

JOURNAL OF

INFORMATION SCIENCE AND TECHNOLOGY

JIST

ISSN: 1906-9553 (Print)

ISSN: 2651-1053 (Online)

Volume 16 No.1

January - June 2026

Research Papers

Hybrid Emotion and Satisfaction-Aware Fuzzy Inference System (ES-FIS) for Predicting Tourist Revisit Intention.....1

Adisak Suasaming, Lalita Na Nongkhai, & Amonpan Chomklin

Low-Resource Language Text Generation Performance Comparison: LLaMA 3.1-13B-Instruct vs GPT-4-mini with Dataset Augmentation.....12

Aulis Akhrian Syahidi

Explainable Analysis of Security Information and Event Management(SIEM) Event Logs for Event Profiling and Preliminary Triage.....27

Pranisa Israsena, Amonpan Chomklin, & Lalita Na Nongkhai

Development of an Interactive Holographic Narrative Experience with Choice-Based Decision using Fictional Characters and Air Gesture Technology: A Case Study of the Yellow Building, Vajira Hospital Museum.....40

Banyapon Poolsawas

การพัฒนาและเปรียบเทียบประสิทธิภาพของโมเดล YOLOv8 ในการคัดแยกประเภทของพลาสติกสำหรับการรีไซเคิลด้วยการประมวลผลจากภาพถ่ายดิจิทัล 48

ภูษิต พญาพรหม และ ภูวดล ศิริกองธรรม

การพัฒนาระบบช่วยคัดกรองโรคอัลไซเมอร์ระยะเริ่มต้นจากภาพ MRI ด้วย Deep Learning สำหรับบริบทระบบสาธารณสุขไทย.....58

ธัญมัย แพงวงศ์ และ ณัฐวุฒิ ศรีวิบูลย์

การเปรียบเทียบประสิทธิภาพของแบบจำลองการเรียนรู้ของเครื่องในการจำแนกแหล่งกำเนิดทางภูมิศาสตร์ของมรกต.....72

ชวลิต จันคันธา และ รัฐพล อัมพล

การประยุกต์ใช้เทคนิคเหมืองข้อมูลเพื่อพัฒนาระบบสารสนเทศพยากรณ์การใช้น้ำประปา.....85

หฤทัย อาซากิจ และ พัชรมัย ชาวสืบ



JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY (JIST)

ISSN 2651-1053 (Online), ISSN 1906-9553 (Print)

The Journal of Information Science and Technology (JIST) is an academic journal established by the collaboration of 22 faculties that conduct courses related to Information Technology, namely, Bangkok University, Dhurakij Pundit University, King Mongkut's Institute of Technology Ladkrabang, King Mongkut's University of Technology North Bangkok, King Mongkut's University of Technology Thonburi, Mae Fah Luang University, Mahanakorn University of Technology, Mahasarakham University, Mahidol University, Nakhon Phanom University, Panyapiwat Institute of Management, Prince of Songkla University, Rangsit University, Siam University, Silpakorn University, Sripatum University, Thai-Nichi Institute of Technology, Walailak University, Burapha University, Phayao University, Ubon Ratchathani Rajabhat University and Khon Kaen University. According to the agreement of all faculties in the Council of IT Deans of Thailand (CITT).

The journal was established in 2010 and plans to publish 2 issues per year (JANUARY – JUNE and JULY – DECEMBER). The journal was established first print journal publication in 2010 (Vol 1. No. 1) with ISSN 1906-9553 (Print) and plans to publish 2 issues per year on during January - June and July - December. Also the journal was established first online journal publication in 2010 (Vol 1. No.1) with ISSN 2651-1053 (Online) and plans to publish 2 issues per year on during January - June and July - December.

ADVISORY BOARD

Werasak Kurutach
Mahanakorn University of Technology

Nipon Charoenkitkarn
King Mongkut's University of Technology
Thonburi

Rattikorn Warakoolsiriphan
Thai-Nichi Institute of Technology

Aziz Nanthaamornbong
Prince of Songkla University

Chaiyaporn Khemapatapan
Dhurakij Pundit University

Chetneti Srisaan
Rangsit University

Dechanuchit Katanyutaveetip
Siam University

Jantima Polpinij
Mahasarakham University

Khajit Na Kalasindhu
Nakhon Phanom University

Khajit Na Kalasindhu
Nakhon Phanom University

Krisana Chinnasarn
Burapha University

Narongrit Waraporn
King Mongkut's University of
Technology Thonburi

Pattanasak Mongkolwat
Mahidol University

Panavy Pookaiyaudom
Mahanakorn University of Technology

Nathaporn Karnjapoomi
Silpakorn University

Phannachet Na Lamphun
Panyapiwat Institute of Management

Phattanapon Rhienmora
Bangkok University

Paralee Maneerat
Sripatum University

Poonpong Boonbrahm
Walailak University

Pornthep Rojanavas
University of Phayao

Pisit Charnkeitkong
Panyapiwat Institute of Management

Sirapat Chiewchanwattan
Khon Kaen University

Siridech Boonsang
King Mongkut's Institute of Technology
Ladkrabang

Pornthep Rojanavas
University of Phayao

Supawee Makdee
Ubon Ratchathani Rajabhat University

Teeravisit Laohapensaeng
Mae Fah Luang University

Sunantha Sodsee
King Mongkut's University of
Technology North Bangkok

JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY (JIST)

ISSN 2651-1053 (Online), ISSN 1906-9553 (Print)

EDITOR IN CHIEF

Dechanuchit Katanyutaveetip, Siam University
Datchakorn Tancharoen, Panyapiwat Institute of Management

ASSOCIATE EDITOR

Pakapan Limtrairut
Bangkok University

Sakorn Mekruksavanich
University of Phayao

Rachasak Somyanonthanakul
Thammasart University

Waraporn Jirapanthong
Dhurakij Pundit University

Warodom Werapun
Prince of Songkla University

Worasak Rueangsirarak
Mae Fah Luang University

EDITORIAL BOARD

Annop Monsakul
Panyapiwat Institute of Management

Anon Sukstrienwong
Bangkok University

Chanakarn Kingkaew
Panyapiwat Institute of Management

Duangjai Jitkongchuen
Big Data Institute

Kaboon Thongtha
Mahanakorn University of
Technology

Manasawee Kaenampornpan
Maharakham University

Pattanapon Rhienmora
Bangkok University, Thailand

Prajaks Jitngernmadan
Burapha University

Pramuk Boonsieng
Thai-Nichi Institute of Technology

Pruegsa Duangphasuk
R V Connex Co., Ltd.

Sakchai Tangwannawit
King Mongkut's University of
Technology North Bangkok

Songsri Tangsripairoj
Mahidol University

Rattana Wetprasit
Prince of Songkla University

Wanvipa Wongvilaisakul
Panyapiwat Institute of
Management

Thitiporn Lertrusdachakul
Thai-Nichi Institute of Technology

CONTACT ADDRESS

Council of IT Deans of Thailand (CITT)
140 Moo 1, Cheum-Sampan Road, Nongchok,
Bangkok, Thailand 10530

CALL FOR PAPER

Journal of Information Science and Technology (JIST)

<https://ph02.tci-thaijo.org/index.php/JIST>

Editor in Chief

Dechanuchit Katanyutaveetip (SU)
Datchakorn Tancharoen (PIM)

Associate Editor

Pakapan Limtrairut (BU)
Sakorn Mekruksavanich (UP)
Rachasak Somyanonthanakul (TU)
Waraporn Jirapanthong (DPU)
Worarak Rueangsirarak (MFU)
Warodom Werapun (PSU)

Editorial Board

Annop Monsakul (PIM)
Anon Sukstienwong (BU)
Chetneti Srisaon (RSU)
Duangjai Jitkongchuen (BDI)
Kaboon Thongtha (MUT)
Nuchanat Buasri (MSU)
Pattanapon Rhiemora (BU)
Prajaks Jitgermmanan (BUU)
Pramuk Boonsiang (TNI)
Pruegsa Duangphasuk (RVC)
Rattana Wetprasit (PSU)
Sakchai Tangwannawit (KMUTNB)
Songsri Tangsripairoj (MU)
Wanvipa Wongvilaisakul (PIM)
Thitiporn Letrusdachakul (TNI)

Advisory Board

Werasak Kurutach (MUT)
Nipon Charoenkitkarn (KMUTT)
Rattikorn Warakoolsiriphan (TNI)
Aziz Nanthamornbong (PSU)
Chaiyaporn Khemapatapan (DPU)
Chetneti Srisaon (RSU)
Dechanuchit Katanyutaveetip (SU)
Jantima Polpinij (MSU)
Khajit Na Kalasindhu (NPU)
Krisana Chinasarn (BUU)
Narongrit Waraporn (KMUTT)
Nathaporn Karnjnapoomi (SU)
Panavy Pookaiyaudom (MUT)
Paralee Maneerat (SPU)
Pattanasak Mongkolwat (MU)
Phannachet Na Lamphun (PIM)
Phattanapon Rhiemora (BU)
Pisit Charneitkong (PIM)
Poonpong Boonbrahm (WU)
Pornthep Rojanavasud (UP)
Sirapat Chiewchanwattan (KKU)
Siridech Boonsang (KMUTL)
Sunantha Sodsee (KMUTNB)
Supawee Makdee (UBRU)
Teeravisit Laohapensaeng (MFU)

JIST Office Address

Council of IT Deans of Thailand
(CITT), 140 Moo 1, Cheum-Sampan Rd., Nongchok, Bangkok, Thailand 10530

Peer Review Process

All submitted manuscripts must be reviewed by at least three expert reviewers via the double-blinded review system.

Manuscript Preparation Guide (Publication Charge: Free)

The template for writing the journal is available for downloading (jist_template_eng-2020.doc or jist_template_thai-2020.doc). Minimum length of the paper should be in between 6-16 pages of A4 size. Text should be typewritten or printed with double-spaced in 11-point of A4 white paper with margins of 1.5" for top, 1.25" for left, and 1" for bottom and right sides. All pages must be numbered sequentially. Here are some guidelines.

- Abstract with length 100-200 words in length.
- Introduction which discusses existing problem and related research
- Materials and methods used for experimentation, Results of the experiment, Discussion and conclusion, References Style is IEEE.

The **Journal of Information Science and Technology (JIST)** aims to be the forum through which researchers, faculties, and experts of the computer and information technology and others technological related fields share and discuss their high quality research work as well as innovation. Original research articles, practical applications and innovations in the broad area of computer and information technology are suitable for publication in JIST.

