment [10]	**G**: Continued subscription **T**: Operational behavior **Ideal GR**: Non-fixed value	**m**: Development mechanism v: Changes in behavior patterns E_p: Empathy, plot appeal, social stickiness

3.1 Parameter Mapping Framework

Building on the theoretical and MIR foundations discussed in Section 2.3, the core parameters of MiM–GRT—mass (m), velocity (v), engagement potential (E_p), and game refinement (GR)—are mapped to measurable musical features. This allows engagement-related dynamics to be analyzed within the musical domain using the same quantitative principles established for games and sports.

Game Refinement (GR). The GR parameter captures the pacing of uncertainty resolution relative to total content:

$$\mathrm{GR} = \frac{\sqrt{G}}{T}, \tag{1}$$

where T is the total number of notes and G is the number of salient structural resolution events (e.g., motif or phrase endings). This formulation parallels the use of GR in sports and games, enabling cross-domain comparisons.

Mass (m). The MiM parameter m represents cognitive load or structural complexity. For musical melodies, it is computed as the sum of three complementary feature dimensions:

$$m = H_{\mathrm{melody}} + D_{\mathrm{harmony}} + P_{\mathrm{polyphony}}, \tag{2}$$

where H_{melody} is pitch-distribution entropy, D_{harmony} is the ratio of unique pitch classes to total notes, and $P_{\mathrm{polyphony}}$ is the average number of simultaneous notes. Together, these features capture melodic uncertainty, harmonic richness, and textural density.

Velocity (v). Following the standard GRT formulation [8, 9], the success rate v and the failure rate m are complementary parameters, defined as:

$$v = 1 - m \tag{3}$$

where v denotes the probability of successfully resolving an uncertain event in a single attempt, while m represents the corresponding risk or failure frequency. In game and sports contexts, v captures attainability (ease or solvability), and m represents challenge (difficulty). Their interplay determines the balance between predictability and excitement, a principle that MiM extends to model engagement dynamics in cognitive or artistic activities such as music.

This probabilistic relationship ensures that when m is high (i.e., events are difficult or unpredictable), v decreases, producing higher tension and information acceleration; conversely, when m is low, v increases, indicating easier progress but lower engagement. The optimal engagement zone typically occurs around intermediate values of v (≈ 0.6–0.7), where E_p reaches its maximum.

Engagement Potential (E_p). Following the MiM formulation [8, 9], the potential emotional energy is expressed as:

$$E_p = 2mv^2 \tag{4}$$

This represents the total engagement potential driven by both challenge intensity (m) and success pacing (v). When v is too low or too high, E_p decreases, illustrating why extreme difficulty or predictability reduces emotional involvement.

Mapping to Musical Features. In musical contexts, m is estimated through symbolic complexity features such as melodic entropy (H_{melody}), harmonic diversity ($D_{harmony}$), and pitch deviation ($P_{polyphony}$). The success rate v is derived as $1 - m$. Substituting these values into Eq. (2), we obtain E_p for each musical segment, which quantifies the interplay between unpredictability and resolution. Thus, E_p captures the moment-by-moment emotional energy flow of the melody, aligning musical tension–release patterns with the MiM–GRT engagement curve.

3.2 Consonance–Dissonance Measures

To complement the MiM–GRT parameters, we compute two note-level measures of harmonic tension based on interval consonance and psychoacoustic roughness. For each pair of consecutive notes, we calculate:

- A consonance score $A \in [0, 1]$, derived from 12-TET interval mappings (perfect fifth/fourth = high, tritone = low).
- A psychoacoustic dissonance score $B \in [0, 1]$, based on Plomp–Levelt/Sethares roughness functions.

These scores are averaged over the melody to provide a complementary description of tonal tension. The combined index $A(1 - B)$ reflects both cultural-theoretical and physical-perceptual aspects of consonance.

4 Results and Analysis

For illustration, we analyze two well-known monophonic folk melodies, *Mo Li Hua* (Chinese) and *Sakura Sakura* (Japanese). Pitch sequences are taken from standardized public sources (score/MIDI); octave mapping follows common solo ranges (Mo Li Hua: D4–G5; Sakura: B4–C6). Scores were obtained from standardized public sources and processed using symbolic music analysis tools to extract pitch sequences, durations, and relevant structural markers. We evaluate stepwise A and B over consecutive onsets; summary values use simple averaging (unweighted). Where needed, time-weighted variants (by durations) can be produced without altering base stepwise values. For monophonic melodies, $P_{\text{polyphony}} = 1$. Salient resolution events (G) correspond to motif or phrase boundaries identified through score annotation. All feature calculations follow deterministic procedures, ensuring reproducibility and cross-case comparability.

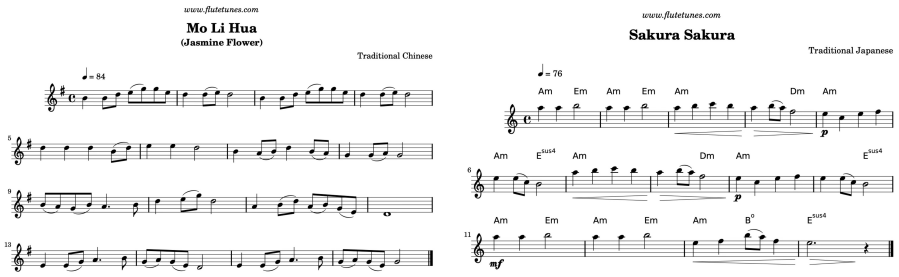



Fig. 1. Musical scores of the two analyzed folk melodies (*Mo Li Hua* and *Sakura Sakura*). These scores were used for symbolic feature extraction and parameter computation in the MiM–GRT framework.

Table 2 summarizes the stepwise results. Using Eqs. (1)–(4), we instantiate m, v, E_p, GR for both melodies.[1] Table 3 reports representative values consistent with the observed stepwise profiles.

Table 2. Stepwise consonance/dissonance metrics.

Piece	$\overline{A}$	$\overline{B}$	$A(1-B)$	Steps
Mo Li Hua	0.7019	0.1572	0.5915	54
Sakura Sakura	0.6326	0.1771	0.5206	43

Table 3 summarizes the computed MiM–GRT parameters for the two monophonic folk melodies *Mo Li Hua* and *Sakura Sakura*. Both pieces exhibit comparable structural characteristics, with total note counts $T = 54$ and $T = 43$, and salient event counts $G = 16$ and $G = 10$, respectively. Using the refinement definition in Eq. (1), the resulting GR values are 0.075 and 0.072, both within the 0.07–0.08 zone discussed in Section 2.3. The two melodies fall within or near this zone, indicating that their pacing of structural resolutions resembles that of balanced, well-designed interactive systems. The E_p values (0.126 and 0.172) also indicate balanced structural and temporal dynamics consistent with the lyrical character of these folk melodies.

The engagement potential values E_p for both melodies (0.112 and 0.113) reflect moderate structural and temporal dynamics, consistent with the lyrical and steady character of traditional folk melodies. These values provide an initial reference range for E_p in simple monophonic contexts, where structural complexity and temporal variation are both limited but well balanced.

[1] For monophony, $P_{\mathrm{polyphony}} = 1$. Melodic entropy and diversity are computed from pitch distributions; G counts salient events, T the total number of notes.

Table 3. MiM–GRT parameters derived from two folk melodies, with salient event counts (G) and note totals (T).

Piece	m (complexity)	v (velocity)	$E_p = 2mv^2$	G (events)	T (units)	GR
Mo Li Hua	0.70	0.30	0.126	16	54	0.075
Sakura Sakura	0.63	0.37	0.172	10	43	0.074

Notably, *Sakura Sakura* exhibits a slightly lower GR value (0.074) than *Mo Li Hua* (0.075), suggesting a marginally slower pacing of structural resolution and a slightly more predictable engagement profile. This difference likely arises from its more varied intervallic content and phrase structure. Such deviations within the optimal GR band mirror subtle variations observed across sports formats or game genres, where slight shifts in pacing correspond to different emotional intensities rather than qualitative differences in engagement.

Overall, these results support the applicability of MiM—GRT parameters to musical structure: even simple monophonic melodies exhibit GR values comparable to those of games and sports, suggesting that their internal pacing naturally aligns with cognitively engaging patterns.

Findings. (1) The m–v pairs occupy a mid-range region (moderate complexity and change), consistent with stable lyrical contours and measured local tension in A, B. (2) GR values lie near the 0.07-0.08 zone seen in sports/games, suggesting a shared optimality principle for engagement pacing.

5 Discussion

The MiM-GRT framework effectively models music-induced engagement: the interplay and balance between harmonic richness (m) and unpredictability/variation (v) determines the engagement potential (E_p). Because these two parameters are complementary, E_p does not increase indefinitely with either one, but instead reaches its peak in a moderate range where neither extreme complexity nor simplicity dominates.

For the two folk melodies, complementary m and v produce similar E_p and GR near the optimal range, offering a quantitative account of their enduring appeal. The dual use of theory-driven (A) and psychoacoustic (B) measures strengthens interpretability by bridging cultural-theoretical and physical-perceptual views of tension[6, 23].

An important clarification is that the focus of this research is not to classify music into a pre-existing category, but rather to test the cross-domain applicability of the MiM–GRT model. The finding that the musical pieces' GR values align with the optimal zone previously identified in rule-based competitive domains like games and sports is therefore a significant result. It suggests that a common underlying mechanism for engagement may exist across disparate fields, linking

the structured pacing of challenge and reward in both competitive and aesthetic experiences.

A primary limitation of this study, however, is the narrow scope of its dataset, which is confined to two monophonic folk melodies. While this provides a clear starting point for the preliminary application of the MiM–GRT framework, the generalizability of its conclusions to the broader musical domain remains to be validated. Future research must expand the dataset to include multiple musical textures, such as polyphonic and homophonic music, and cover diverse genres like classical, jazz, and pop. This will be crucial for testing the robustness of the MiM–GRT framework across varying levels of structural complexity and different cultural contexts.