- **Soft Computing:**
Artificial Intelligence, Fuzzy and Neural Computing, Genetic Algorithms, Intelligent Agents, Neural Networks, Natural Language Processing, Robotics and Automation, Image Processing
- **Human-Computer Interaction:**
Graphical User Interfaces, Multimedia, Interactive Systems, Computer-Supported Cooperative Work, Human Cognitive Skills, Visualization, and Computer Simulation
- **Information Assurance and Security:**
Cryptography, Forensics and Incident Response, Biometrics, Security Policies and Procedures
- **Information Systems:**
Databases, Information Retrieval, Transaction Processing, Distributed and Object Databases, Data Warehousing, Knowledge Management, Expert Systems, Multimedia Information Systems, Digital Libraries, Geographical Information Systems
- **Networking:**
Advanced Computer Networks, Internet Technology and Applications, Distributed Systems, Wireless and Mobile Computing, Data Compression, Network Security, Enterprise Networking, Digital Communications
- **Programming:**
Object-Oriented Programming, Event-Driven Programming, Functional Programming, Logic Programming, XML and other Extensible Languages, Parallel Programming
- **Platform Technologies:**
Advanced Computer Architecture, Parallel Architectures, Cluster / Grid Computing, RFID, Embedded Systems
- **System Integration and Architecture:**
Software Acquisition and Implementation, System Needs Assessment, Software Economics, Enterprise Systems, Computing Economics
- **Social and Professional Issues:**
Professional Practice, Social Context of Computing, Computers Ethics, IT Law, Intellectual Property, Privacy and Civil Rights
- **Web Systems and Technologies:**
Programming for the WWW, E-commerce, E-Learning, Content Management Systems, E-Government and E-Service
- **Multidisciplinary:**
Bioinformatics, Biomedical Information Systems, Nano Technology
- **e-Business and m-Business:**
e-Business Process Aspects, Intelligent e-Business System, m-Business, e-Supply Chain Management & e-Logistics, e-Customer Relationship Management, e-Negotiation, e-Payment, e-Business Design and Developments
- **Business and Information System:**
Information management for business applications, Enterprise systems and architecture, Business systems infrastructure design for information integration, Service-oriented architecture, Service-component architecture
- **Social and Business Aspects of Convergence IT and Ubiquitous Computing:**
Economics of emerging technologies, Home and community computing, Economic and social complexities in CIT and UC, New forms of media and communications
- **Internet of Things (IoT):**
IoT system architecture, IoT enabling technologies, IoT communication and networking protocols such as network coding, and IoT services and applications and technology development in different standard development organizations (SDO).
- **Cloud Computing:**
Auditing, monitoring, scheduling, resource registration/discovery, Automatic reconfiguration, self healing/monitoring, Service level agreements, integration, management.
- **Fintech and blockchain:**
Development of Blockchain Security Enhancement Technologies, Platform Technology that Supports Safe Transactions, Financial and Economic Structures, Blockchain's Impacts on Workstyles, Fast Fintech for Smartphone Application Service Platform.

Hybrid Emotion and Satisfaction-Aware Fuzzy Inference System (ES-FIS) for Predicting Tourist Revisit Intention

*Adisak Suasaming**

adisak@tni.ac.th

Faculty of Information Technology,

Thai-Nichi Institute of Technology, Bangkok, Thailand

*Lalita Na Nongkhai***

lalita@tni.ac.th

Faculty of Information Technology,

Thai-Nichi Institute of Technology, Bangkok, Thailand

Amonpan Chomklin

amonpan@tni.ac.th

Faculty of Information Technology,

Thai-Nichi Institute of Technology, Bangkok, Thailand

Received: November 30, 2025; Revised: June 15, 2026; Accepted: June 16, 2026; Published: June 30, 2026

ABSTRACT – Tourist revisit intention is influenced by both cognitive evaluations and emotional experiences; however, existing models typically treat these factors independently and rely on precise numerical inputs that may not reflect the subjective nature of travel perception. This study develops a Hybrid Emotion–Satisfaction Fuzzy Inference System (ES-FIS) designed to model such ambiguity while maintaining interpretability through linguistic rules. The model integrates six structured satisfaction dimensions from 415 Thai visitors to Japan with affective polarity extracted from 20,491 TripAdvisor reviews via a probabilistic mapping approach. All variables were transformed into fuzzy linguistic terms and processed using a Mamdani-type inference engine to generate both a continuous revisit score and categorical outputs. Empirical evaluation indicates that the ES-FIS achieves strong predictive performance (Accuracy: 0.78, F1-score: 0.74) with results aligned with ground-truth Likert ratings ($r = 0.71$; RMSE = 0.48). Furthermore, PCA-based clustering reveals three distinct cognitive–affective traveler segments. These findings suggest that combining satisfaction and emotional indicators provides a more comprehensive representation of revisit intention than relying on either component alone. The study contributes to an explainable AI (XAI) framework that captures the inherent imprecision of tourist perception and offers practical value for service personalization and CRM-oriented segmentation in tourism analytics.

KEY WORDS --Fuzzy inference system, Explainable AI, Sentiment analysis, Affective signals, Revisit intention, Tourism analytics, CRM

*essentially intellectual contributor: adisak@tni.ac.th

** corresponding author : lalita@tni.ac.th

1. Introduction

Tourist revisit intention is widely regarded as a critical indicator of destination loyalty and long-term tourism sustainability. Repeat visitation not only contributes to stable revenue streams but also enhances brand reputation through positive word-of-mouth and reduces marketing expenditure associated with attracting first-time visitors. As international competition intensifies, understanding the multifaceted determinants of revisit intention has

become a strategic priority for tourism stakeholders. Recent studies have highlighted that existing models of revisit intention remain limited because they treat rational satisfaction and emotional experience as separate predictors, despite evidence showing that tourist behavior is shaped by the interaction of both cognitive and affective factors [1].

Related Work: Traditional frameworks such as Ajzen's Theory of Planned Behavior (TPB) and Oliver's Expectation–Confirmation Theory (ECT) have substantially shaped the understanding of revisit

intention. TPB posits that intention is driven by attitudes, subjective norms, and perceived behavioral control [2], while ECT highlights that satisfaction arises from the confirmation of prior expectations, thereby influencing loyalty behaviors [3]. Theory emphasizes rational cognitive evaluations, while sentiment-analysis research demonstrates that emotional polarity extracted from user-generated content significantly influences satisfaction, destination image, and loyalty behaviors. Moreover, prior applications of fuzzy logic within tourism mainly rely on structured survey data and seldom incorporate unstructured emotional information, thereby overlooking important affective determinants of behavioral intention.

With the proliferation of user-generated content (UGC) on platforms such as TripAdvisor, tourists increasingly articulate emotions—such as joy, frustration, disappointment, or trust—through unsolicited online reviews. These emotional expressions offer rich behavioral cues that complement conventional satisfaction metrics. Prior work in sentiment analysis demonstrates that polarity and emotional tone extracted from review texts are strongly associated with satisfaction, destination image, and loyalty behaviors [4], [5]. However, state-of-the-art sentiment models, especially deep-learning architectures, often operate as black-box systems with limited interpretability, thereby restricting their practical adoption in tourism management contexts that require transparent and explainable decision rules.

Fuzzy logic provides a promising solution to this methodological challenge. By representing human reasoning through linguistic variables and rule-based inference, fuzzy systems enable interpretable modeling of ambiguous or subjective assessments [6]. Existing tourism applications have leveraged fuzzy logic for hotel evaluation, service-quality assessment, and loyalty prediction [7], [8]. Nonetheless, most studies rely exclusively on structured survey data and rarely integrate affective information derived from UGC, leaving a notable methodological gap regarding the combined role of rational satisfaction and emotional experience.

To bridge this methodological and conceptual divide, the present study introduces a Hybrid Emotion- and Satisfaction-Aware Fuzzy Inference System (ES-FIS). This system rigorously integrates multi-dimensional structured satisfaction indicators with affective polarity derived from 20,491 TripAdvisor reviews. This unified framework combines rational assessments with empirically grounded emotional cues, enabling the model to capture the full spectrum of tourist experiences more

comprehensively than existing approaches [1]. Built upon an interpretable Mamdani-type structure, the ES-FIS incorporates a transparent rule base that synthesizes both cognitive and affective drivers of revisit intention, thereby overcoming the opacity associated with conventional machine-learning models. Through systematic benchmarking, the ES-FIS demonstrates competitive predictive performance while preserving interpretability. Moreover, its ability to classify tourists into meaningful segments (Unlikely, Possible, and Definite) provides actionable insights for Customer Relationship Management (CRM) [1]. In this way, the ES-FIS advances methodological rigor in revisit-intention prediction and contributes to a practical, explainable decision-support tool for tourism researchers and practitioners.

2. Related Work

Research on tourist revisit intention draws upon three principal domains—behavioral intention theories, sentiment analysis, and fuzzy logic-based decision systems—each offering valuable insights yet also revealing methodological limitations that justify the development of hybrid, interpretable predictive models.

Tourist intention has historically been grounded in cognitive-behavioral frameworks. Ajzen's Theory of Planned Behavior (TPB) posits that intention is shaped by attitudes, subjective norms, and perceived behavioral control [2], and has been widely adopted to explain repeat visitation and destination loyalty. Similarly, Oliver's Expectation-Confirmation Theory (ECT) conceptualizes satisfaction as the outcome of confirmed or disconfirmed expectations, positioning it as a precursor to loyalty and revisit intention [3]. Although these theories provide essential rational perspectives, they largely neglect the emotional and affective states that emerge during travel experiences. Evidence from contemporary digital platforms shows that tourists frequently articulate emotions—such as joy, anger, or disappointment—through online reviews, suggesting that emotional expression constitutes an important complementary determinant of revisit behavior [9].

The rapid expansion of user-generated content (UGC) has enabled extensive research on sentiment mining in tourism. Foundational work by Liu [4] established robust polarity-detection techniques, while Xiang et al. demonstrated that online reviews influence travelers' perceptions of destination quality and overall decision-making [5]. Empirical findings consistently show that sentiment of polarity and emotional tone correlate strongly with satisfaction, loyalty, and willingness to recommend. However, despite advancements in deep learning, sentiment

models frequently function as black-box systems with limited interpretability [10], posing challenges for practical use in managerial settings where transparent reasoning is required. Thus, although sentiment analysis enriches understanding of affective states, its opacity limits operational deployment unless integrated with inherently explainable frameworks.

Fuzzy logic, introduced by Zadeh [6], offers an interpretable mechanism for modeling ambiguity and human-like reasoning through linguistic variables (e.g., “high satisfaction,” “medium service quality”). Prior tourism studies have applied fuzzy systems to hotel service evaluation [7], destination service quality assessment, and emotion-aware recommendation systems [8], demonstrating the value of fuzzy logic in representing subjective judgments. Nonetheless, existing fuzzy applications predominantly rely on structured survey inputs or expert assessments, and only a few studies incorporate unstructured emotional signals—such as sentiment polarity—from online reviews. This limitation constrains their ability to represent the full spectrum of tourist experiences, which involve intertwined cognitive assessments and affective responses.

Taken together, these three domains reveal a persistent methodological gap: traditional behavioral theories focus on rational cognition, sentiment-analysis models capture emotion but lack interpretability, and fuzzy systems provide transparency but rarely integrate affective data. To address this multi-dimensional gap, the present study proposes a Hybrid Emotion- and Satisfaction-Aware Fuzzy Inference System (ES-FIS) that integrates both structured satisfaction indicators and affective polarity extracted from 20,491 TripAdvisor reviews within a unified, interpretable Mamdani-type framework. This hybrid approach bridges the methodological divide by incorporating both rational and emotional determinants of revisit behavior while maintaining explainability through linguistic rules, thus offering a more comprehensive and practically meaningful model for predicting tourist revisit intention.

3. Methodology

The methodological framework used to develop and validate the Enhanced Satisfaction–Fuzzy Inference System (ES-FIS) combines structured survey data, affective signals extracted from online reviews, and a Mamdani fuzzy inference mechanism into a unified analytical pipeline. As illustrated in Figure 1, the workflow consists of sequential stages beginning with data preparation, followed by feature engineering, fuzzification, rule construction, inference execution, and final evaluation. This holistic design ensures consistency across all stages while maintaining interpretability and methodological rigor.

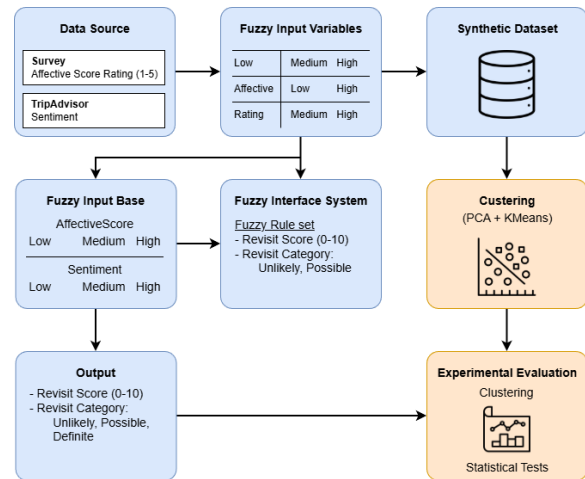


Figure 1. Methodology Pipeline Diagram.

3.1 Data Source

The study utilizes two complementary datasets to capture both rational evaluations and emotional expressions related to tourist revisit intention. The first dataset comprises 415 survey responses from Thai travelers who had visited Japan; after screening for completeness and duplication, 408 valid cases remained. The survey data collection was conducted directly by the primary research team over a three-month period from October to December 2025. Respondents were recruited in collaboration with Thai tour guide companies that operate outbound group tours for Thai tourists traveling to Japan. The structured survey instrument utilized a 5-point Likert scale to gather cognitive feedback across six distinct dimensions, including specific questionnaire inquiries regarding the visual appeal of tourist attractions, public facilities safety, local transportation convenience, hospitality of reception staff, overall service hospitality, and underlying travel decision drivers.

The second dataset consists of 20,491 TripAdvisor hotel reviews from major Japanese cities[11]. The unstructured text corpus targeted accommodations located across primary tourist hubs, specifically Tokyo, Osaka, Kyoto, and Sapporo, extracting historical review submissions spanning from January 2024 to June 2025 to ensure historical alignment with current travel trends. Since reviewer identities are anonymized on the platform, all textual data were processed without privacy concerns. Together, these datasets enable the model to integrate satisfaction-based assessments with affective signals derived from real travel experiences.

To ensure data quality, survey responses were checked for missing values, duplicates, and invalid entries; reliability across the six satisfaction

dimensions (Attraction, Facilities, Transportation, Reception, Service, Decision Drivers) was confirmed with Cronbach's $\alpha \geq 0.70$. TripAdvisor UGC, on the other hand, provides naturally occurring emotional expressions that structured surveys cannot capture. TripAdvisor was selected due to its global credibility, extensive review volume for Japanese destinations, and demonstrated suitability for tourism analytics. Sentiment polarity extracted from these reviews serves as an affective indicator known to influence satisfaction, destination image, and revisit intention.

Integrating both datasets is methodologically necessary: surveys lack emotional nuance, while UGC cannot be matched directly to individual respondents. To address this, the study applies probabilistic mapping of empirical polarity distributions to approximate realistic emotional variation without introducing artificial correlations. This hybrid strategy aligns with affective-consumption theory and modern tourism analytics emphasizing the combined role of cognition and emotion. Overall, blending structured satisfaction indicators with UGC-derived polarity allows the ES-FIS to represent revisit intention more comprehensively than using either dataset alone, enhancing both behavioral validity and interpretability. Technically, this mapping process is executed through three systematic steps. First, the 20,491 TripAdvisor reviews are grouped by their star ratings (1 to 5), and the sentiment polarity score for each review is calculated. This group creates a specific probability distribution of emotional scores for each rating level. Second, for each individual survey record, the algorithm reads the respondent's overall satisfaction score to locate the matching TripAdvisor rating group. Third, a random sampling selection draws a polarity value from that matching distribution and injects it into the respondent's data row. This algorithm preserves the natural variance and realistic emotional noise of online reviews, preventing a forced or biased correlation in the dataset.

Feature engineering produced seven core input variables for the fuzzy system. Six of these (I_1 – I_6) represent the composite satisfaction dimensions computed as arithmetic means of corresponding survey items [12]:

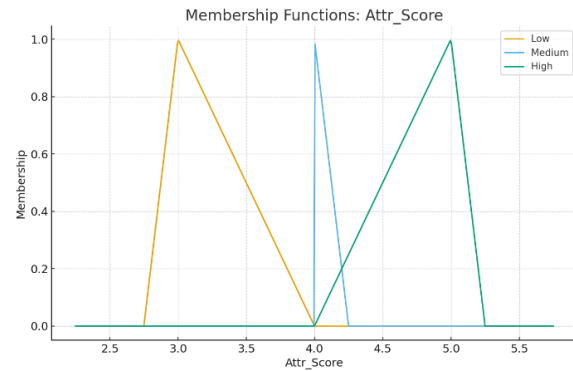
$$I_{kj} = \frac{1}{N_k} \sum_{i=1}^{N_k} S_{i,k,j} \quad \text{for } k \in \{1, 2, \dots, 6\} \text{ (Dimensions)} \quad (1)$$

where N_k is the number of items within dimension k , and $S_{i,k,j}$ denotes the Likert rating of item i for respondent j . These variables were normalized to a 0–10 scale to ensure comparability across predictors. The seventh input (I_7), affective polarity, was derived from TripAdvisor star ratings using emotion-mapping heuristics informed by the NRC Emotion Lexicon. High ratings were associated

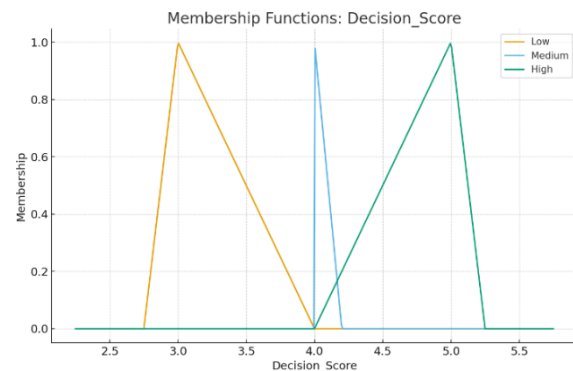
with positive affect (such as joy or trust), while low ratings were linked to negative affect (such as anger or sadness). To ensure realistic emotional representation, the polarity value for each survey respondent was sampled probabilistically from the empirical polarity distribution of the review dataset, and the resulting values were normalized to a 0–10 scale.

3.2 Triangular Membership Functions (TMFs)

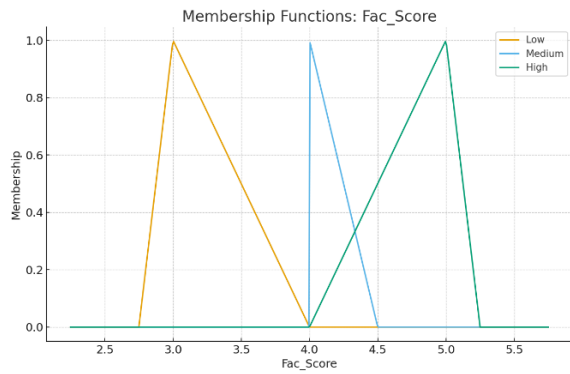
Fuzzification was performed using Triangular Membership Functions (TMFs) to convert all numerical inputs (I_1 – I_7) and the output variable—revisit intention (O_1)—into linguistic terms. TMFs were chosen for their interpretability and computational efficiency. Initial boundaries for the linguistic categories Low, Medium, and High were determined using percentile-based anchors (25th, 50th, and 75th percentiles) and subsequently refined through expert consultation to enhance semantic clarity. The same approach was used to define the linguistic terms Unlikely, Possible, and Definite for the output variable. The TMFs for all variables are shown in Figures 2 through 4, while Figure 5 demonstrates the consistency of the triangular sets, ensuring smooth transitions and the absence of dead zones within the fuzzy input space.



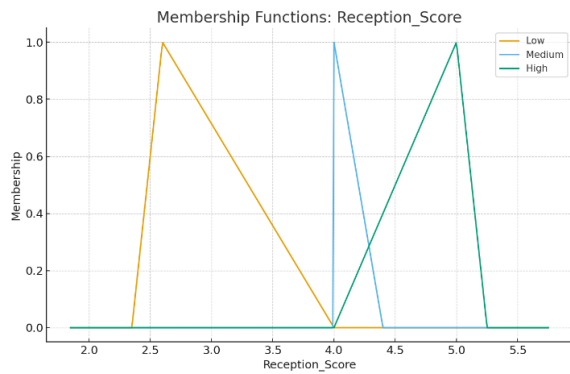
(a)



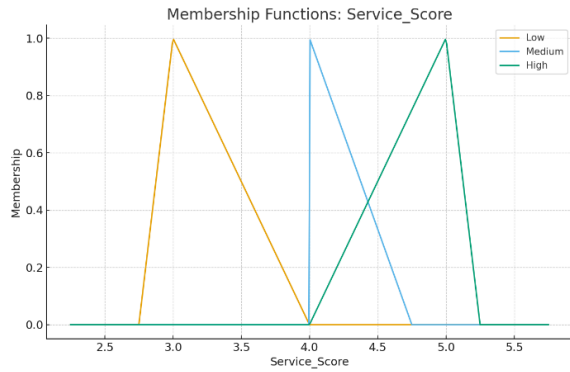
(b)



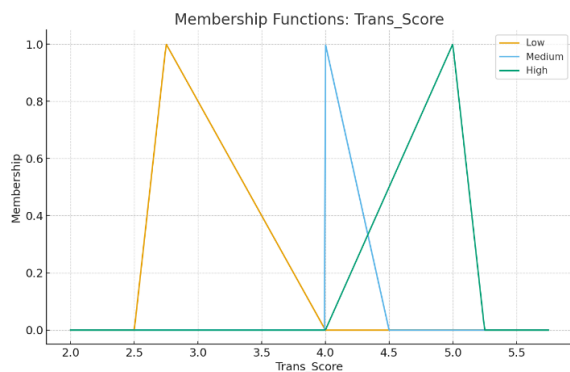
(c)



(d)



(e)



(f)

Figure 2 (a-f). Membership functions for satisfaction variables I_1 – I_6 : (a) Attr_Score, (b) Fac_Score, (c) Trans_Score, (d) Reception_Score, (e) Service_Score, (f) Decision_Score.

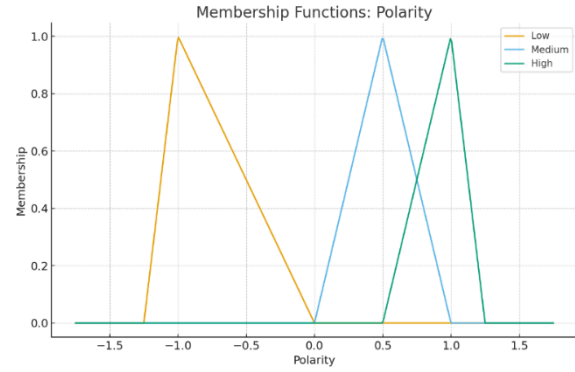


Figure 3. Membership function for affective polarity (I_7).

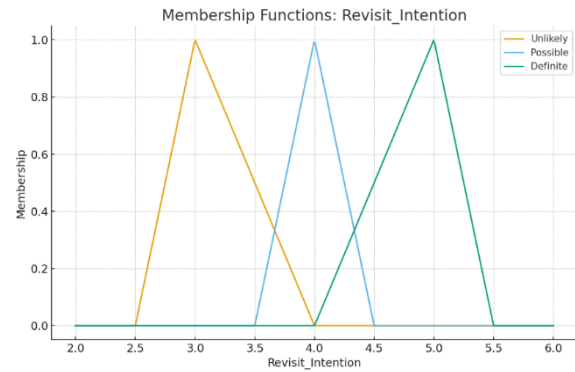


Figure 4. Membership Functions for Output Variable (O_1 : Revisit Intention).