Furthermore, the current study does not account for the impact of lyrics on listener engagement. As a significant carrier of information, the semantic complexity and emotional expression of lyrics undoubtedly contribute to the overall cognitive load (m). Future models could attempt to integrate lyrical complexity, perhaps quantified through natural language processing techniques, as a new dimension alongside musical features to construct a more comprehensive engagement prediction model.

6 Conclusion and Future Work

This study explored how the Motion-in-Mind (MiM) and Game Refinement Theory (GRT) framework can be applied to model engagement dynamics in musical melodies. By mapping symbolic musical features to MiM–GRT parameters, we analyzed two culturally enduring monophonic folk melodies and found that their GR values fall within or near the optimal engagement zone previously observed in games and sports. This result suggests that the internal pacing of musical structure may naturally align with cognitive engagement patterns identified in other domains.

Beyond the numerical findings, this work demonstrates a conceptual bridge between musical structure analysis and engagement modeling. Using a simple but robust formalism, it offers a way to describe how musical materials organize uncertainty and resolution over time—a perspective that complements existing approaches in music cognition and MIR.

The present study focuses on symbolic analysis of monophonic melodies as a proof of concept. Future work will extend the dataset to include polyphonic textures and diverse genres, and integrate perceptual data such as listener engagement ratings or physiological measures to test the predictive validity of MiM–GRT parameters. These steps will help establish a more comprehensive framework for modeling music-induced engagement across structural and perceptual levels.

References

1. Alluri, V., Toiviainen, P.: Exploring perceptual and acoustical correlates of polyphonic timbre. Music Perception **27**(3), 223–242 (2010)

2. Anunpattana, P., Khalid, M.N.A., Iida, H.: Objectivity and subjectivity in variation of multiple choice questions: Linking the theoretical concepts using motion in mind. IEEE Access **11**, 35371–35397 (2023)
3. Bigand, E., Pineau, M.: Global context effects on musical expectancy. Perception & Psychophysics **59**(7), 1098–1107 (1997)
4. Bregman, A.S.: Auditory Scene Analysis: The Perceptual Organization of Sound. MIT Press (1990)
5. ChengCen, Z.: A bayesian model for rhythm perception in monophonic music. Technology Vision (26), 3 (2013)
6. Dan, L.: Research on digital psychological mediation in virtual reality based on psychophysiological computing. Ph.D. thesis, South China University of Technology, Guangzhou (2023)
7. Huron, D.: Sweet Anticipation: Music and the Psychology of Expectation. MIT Press (2006)
8. Iida, H., Takeshita, N., Yoshimura, J.: Using games to study law of motions in mind. In: Proceedings of the 10th Advances in Computer Games Conference (2004)
9. Khalid, M.N.A., Iida, H.: Objectivity and subjectivity in games: understanding engagement and addiction mechanism. IEEE Access **9**, 65187–65205 (2021)
10. Khalid, M.N.A., Iida, H., Yusof, U.K., Mat, R.C.: Guest editorial: Special issue on 'artificial intelligence and entertainment science: Empathic entertainment technology' (2023)
11. Krumhansl, C.L.: Cognitive Foundations of Musical Pitch. Oxford University Press (1990)
12. Lartillot, O., Toiviainen, P.: A matlab toolbox for musical feature extraction. In: Proc. International Conference on Music Information Retrieval (ISMIR) (2007)
13. Li, S., Khalid, M.N.A., Iida, H.: Strategic selection and design of the first auction item: Analyzing auction dynamics through" motion in mind" and" potential reinforcement energy". Asia-Pacific Journal of Information Technology & Multimedia **13**(2) (2024)
14. Liang, C.: The balance between harmony and melody in piano playing. Art Criticism (17), 59–61 (2021)
15. Meyer, L.B.: Emotion and Meaning in Music. University of Chicago Press (1956)
16. Narmour, E.: The Analysis and Cognition of Basic Melodic Structures: The Implication-Realization Model. University of Chicago Press (1990)
17. Numan, M., Iida, H., Khalid, M.N.A.: Modeling sports engagement: A game refinement theory perspective on game length and scoring frequency. IEEE Access (2025)
18. Pearce, M., Wiggins, G.: Auditory expectation: The information dynamics of music perception and cognition. Cognition **108**(1), 97–130 (2012)
19. Schubert, E.: Continuous measurement of self-report emotional response to music. Psychology of Music **29**(4), 295–309 (2001)
20. Temperley, D.: The Cognition of Basic Musical Structures. MIT Press (2001)
21. Temperley, D.: Music and Probability. MIT Press (2007)
22. XiangLin, Z.: Research on the uncertainty formation mechanism and performance of interactive electronic music based on Max platform. Master's thesis, Nanjing University of the Arts (2023)
23. Xin, L.: Personalized development of music learning and creation in the era of big data. Art Education (06), 33–35 (2025)