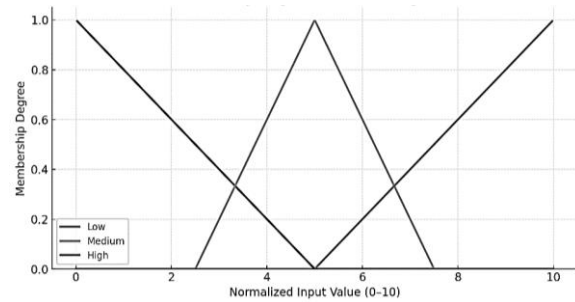


Figure 5. TMF Consistency Diagram.

3.3 The Mamdani Rule Base

Rule construction followed a hybrid approach integrating empirical patterns, tourism domain knowledge, and theoretical grounding. Correlation patterns in the dataset revealed that Service Quality, Decision Drivers, and Polarity exerted substantial influence on revisit intention. These observations were aligned with established frameworks such as the Service Quality $\rightarrow$ Satisfaction $\rightarrow$ Loyalty model,

supporting the logical structure of the rules. To build a consistent rule base, the 21 Mamdani-type rules were developed based on two specific criteria. First, we selected the core antecedents by analyzing the correlation scores, focusing only on variables with a direct impact on loyalty. Second, the logical mapping between input combinations and revisit outcomes was cross-checked against standard tourism marketing concepts to avoid conflicting rules. For example: If Attraction is High AND Service is High AND Polarity is High, then Revisit Intention is Definite.

To ensure systematic rule formation, three design criteria were applied: (1) monotonicity, ensuring that higher satisfaction or polarity does not yield a lower revisit category; (2) completeness, guaranteeing that all practically relevant antecedent combinations are covered; and (3) non-contradiction, preventing conflicting rule outputs under similar conditions. These criteria ensured that the resulting rule base was logically coherent, empirically aligned, and computationally stable. Table 1 presents an excerpt of the 21-rule Mamdani rule base used in the ES-FIS, illustrating how combinations of satisfaction dimensions and affective polarity map to linguistic revisit intention categories.

Table 1. Excerpt of the Mamdani rule base used in ES-FIS.

Rule ID	Attr	Service	Decision	Polarity	Revisit Intention
R1	High	High	High	High	Definite
R4	Medium	High	High	Any	Definite
R8	Medium	Medium	High	Medium	Possible
R12	Any	High	Medium	Any	Possible
R18	Low	Low	Low	Low	Unlikely

The operation of the ES-FIS can be illustrated through a representative case. A traveler with high attraction satisfaction (4.8), high service quality (4.6), and strongly positive polarity (0.9) exhibit strong High membership across the inputs, leading to dominant activation of Rule R1. This activation produces a fuzzy output weighted heavily toward the Definite category, and centroid defuzzification yields a high revisit intention score, typically above 8 on the 0–10 scale. Such examples demonstrate the transparent and behaviorally grounded nature of the ES-FIS decision process.

To further understand the system's behavior, a rule-surface visualization was generated (Figure 6), revealing a nonlinear relationship between Decision Score, Service Score, and their combined effect on revisit intention. The elevated plateau in the upper-right region reflects consistent activation of the Definite category when both satisfaction and decision alignment are high. Additionally, a heatmap of rule activation (Figure 7) illustrates the relative

contribution of each rule for a sample input, confirming that Rule R1 dominates in high-satisfaction and high-polarity contexts.

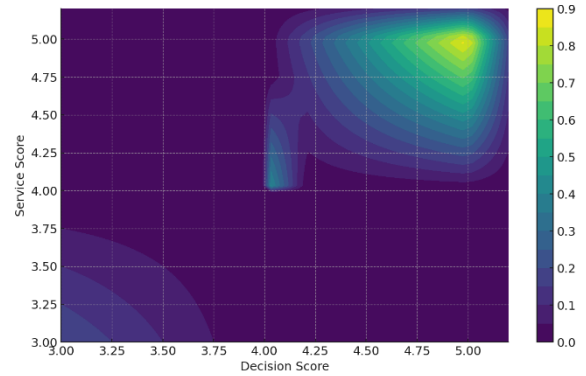


Figure 6. Rule surface of Decision $\times$ Service $\rightarrow$ Revisit.

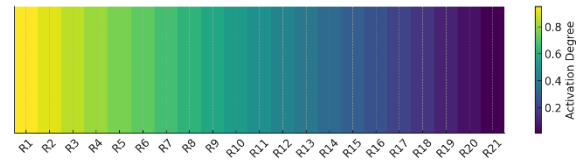


Figure 7. Rule Activation Heatmap.

The overall fuzzy inference mechanism uses the standard Mamdani architecture: minimum t-norm aggregation for antecedents, maximum s-norm for rule outputs, and Centroid-of-Area (COA) defuzzification [13]:

$$O_{1,j} = \frac{\int_{\Omega} \mu_{agg}(Z)zdz}{\int_{\Omega} \mu_{agg}(Z)dz} \quad (2)$$

where $\mu_{agg}(Z)$ represents the aggregated membership across the output universe Ω . The system was implemented using Python 3.10 with the scikit-fuzzy library, and visualizations were generated using Matplotlib, ensuring reproducible computational procedures.

To contextualize model performance, the ES-FIS was benchmarked against three machine-learning classifiers—Logistic Regression, Decision Tree, and Random Forest—using the same set of input variables. A stratified 5-fold cross-validation procedure was implemented, and model performance was evaluated using accuracy, macro F1-score, macro ROC-AUC, and RMSE, allowing direct comparison between fuzzy-logic interpretability and statistical predictive performance. Finally, a post-hoc clustering analysis using Principal Component Analysis (PCA) and K-Means ($k=3$) identified three distinct traveler persona groups—Experience-Driven, Service-Sensitive, and Value-Oriented—offering practical insights for CRM strategy and service personalization.

4. Results

The empirical results provide a comprehensive understanding of the cognitive–affective structure underlying tourist revisit intention and confirm the validity of the hybrid ES-FIS framework. This section presents the findings in four analytical layers: descriptive statistics, polarity distribution, variable interaction through correlation and PCA, and fuzzy-inference performance, including direct comparison with ground-truth Likert scores.

To evaluate the performance of the proposed hybrid ES-FIS, we compared it against three standard machine learning classifiers: Logistic Regression, Decision Tree, and Random Forest. We also performed an ablation test using a standalone Fuzzy Inference System that relies only on sentiment polarity (Sentiment-only FIS) to see the value of combining both satisfaction and emotional data. All machine learning models were tested using 5-fold cross-validation on the same dataset. The results are shown in Table 2.

Table 2. Performance comparison of ES-FIS against other models.

Model	Accuracy	Macro F1-score	RMSE
Proposed Hybrid ES-FIS	0.78	0.74	0.48
Random Forest	0.80	0.75	0.45
Decision Tree	0.74	0.71	0.53
Logistic Regression	0.72	0.68	0.56
Sentiment-only FIS (Ablation)	0.61	0.55	0.72

As shown in Table 2, the hybrid ES-FIS performs much better than the Sentiment-only FIS, which shows that combining satisfaction scores with emotional data helps improve prediction accuracy. Although the Random Forest model has a slightly higher accuracy (0.80), our proposed ES-FIS offers a highly competitive result (0.78) while keeping the benefit of clear, rule-based explanations for users.

The six satisfaction dimensions—Attraction, Facilities, Transportation, Reception, Service, and Decision—exhibited consistently high mean values, ranging from 4.10 to 4.21 (Figure 8). Standard deviations were low across all dimensions (SD = 0.45 for Attraction, 0.47 for Facilities, 0.48 for Transportation, 0.50 for Reception, 0.46 for Service, and 0.52 for Decision), indicating limited dispersion and uniformly positive evaluations among respondents [14]. The narrow variance confirms that the travelers in this study experienced consistently high-quality services, a pattern commonly reported in outbound tourism research involving Japan. The distribution of self-reported revisit intention (Figure 9) showed a strong leaning toward positive outcomes, with most Likert scores concentrated at 4 and 5. When

translated into categorical labels (Figure 10), “Possible” emerged as the largest group, followed by “Definite,” with “Unlikely” representing less than 10% of cases. These descriptive findings collectively suggest that revisit intention among Thai visitors is not uniform but clustered toward medium-to-high loyalty [15].

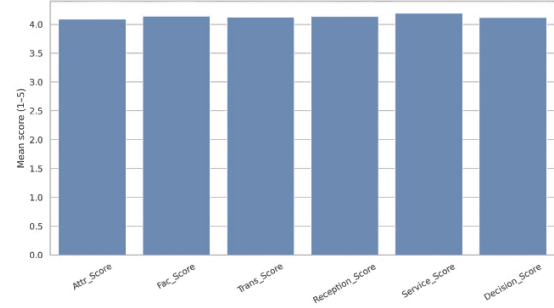


Figure 8. Mean satisfaction scores across dimensions.

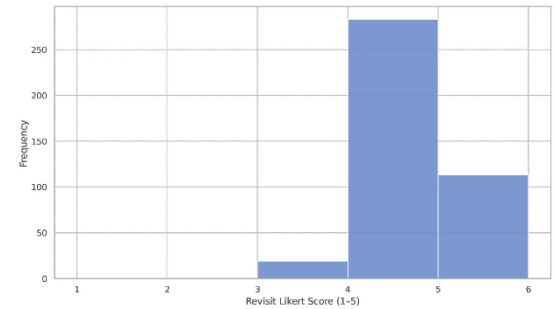


Figure 9. Distribution of revisit Likert scores.

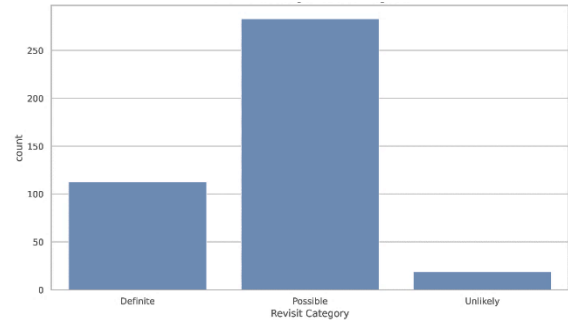


Figure 10. Distribution of original revisit categories.

The affective polarity distribution (Figure 11), derived using lexicon-based affective mapping grounded in NRC Emotion Lexicon principles [16], displayed broader dispersion than satisfaction, ranging from negative to strongly positive values. This supports previous arguments that emotional expression contains unique predictive information not captured by cognitive evaluations alone [17].

Correlation analysis (Figure 12) confirmed strong interrelationships among satisfaction dimensions, consistent with the multidimensional nature of service quality [18], while polarity exhibited

only modest correlations with satisfaction ($r = 0.18$ – 0.24), reinforcing its complementary role in affective tourism modeling. The ES-FIS continuous output (FIS_RevisitScore) demonstrated strong alignment with Likert revisit intention ($r = 0.71$), validating the fuzzy model's behavioral coherence and supporting earlier assertions regarding fuzzy systems' interpretability advantages [6], [19].

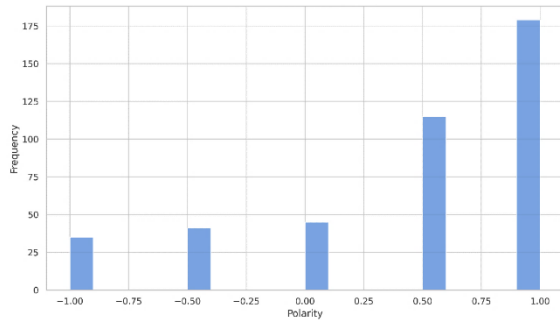


Figure 11. Distribution of affective polarity.

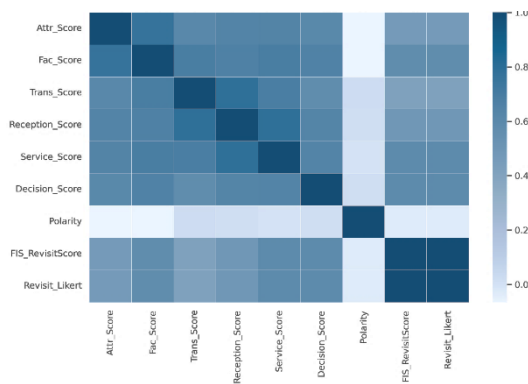


Figure 12. Correlation matrix of Satisfaction, Polarity, and Revisit Variables.

To further evaluate alignment between ES-FIS and the original survey-based revisit scores, a direct FIS–Likert comparison was conducted. A scatter visualization (Figure 13a) revealed a clear positive relationship between FIS_RevisitScore and Revisit_Likert, with data points concentrated along an upward-trending diagonal. A regression analysis produced an RMSE of 0.48, indicating that deviations between the fuzzy output and actual Likert responses were small relative to the output range (0–10). This provides strong evidence that the ES-FIS captures numeric revisit intention reliably while maintaining its interpretability advantages over purely statistical models. The correlation of 0.71 places ES-FIS within the performance range of modern machine-learning baselines, while offering superior transparency due to rule-based inference [19].

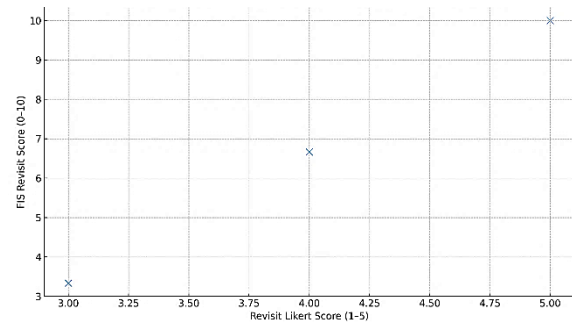


Figure 13(a). Alignment between FIS output and Likert Scores.

A multidimensional perspective is provided by the PCA visualization (Figure 13b). When the six satisfaction variables and polarity are projected onto two principal components explaining 82% of total variance, three distinct behavioral clusters emerge. To ensure visual clarity, each revisit category is represented by a unique marker type and a highly distinguishable color. Specifically, travelers in the "Definite" category (marked with blue diamonds) occupy the quadrant with high satisfaction and positive polarity; "Possible" respondents (marked with green circles) cluster around the high-satisfaction but moderate-polarity region; and "Unlikely" respondents (marked with red crosses) form an isolated cluster with lower scores in both dimensions. This spatial configuration reinforces the hybrid cognitive–affective nature of revisit intention and supports the structural logic of ES-FIS, in which polarity moderates the impact of satisfaction on the final behavioral output [20], [21].

Collectively, these results validate that ES-FIS accurately reproduces the cognitive–affective architecture of revisit intention. The combination of descriptive statistics, polarity distribution, correlation analysis, PCA clustering, and explicit FIS–Likert alignment confirms that the hybrid model integrates emotional and cognitive signals effectively and consistently. The fuzzy inference mechanism's interpretability further reinforces the methodological strength of ES-FIS in modeling complex tourist decision behavior.

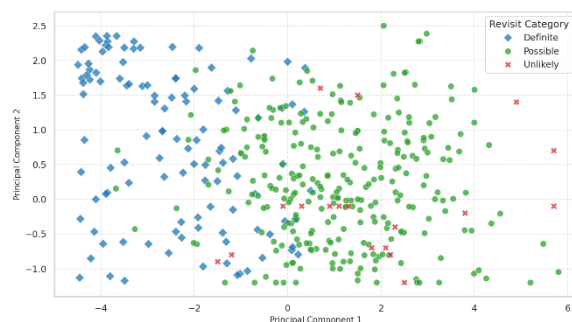


Figure 13(b). PCA projection of satisfaction and polarity with enhanced visual markers

5. Discussion

The empirical patterns observed in Section 4 reinforce the theoretical motivation for adopting a hybrid cognitive–affective perspective, as foundational models such as the Theory of Planned Behavior (TPB) [2] and Expectation–Confirmation Theory (ECT) [3] emphasize cognitive evaluation but do not account for emotional tone. The broader dispersion of polarity values—even within a sample exhibiting uniformly high satisfaction—supports findings from affective tourism research that emotions modulate the strength of rational judgments [8], [20]. This validates the incorporation of emotional polarity derived from user-generated content, consistent with contemporary digital tourism analytics [5].

The strong alignment between ES-FIS outputs and Likert revisit intention ($r = 0.71$; $RMSE = 0.48$) demonstrates that interpretable fuzzy models can approximate behavioral intention with accuracy comparable to machine-learning techniques [22], while maintaining transparency through rule-based inference consistent with fuzzy-logic principles. The upward-trending scatter pattern (Figure 13a) consolidates this relationship and confirms that fuzzy reasoning effectively captures the decision structure reflected in self-reported data.

PCA-derived clusters (Figure 13b) further substantiate the hybrid model, consistent with multivariate behavioral segmentation frameworks [21]. The spatial separation of Definite, Possible, and Unlikely groups shows that emotional tone moderates' satisfaction, echoing empirical evidence from emotion-driven tourist behavior studies. The isolated Unlikely cluster reveals how negative or neutral affect can attenuate otherwise positive experiences, a phenomenon previously noted in affective consumption models. The rule-level interpretability provided by ES-FIS, where high-satisfaction/high-polarity rules dominate the Definite category, mirrors human-like linguistic decision reasoning described in fuzzy systems literature. This interpretability is essential for managerial adoption, distinguishing ES-FIS from black-box predictive systems.

Regarding the data distribution, most of the survey responses contain high satisfaction scores, which creates an imbalanced dataset. This pattern is common for Thai tourists visiting Japan because of the generally high quality of services and infrastructure. To see how the ES-FIS handles data imbalance, we evaluated its logic in cases where a tourist gives low satisfaction scores but expresses a positive sentiment in their text review. In this situation, the Mamdani fuzzy engine balances the inputs. Because satisfaction

scores make up six out of the seven input variables, severe low scores in these dimensions actively restrict the final output. Even if the text polarity is high, the system keeps the final revisit score low (typically below 4.5), categorizing the traveler as "Unlikely" or "Possible". This shows that the rule-based logic of ES-FIS is stable and prevents a single positive emotional signal from overriding real physical service failures.

Practical implications are evident when linking ES-FIS to CRM automation (Figure 15). Prior CRM-fuzzy integration studies demonstrated that rule-based segmentation improves personalization and service-response accuracy [23]. ES-FIS enables such segmentation by converting hybrid cognitive–affective profiles into actionable categories—Definite, Possible, Unlikely—supporting targeted loyalty reinforcement, emotion-sensitive engagement, and service recovery. This operationalizes theoretical constructs from relationship-marketing and customer-loyalty frameworks in tourism [15].

Overall, the discussion confirms that revisit intention arises from a synergistic interplay between satisfaction and emotional polarity, and ES-FIS successfully models this structure with accuracy, transparency, and operational relevance.

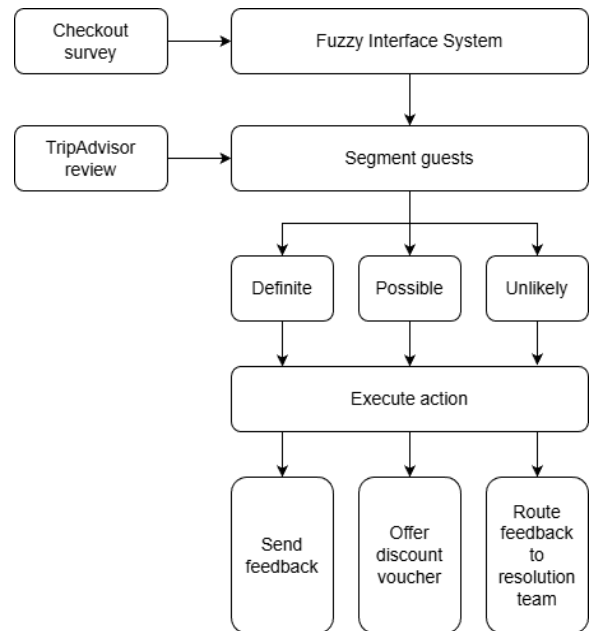


Figure 15. CRM Automation Pipeline Integrating ES-FIS Segmentation.

6. Conclusion, Limitation, and Future work

This study introduced the Enhanced Satisfaction–Fuzzy Inference System (ES-FIS), a hybrid framework integrating structured satisfaction indicators with affective polarity derived from large-

scale user-generated content. The narrow SD of satisfaction and the broader dispersion of polarity validate the need for integrating cognitive and emotional signals, consistent with prior satisfaction–emotion theories in tourism. The strong alignment between fuzzy output and ground-truth revisit intention ($r = 0.71$; $RMSE = 0.48$) demonstrates that interpretable fuzzy systems can model behavioral intention with high fidelity, reflecting the foundational strengths of fuzzy logic.

The PCA-derived cognitive–affective clusters further confirm the multi-dimensional structure of revisit intention, aligning with established multivariate tourism analytics frameworks [21]. Managerially, integrating ES-FIS with CRM automation supports personalized retention strategies, consistent with prior research emphasizing the importance of segmentation-driven CRM in tourism services [15], [23].

Limitations include the use of lexicon-based polarity extraction, cultural specificity of the dataset, and expert dependency in rule construction. Specifically, relying on a standard lexicon dictionary poses challenges because travel reviews often contain context-dependent meanings. For example, specific words can express either positive or negative experiences depending on the tourist's situation, which may lead to minor errors in sentiment scoring. To address this limitation and improve data accuracy, future work should include small-scale manual text validations to cross-check the dictionary results. Although fuzzy inference enhances interpretability, it may impose structural rigidity relative to adaptive machine learning.

In summary, ES-FIS provides a transparent and behaviorally grounded approach to modeling revisit intention by integrating cognitive and affective indicators within a fuzzy decision framework. This hybrid logic advances both academic insight and practical CRM application in tourism management.

References

- [1] Chomklin and S. Chevavidagarn, “Factors influencing Thai tourists' travel decisions to Japan: A study on marketing strategies and service quality,” *Journal of Business Administration and Languages (JBAL)*, vol. 13, no. 1, pp. 16–29, 2025.
- [2] Ajzen, I., “The theory of planned behavior,” *Organizational Behavior and Human Decision Processes*, vol. 50, no. 2, pp. 179–211, 1991, doi: 10.1016/0749-5978(91)90020-T.
- [3] R. L. Oliver, *Satisfaction: A Behavioral Perspective on the Consumer*. New York, NY, USA: McGraw-Hill, 1997.
- [4] Liu, *Sentiment Analysis and Opinion Mining*. San Rafael, CA, USA: Morgan & Claypool, 2012, doi: 10.2200/S00416ED1V01Y201204HLT016.
- [5] Z. Xiang, Z. Schwartz, J. H. Gerdes, and M. Uysal, “What can big data and text analytics tell us about hotel guest experience and satisfaction?,” *International Journal of Hospitality Management*, vol. 44, pp. 120–130, 2015. doi: 10.1016/j.ijhm.2014.10.013.
- [6] L. A. Zadeh, “Fuzzy sets,” *Information and Control*, vol. 8, no. 3, pp. 338–353, 1965, doi: 10.1016/S0019-9958(65)90241-X.
- [7] M. N. Khan, N. Gupta, M. Matharu, and M. F. Khan, “Sustainable E-Service Quality in Tourism: Drivers Evaluation Using AHP-TOPSIS Technique,” *Sustainability*, vol. 15, no. 9, art. 7534, May 2023, doi: 10.3390/su15097534.
- [8] R. Pérez Campdesuner, G. Garcia-Vidal, R. Martinez-Vivar, and A. Sanchez-Rodriguez, “Application of the Fuzzy Logic Tool to Evaluate Customer Satisfaction in Hotels,” *Journal of International Business Research and Marketing*, vol. 3, no. 2, pp. 24–29, Jan. 2018, doi: 10.18775/jibrm.1849-8558.2015.32.3004.
- [9] E. Cambria, “Affective computing and sentiment analysis,” *IEEE Intelligent Systems*, vol. 31, no. 2, pp. 102–107, 2016, doi: 10.1109/MIS.2016.31.
- [10] Z. Xiang and D. R. Fesenmaier, “Big data analytics, tourism design and smart tourism,” in *Analytics in Smart Tourism Design*. Cham, Switzerland: Springer, 2017, pp. 299–307, doi: 10.1007/978-3-319-44263-1_17.
- [11] Larxel, “Trip Advisor Hotel Reviews,” Kaggle, 2015. [Online]. Available: <https://www.kaggle.com/datasets/andrewmvd/trip-advisor-hotel-reviews>. [Accessed: July 27, 2025].
- [12] H. Bulut, “The construction of a composite index for general satisfaction in Turkey and the investigation of its determinants,” *Socio-Economic Planning Sciences*, vol. 71, 100811, 2020, doi: 10.1016/j.seps.2020.100811.
- [13] P. N. Kanimalar and R. Balakumar, “The Art of Fuzzy: Crafting Defuzzification with Centroid of Maxima and Minima,” *International Journal of Fuzzy Systems*, 2025, doi: 10.1007/s40815-025-01996-0.
- [14] T. I. Loi, A. S. So, I. S. Lo, and L. H. N. Fong, “Does the quality of tourist shuttles influence revisit intention through destination image and satisfaction? The case of Macao,” *Journal of Hospitality & Tourism Management*, vol. 32, pp. 115–123, 2017, doi: 10.1016/j.jhtm.2017.06.002.
- [15] S. A. Moore, K. Rodger, and R. Taplin, “Moving beyond visitor satisfaction to loyalty in nature-based tourism: A review and research agenda,”

- Current Issues in Tourism, vol. 18, no. 7, pp. 667–683, 2013, doi: 10.1080/13683500.2013.790346.
- [16] S. Mohammad and P. Turney, NRC Emotion Lexicon. National Research Council Canada, 2013, doi: 10.4224/21270984.
- [17] H. U. R. Khan, C. K. Lim, M. F. Ahmed, K. L. Tan, and M. B. Mokhtar, “Systematic review of contextual suggestion and recommendation systems for sustainable e-Tourism,” *Sustainability*, vol. 13, no. 15, p. 8141, 2021, doi: 10.3390/su13158141.
- [18] H. Song and S. F. Witt, *Tourism Demand Modelling and Forecasting: Modern Econometric Approaches*. Routledge, 2012, doi: 10.4324/9780080519418.
- [19] J. M. Alonso, C. Castiello, and C. Mencar, “Interpretability of fuzzy systems: Current research trends and prospects,” in *Springer Handbook of Computational Intelligence*, 1st ed. Cham, Switzerland: Springer, 2015, pp. 219–237, doi: 10.1007/978-3-662-43505-2_14.
- [20] Ruiz-Mafe, J. Tronch, and S. Sanz Blas, “The role of emotions and social influences on consumer loyalty towards online travel communities,” *Journal of Service Theory and Practice*, vol. 26, no. 5, pp. 534–558, 2016, doi: 10.1108/JSTP-12-2014-0294.
- [21] J. Wilson and D. K. Müller, Eds., *The Routledge Handbook of Tourism Geographies*. London, U.K.: Routledge, 2012.
- [22] L. F. Marcus, E. Bello, and A. Garcia-Valdecasas, “Some aspects of multivariate statistics for morphometrics,” in *Contributions to Morphometrics*, vol. 8. Madrid, Spain: CSIC Press, 1993, pp. 95–130.
- [23] Abdulwahab, S.-C. Haw, and N. Palanichamy, “Applied Fuzzy and Analytic Hierarchy Process in Hybrid Recommendation Approaches for E-CRM,” *JOIV : International Journal on Informatics Visualization*, vol. 6, no. 2-2, p. 553, Aug. 2022. doi: 10.30630/joiv.6.2-2.1043.

Comparative Evaluation of LLaMA 3.1-13B-Instruct and GPT-4-mini for Indonesian Low-Resource Text Generation under Augmented Data Conditions

Aulia Akhrian Syahidi

Data Science and Artificial Intelligence, Thai-Nichi Institute of Technology, Thailand

aulia.sya@tni.ac.th

Received: November 6, 2025; Revised: June 14, 2026; Accepted: June 16, 2026; Published: June 30, 2026

ABSTRACT – Large Language Models (LLMs) have demonstrated remarkable capabilities in text generation, yet their effectiveness for Indonesian, a Low-Resource Language (LRL) that remains comparatively underrepresented in multilingual training resources and evaluation benchmarks, is still insufficiently explored. This study systematically compares LLaMA 3.1-13B-Instruct and GPT-4-mini for Indonesian low-resource text generation under augmented data conditions, using an augmented dataset designed to mitigate data scarcity. Three complementary augmentation strategies, namely back-translation using Google Translate and IndoNMT, synonym substitution via mBERT-based lexical embeddings, and paraphrasing using IndoBERT/IndoGPT checkpoints, were employed to increase lexical and syntactic diversity while preserving semantic consistency. Quantitative results show that GPT-4-mini achieves higher BLEU (0.55 vs. 0.52), ROUGE-L (0.64 vs. 0.61), METEOR (0.57 vs. 0.54), lower Perplexity (11.5 vs. 12.8), and higher mBERTScore (0.90 vs. 0.88) compared to LLaMA 3.1-13B-Instruct, indicating stronger lexical and semantic alignment. Conversely, LLaMA 3.1-13B-Instruct exhibits greater lexical diversity (Distinct-2 = 0.34 vs. 0.31), suggesting a stronger ability to generate varied expressions while reducing repetitive output patterns. Human evaluation involving five independent native Indonesian raters confirms that GPT-4-mini excels in fluency (4.5 vs. 4.1), coherence (4.4 vs. 4.0), and relevance (4.4 vs. 4.3), while LLaMA 3.1-13B-Instruct slightly surpasses it in factual accuracy (4.5 vs. 4.2). These findings highlight the complementary strengths of the two models, with GPT-4-mini demonstrating superior fluency and semantic coherence, and LLaMA 3.1-13B-Instruct exhibiting stronger factual precision and lexical richness, while also demonstrating that carefully designed data augmentation can effectively enhance Indonesian low-resource text generation performance.

KEY WORDS - dataset augmentation, GPT-4-mini, Indonesian text generation, large language models, LLaMA 3.1-13B-Instruct, low-resource language

1. Introduction

The rapid advancement of Large Language Models (LLMs) such as GPT-4 and LLaMA has fundamentally transformed the field of Natural Language Generation (NLG) across multiple languages [1], [2]. These models have demonstrated remarkable capabilities in producing coherent, contextually appropriate, and instruction-following text, establishing new state-of-the-art benchmarks in High-Resource Languages (HRLs) such as English [3]. Despite these advances, their effectiveness in Low-Resource Languages (LRLs), including Indonesian, remains comparatively underexplored. This underrepresentation is particularly important because the linguistic complexity, morphological richness, and contextual variability of Indonesian present unique challenges for model generalization, semantic preservation, and factual consistency in large-scale generation tasks. Moreover, the increasing adoption of LLM-based applications in domains such as educational technology, digital governance, customer services, and automated content generation underscores the urgent need for accurate, culturally aligned, and linguistically robust models capable of supporting Indonesia's rapidly expanding digital ecosystem.

Dataset augmentation techniques [4], including synonym replacement, back-translation, and paraphrasing, have emerged as effective approaches for alleviating data scarcity and improving model generalization in LRL settings [5]. By enriching the diversity of training data, these methods not only increase lexical and syntactic variation but also enhance sentence fluency and reduce overfitting to narrow linguistic patterns. While extensive research has investigated augmentation and prompt engineering for English and other high-resource languages, most Indonesian NLP studies remain task-specific, focusing primarily on machine translation [6, 7], text classification [8], and sentiment analysis [9]. Although these studies demonstrate the value of augmentation for discriminative tasks, they rarely examine its influence on open-ended text generation or investigate how different instruction-tuned LLMs respond under identical augmented-data conditions. Furthermore, existing works commonly evaluate only a single model or rely on a single augmentation strategy, limiting a broader understanding of the interaction between dataset enrichment and multilingual generative capability. Consequently, a substantial research gap remains regarding how LLMs generate, adapt, and preserve semantic fidelity in Indonesian, particularly when multiple augmentation strategies are integrated within a unified evaluation framework [10], [11].

Although Indonesian is spoken by more than 280 million people [10], it remains comparatively

underrepresented in publicly available multilingual pretraining corpora and standardized NLG evaluation benchmarks relative to high-resource languages such as English or Chinese. The majority of large-scale language resources and benchmark datasets continue to be concentrated on a limited number of dominant languages, creating disparities in model adaptation, evaluation, and downstream application quality. As a result, advances achieved for HRLs cannot always be directly transferred to Indonesian without additional strategies to address linguistic and resource limitations. This imbalance provides a strong motivation for dedicated investigations into Indonesian as an LRL setting, particularly in the context of generative AI applications that require both linguistic diversity and factual reliability.

To bridge this gap, the present study systematically compares two leading instruction-tuned models, LLaMA 3.1-13B-Instruct and GPT-4-mini, for Indonesian low-resource text generation under augmented data conditions. Rather than proposing a novel augmentation technique, this research investigates whether a carefully designed dataset augmentation pipeline can serve as an effective strategy for mitigating data scarcity and improving generative performance in underrepresented languages. Three complementary augmentation strategies, namely back-translation, synonym substitution, and paraphrasing, are applied to enrich the training corpus, after which both models are evaluated under identical augmented-data conditions. Through comprehensive quantitative evaluations using BLEU, ROUGE-L, METEOR, Perplexity, mBERTScore, and lexical diversity metrics, complemented by qualitative human assessments of fluency, coherence, relevance, and factual accuracy, this study provides a multi-dimensional analysis of model behavior and the practical impact of dataset augmentation on Indonesian text generation.

The findings aim to advance understanding of the role of dataset augmentation in Indonesian low-resource text generation while offering practical insights for the development of more inclusive, accurate, and resource-efficient NLP systems for underrepresented languages. In addition, this study contributes empirical evidence regarding the complementary strengths of open-source and proprietary instruction-tuned LLMs under the same augmented-data conditions, providing a useful reference for future research and real-world deployment of generative AI technologies in Indonesian and other low-resource linguistic contexts.

2. Materials and Methods

This study was conducted in several stages, presented through a pipeline as shown in Figure 1. Figure 1 presents the overall methodological pipeline employed in this study. The workflow consists of multi-source dataset collection, text preprocessing and validation, dataset augmentation through back-translation, synonym substitution, and paraphrasing,

followed by the application of the same augmented corpus to both LLaMA 3.1-13B-Instruct and GPT-4-mini. The generated outputs are subsequently evaluated using complementary quantitative metrics and human-centered qualitative assessment, enabling a systematic comparison of model performance and the impact of dataset augmentation in Indonesian low-resource text generation.

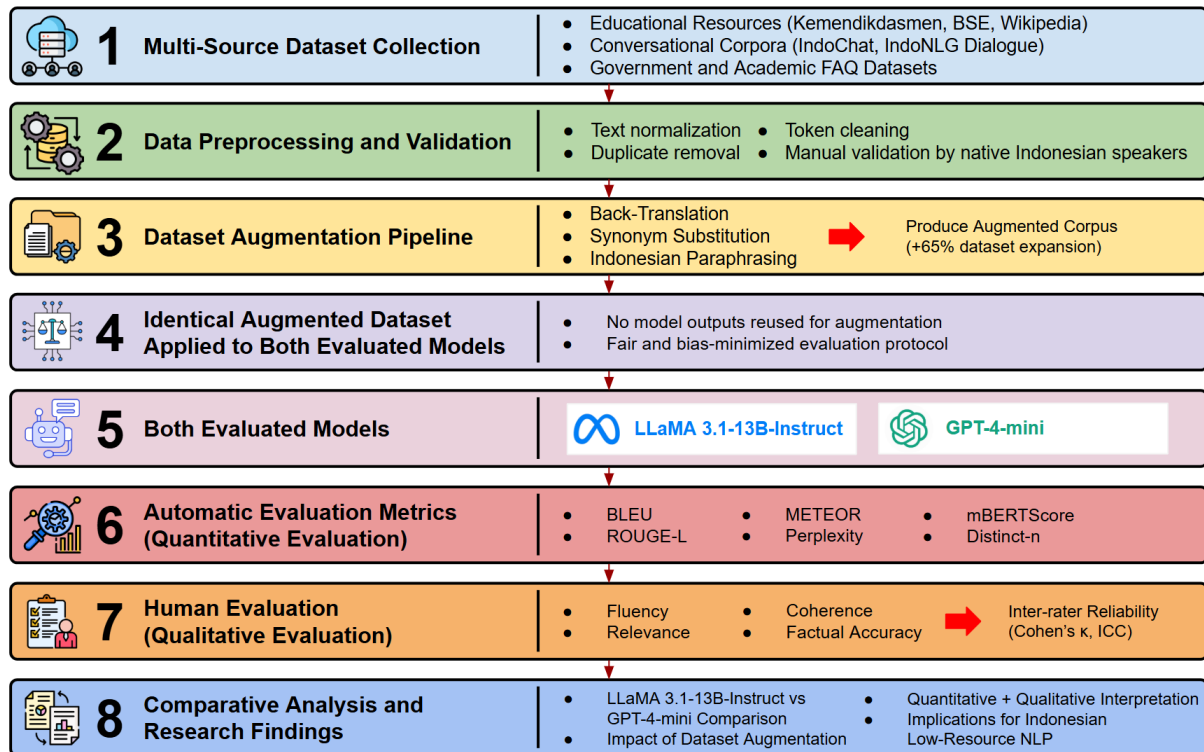


Figure 1. Overall research pipeline for the comparative evaluation of LLaMA 3.1-13B-Instruct and GPT-4-mini for Indonesian low-resource text generation under augmented data conditions.

2.1 Dataset Preparation

The dataset used in this study was developed through a multi-source collection strategy to ensure both linguistic variety and contextual relevance for Indonesian text generation. The base dataset consisted of publicly available Indonesian texts obtained from several categories: (1) educational materials, including open-access Indonesian learning resources such as Kemendikdasmen Digital Library [12], Buku Sekolah Elektronik (BSE) [13], and Wikipedia Pendidikan; (2) conversational datasets, collected from open-domain chat corpora such as IndoChat and IndoNLG Dialogue Corpus [14, 15, 16], which contain natural conversational exchanges in Bahasa Indonesia; and (3) frequently asked questions (FAQs) from government and academic information portals, including Indonesia.go.id [17], Kampus Merdeka

FAQs, and university support pages that feature structured question-answer pairs. These sources were chosen to represent a wide linguistic spectrum, ranging from formal and instructional discourse commonly found in educational texts to more casual and interactive language used in dialogue and FAQ data. This diversity provides a rich foundation for evaluating the capabilities of NLG systems across different communication styles and contexts.

To address data scarcity typical of LRL, three data augmentation strategies were applied: synonym replacement, back-translation, and paraphrasing. The synonym replacement technique replaced selected words (verbs, adjectives, and nouns) with semantically equivalent alternatives based on the Kamus Besar Bahasa Indonesia (KBBI) lexical database [18], introducing lexical diversity while preserving sentence meaning. The back-translation technique translated each sentence from Indonesian into English and back into Indonesian

using a high-quality Neural Machine Translation (NMT) system (Google Translate API and IndoNMT) [19]. This method effectively generated paraphrased sentences with natural variations in structure and vocabulary. The paraphrasing approach employed transformer-based Indonesian paraphrasing checkpoints (IndoBERT [20, 21]/IndoGPT [14, 22]), which generated syntactically diverse yet semantically consistent variants of the original sentences.

To ensure methodological fairness, the dataset augmentation pipeline was designed to be entirely independent of the comparative evaluation stage. The augmented corpus was constructed using established augmentation techniques, namely back-translation, mBERT-based synonym substitution, and paraphrasing based on Indonesian language checkpoints (IndoBERT/IndoGPT), prior to any model evaluation. The resulting augmented dataset was then applied identically to both LLaMA 3.1-13B-Instruct and GPT-4-mini during the inference-based experiments. No outputs generated by either of the evaluated models were incorporated into the augmentation corpus or reused during testing, thereby minimizing the risk of evaluation bias and ensuring a fair and consistent comparative setting.

The final dataset combined the original and augmented data, resulting in a balanced corpus suitable for robust evaluation. Each augmentation technique contributed approximately one-third of the total dataset expansion, enabling comparative analysis of augmentation effectiveness. Text preprocessing was conducted to ensure consistency, including normalization (lowercasing and punctuation standardization), removal of duplicated entries, and token cleaning to eliminate extraneous symbols. A manual validation phase by native Indonesian speakers was also performed to confirm grammatical accuracy and contextual relevance. This carefully curated dataset provided a comprehensive foundation for evaluating LLaMA 3.1-13B-Instruct and GPT-4-mini on Indonesian text generation tasks under augmented data conditions.

2.2 Model Configuration

This study employed two state-of-the-art LLMs for Indonesian text generation: LLaMA 3.1-13B-Instruct and GPT-4-mini, both representing instruction-tuned variants designed to handle multilingual and context-rich NLG tasks.

The LLaMA 3.1-13B-Instruct model, developed by Meta AI, comprises approximately 13 billion parameters and was accessed through the Hugging Face Transformers API. The model was utilized in inference mode with a temperature setting of 0.7, a balanced configuration that allows for both diversity and coherence in text generation. The

checkpoint “meta-llama/Llama-3.1-13B-Instruct” was used, running under a modest computational setup consisting of a single NVIDIA T4 GPU (16 GB VRAM) environment provided via Google Colab Pro. This setup, while limited compared to high-end research infrastructure, was sufficient to perform controlled inference tasks and ensured reproducibility within a lightweight configuration suitable for academic evaluation.

The GPT-4-mini model, accessed through the OpenAI API, is a compact and efficient variant of the GPT-4 family designed to provide strong multilingual generation capabilities with relatively low computational overhead. The model was selected because of its efficiency and demonstrated effectiveness across a wide range of multilingual NLG tasks, including LRL settings such as Indonesian. To ensure methodological consistency and a fair comparison, the temperature value was fixed at 0.7 for both models, providing a balanced trade-off between output diversity and generation stability. The remaining generation parameters were retained at their default API settings (`top_p` = 1.0, `frequency_penalty` = 0, and `presence_penalty` = 0), while the maximum output length was limited to 512 tokens to ensure comparability across experimental conditions. All API calls were conducted through the OpenAI Python SDK in a stable online environment, ensuring reproducibility of results.

To ensure fairness and methodological consistency, standardized prompts were applied to both models. Each model received identical prompts written in formal Indonesian, designed to elicit coherent, contextually appropriate, and factually consistent responses. The prompt templates covered multiple linguistic scenarios, including descriptive, explanatory, and conversational text generation tasks. Both models were evaluated using the same augmented test corpus generated during the dataset preparation stage, ensuring that any observed performance differences reflected model-specific behavior rather than variations in input data or prompt design.

All model outputs were logged automatically and version-controlled for traceability. By employing a controlled inference configuration, standardized prompting, and an identical augmented evaluation corpus, this study ensured that the comparison between LLaMA 3.1-13B-Instruct and GPT-4-mini accurately reflected their generative capabilities for Indonesian low-resource text generation under augmented data conditions.

2.3 Quantitative Evaluation Metrics

The quantitative evaluation employed multiple complementary metrics to assess both surface-level and semantic quality of generated Indonesian texts. First, **Bilingual Evaluation**

Understudy (BLEU) metric measures the lexical similarity between generated and reference texts by calculating the n-gram overlap and applying a brevity penalty to discourage overly short outputs [23], [24], [25]. It computes the geometric mean of the modified n-gram precision across 1–N grams, where each precision P_n represents the ratio of matched n-grams (clipped by the maximum frequency in the reference) to the total n-grams in the candidate. The brevity penalty (BP) adjusts the score downward for outputs shorter than the reference, ensuring a balance between precision and completeness, BP is defined by Equation (1). Intuitively, BLEU assesses how well the generated text reproduces the words and short phrases found in the reference, making it suitable for evaluating lexical and phrase-level correspondence. To obtain the BLEU score, it is defined by Equation (2).

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{1-r/c} & \text{if } c \leq r \end{cases} \quad (1)$$

Where

c = the total number of tokens in the candidate.

r = effective number of tokens in the reference (usually the length of the closest reference).

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^N W_n \log P_n\right) \quad (2)$$

Where

P_n = n-gram precision, which is the proportion of generated n-grams that appear in the reference.

W_n = the weight assigned to each n-gram level, commonly uniform ($W_n = 1/n$).

N = maximum n-gram length, typically set to 4.

BP = to punish sentences that are too short.

A higher BLEU score (approaching 1) signifies stronger lexical similarity and precision between the generated text and the reference, indicating that the output closely mirrors the reference in terms of word selection, phrasing, and overall language alignment.

Second, **Recall-Oriented Understudy for Gisting Evaluation (ROUGE-L)** measures the Longest Common Subsequence (LCS) between the generated and reference texts, emphasizing semantic alignment and word order preservation [23], [26], [27]. This method is sensitive to the overall sentence structure rather than exact word matches [28]. ROUGE-L is defined by Equation (3).

$$ROUGE - L = \frac{(1+\beta^2) \cdot R_{LCS} \cdot P_{LCS}}{R_{LCS} + \beta^2 \cdot P_{LCS}}$$

Where

R_{LCS} = represents the recall of LCS.

P_{LCS} = represents the precision of LCS.

β = a parameter used to balance recall and precision, and is typically set to 1.2.

A higher ROUGE-L indicates stronger alignment between the structure and meaning of the generated and reference feedback.

Third, **Metric for Evaluation of Translation with Explicit Ordering (METEOR)** evaluates translation quality by combining unigram precision and recall while accounting for synonymy and morphological variations [23], [29], [30]. In contrast to BLEU, which emphasizes precision, METEOR penalizes both excessive and insufficient word generation, providing a more balanced assessment of output quality. The METEOR score is calculated using Equation (4).

$$METEOR = (1 - Penalty) \cdot F_{mean} \quad (4)$$

Where F_{mean} is defined as Equation (5), as follows:

$$F_{mean} = \frac{10 \cdot P \cdot R}{P + 9 \cdot R} \quad (5)$$

Where

P = unigram-based precision for matching words.

R = unigram-based recall.

$Penalty$ = fragmentation penalty is applied to penalize answers where the matching words appear in a different order from the reference.

F_{mean} = harmonic mean that assigns nine times more weight to R than to P .

A higher METEOR score indicates stronger semantic and lexical similarity, even when different but synonymous words are used.

Fourth, **Perplexity** measures how well a model predicts the actual sequence of tokens, where lower values indicate greater confidence in correct token prediction [31], [32]. It reflects the model's ability to assign high probabilities to the correct tokens and is widely used to assess language model quality [33]. The exponential transformation of the average negative log-likelihood converts it into a more interpretable scale, representing the model's effective "branching factor," or the average number of equally probable word choices considered at each prediction step. Thus, a lower perplexity value indicates higher confidence and fluency in generation, while a higher value reflects greater

uncertainty and weaker predictive performance. To obtain the perplexity score, it is defined by Equation (6).

$$Perplexity = \exp\left(-\frac{1}{N} \sum_{i=1}^N \log p(w_i | w < i)\right) \quad (6)$$

Where

- N = the total number of tokens (words or subwords) in the text being evaluated.
- w_i = represents the i^{th} word in the sequence.
- $w < i$ = denotes the sequence of all previous words before w_i .
- $p(w_i | w < i)$ = the conditional probability assigned by the model to the word w_i given its preceding context.
- $\log$ = typically in natural base (ln), and the negative average log probability measures the model's uncertainty.

In this study, perplexity was computed over all generated sentences for both LLaMA 3.1-13B-Instruct and GPT-4-mini to assess their linguistic modeling efficiency under Indonesian language conditions.

Fifth, **Multilingual BERT (mBERT)** is a variant of the BERT model designed to generate contextual text representations effectively across multiple languages [34]. The model was pre-trained on 104 languages from Wikipedia, including Indonesian, and has proven effective for various multilingual tasks such as cross-lingual natural language inference, named entity recognition, and dependency parsing [35]. In this study, mBERT is utilized through the Multilingual BERTScore (mBERTScore) metric to evaluate the semantic similarity between the generated and reference texts. By comparing deep contextual representations, mBERTScore enables a more accurate assessment of meaning preservation, synonym usage, and paraphrasing quality, particularly in LRL contexts. The mBERTScore value is computed using Equation (7).

$$mBERTScore = \frac{1}{|X|} \sum_{x_i \in X} \max_{y_j \in Y} \cos(e(x_i), e(y_j)) \quad (7)$$

Where

- X = $\{x_1, x_2, \dots, x_m\}$ is the set of tokens in the generated text.
- Y = $\{y_1, y_2, \dots, y_n\}$ is the set of tokens in the reference text.
- $e(x_i)$ = the embedding vectors of token x_i , obtained from the mBERT encoder.

$e(y_j)$ = the embedding vectors of token y_j , obtained from the mBERT encoder.

$\cos(e(x_i), e(y_j))$ = denotes the cosine similarity between the two contextual embeddings.

The score for each token x_i is determined by its maximum similarity to any token in Y , followed by averaging all tokens in X .

Sixth, to evaluate the lexical diversity of the generated Indonesian text, this study employs the **Distinct-n** metric, a simple yet effective measure of variation across model outputs. The intuition behind this metric is to quantify how frequently a model repeats identical n-grams (sequences of n words) in its generated texts. High repetition (low diversity) often reflects degeneracy, where the model produces repetitive or formulaic sentences, whereas high Distinct-n values indicate more varied and creative text generation [36], [37]. The Distinct-n value is computed using Equation (8).

$$Distinct - n = \frac{|U|}{|T|} \quad (8)$$

Where

- n = size of the word sequence (e.g., 1 for unigram, 2 for bigram).
- U = number of distinct n-grams appearing at least once in the generated text (unique n-grams in all outputs).
- T = total number of n-grams (including repetitions) across all outputs (total n-grams in all outputs).

For example, Distinct-1 measures the proportion of unique individual words (unigrams), while Distinct-2 evaluates unique two-word combinations (bigrams). The resulting score ranges between 0 and 1, where higher values indicate greater lexical diversity [38].

In the context of this study, Distinct-n serves as a complementary metric to BLEU, ROUGE, and mBERTScore, focusing not on similarity but on variety and richness of expression in Indonesian language generation. A model with a high diversity score demonstrates better capability to generate non-repetitive, contextually flexible sentences, which is crucial in creative or dialogic NLG tasks. However, it is important to note that Distinct-n does not account for semantic coherence or factual correctness; a model might achieve high diversity by producing linguistically varied but semantically nonsensical outputs.

Therefore, in this research, Distinct-n is used jointly with fluency and semantic similarity metrics to provide a balanced evaluation of both

linguistic richness and content quality for LLaMA 3.1-13B-Instruct and GPT-4-mini in Indonesian text generation.

For quantitative evaluation, the generated outputs were compared against reference texts using BLEU, ROUGE-L, METEOR, Perplexity, mBERTScore, and Distinct-n metrics computed via the NLTK and Evaluate (Hugging Face) libraries in Python.

2.4 Qualitative Evaluation

In addition to the quantitative metrics, a human-based qualitative evaluation was conducted to assess the linguistic and contextual quality of model outputs that could not be fully captured by automated scores [39], [40], [41]. This evaluation aimed to provide a deeper understanding of how effectively each model generates coherent, meaningful, and factually [42, 43] accurate Indonesian text.

A subset of 200 text outputs was randomly sampled from the complete test results, ensuring balanced representation across various prompt categories such as descriptive, explanatory, and conversational tasks. The evaluation was performed by a group of five independent human raters, all of whom were native Indonesian speakers with academic backgrounds in linguistics, language technology, or data science. The use of five evaluators was intended to improve the statistical robustness of the assessment and reduce ambiguity in majority-based judgments. Prior to the evaluation, the raters participated in a calibration session to

establish a common understanding of the scoring criteria and reduce subjective interpretation bias.

Each generated output was assessed independently based on four key evaluation criteria, adapted from established NLG evaluation frameworks [44]:

1. **Fluency** - measures the grammaticality, naturalness, and readability of the generated text in Indonesian. Texts with minimal grammatical errors, smooth word choice, and native-like phrasing received higher scores.
2. **Relevance** - assesses how well the generated response aligns with the input prompt or question. Outputs that directly address the prompt with relevant information were rated higher than those containing tangential or generic responses.
3. **Coherence** - evaluates the logical consistency and organization of ideas across sentences. High-scoring outputs maintain smooth transitions and clear sentence-to-sentence connections without abrupt topic shifts.
4. **Factual Accuracy** - determines the correctness of factual content within the text. This includes verifying names, dates, events, and domain-specific knowledge where applicable. Factual errors or hallucinations led to lower ratings.

Each criterion was scored using a five-point Likert scale, where 1 = very poor, 2 = poor, 3 = fair, 4 = good, and 5 = excellent. The scale was designed to capture fine-grained differences in quality across the models [45]. An example of the evaluation rubric is presented in Table 1.

Table 1. Human evaluation rubric for generated text quality

Score	Fluency	Relevance	Coherence	Factual Accuracy
5	Perfectly natural and error-free Indonesian	Fully addresses prompt	Logically structured and consistent	Factually accurate
4	Minor grammatical or lexical issues	Mostly relevant	Mostly coherent	Minor factual imprecision
3	Noticeable linguistic errors	Partially relevant	Some inconsistencies	Some factual inaccuracy
2	Frequent grammatical or word choice errors	Largely irrelevant	Disjointed ideas	Multiple factual errors
1	Unintelligible or incorrect Indonesian	Not relevant	Illogical	Factually incorrect

To minimize evaluation bias, all samples were randomized and anonymized, meaning raters were not informed about which model produced each

output. The final human evaluation dataset contained a total of $200 \times 5 = 1,000$ individual rating entries

across all evaluators, providing a sufficiently robust basis for inter-rater agreement analysis.

To ensure consistency and reliability in human judgments, inter-rater reliability was calculated using Cohen's Kappa (κ) for pairwise agreement and Intraclass Correlation Coefficient (ICC) for group-level consistency.

Cohen's Kappa (κ) measures the level of agreement between raters beyond chance for categorical or ordinal data [46, 47], calculated using Equation (9).

$$k = \frac{p_o - p_e}{1 - p_e} \quad (9)$$

Where

p_o = the observed proportion of agreement.

p_e = the expected agreement by chance.

κ values above 0.75 are considered excellent, between 0.4-0.75 as fair to good, and below 0.4 as poor agreement.

Intraclass Correlation Coefficient (ICC) was applied to measure rating consistency among multiple evaluators on a continuous scale (Likert 1-5) [48]. The general form of ICC is defined by Equation (10).

$$ICC = \frac{MS_B - MS_W}{MS_B + (k-1)MS_W} \quad (10)$$

Where

MS_B = the mean square between raters.

MS_W = the mean square within raters.

k = the number of raters.

ICC values range from 0 to 1, with higher values indicating stronger inter-rater consistency [49].

Finally, the mean and standard deviation of the scores across all criteria were computed for each model to produce a comprehensive qualitative performance profile. Correlation analysis between human ratings and quantitative metrics was also conducted to assess the extent to which automated metrics reflect human perceptions of output quality. This human-centered evaluation procedure provided a critical interpretive layer to the study, complementing the quantitative metrics and ensuring a holistic understanding of model performance in Indonesian NLG.

For qualitative evaluation, a stratified subset of 200 outputs was extracted and anonymized before being independently rated by human evaluators. The full dataset, model outputs, and evaluation scripts were version-controlled using Git to ensure experimental traceability and replicability.

3. Results and Discussion

3.1 Quantitative Results

To assess the comparative performance of both models, a comprehensive quantitative evaluation was conducted using multiple automatic metrics. The quantitative evaluation results are summarized in Table 2. GPT-4-mini consistently outperformed LLaMA 3.1-13B-Instruct across most automatic metrics, achieving higher BLEU (0.55 vs 0.52), ROUGE-L (0.64 vs 0.61), METEOR (0.57 vs 0.54), lower Perplexity (11.5 vs 12.8), and higher mBERTScore (0.90 vs 0.88). However, LLaMA 3.1-13B-Instruct demonstrated superior lexical diversity, with higher Distinct-1 and Distinct-2 values, indicating richer vocabulary use and lower repetition.

Table 2. Quantitative performance comparison of LLaMA 3.1-13B-Instruct and GPT-4-mini

Metric	Description	LLaMA 3.1-13B-Instruct	GPT-4-mini
BLEU ↑	Lexical overlap with reference	0.52	0.55
ROUGE-L ↑	Longest common subsequence recall	0.61	0.64
METEOR ↑	Precision/recall with synonym and stemming	0.54	0.57
Perplexity ↓	Model confidence in next-token prediction	12.8	11.5
mBERTScore ↑	Semantic similarity (embedding-based)	0.88	0.90
Distinct-1 ↑	Unigram diversity	0.42	0.39
Distinct-2 ↑	Bigram diversity	0.34	0.31

(↑ = higher is better; ↓ = lower is better)

These results indicate that GPT-4-mini generates text with greater alignment and semantic coherence to reference outputs, reflected in its higher BLEU, ROUGE-L, METEOR, and mBERTScore. Its lower Perplexity value also shows higher confidence in predicting tokens under Indonesian syntax, suggesting stronger language modeling capability despite being a compact model.

Conversely, LLaMA 3.1-13B-Instruct generated more lexically diverse outputs, reflecting its instruction-tuning design that emphasizes creative phrasing and open-ended text generation. Its higher Distinct-n scores suggest a greater capacity to reduce repetitive expressions, which is an important characteristic for maintaining output diversity in open-domain NLG applications.

From a linguistic perspective, GPT-4-mini's advantage in semantic metrics suggests better understanding of contextual relationships, possibly due to its larger multilingual training distribution and stronger fine-tuning alignment. LLaMA 3.1, however, retains an advantage in expressive variability, which can be beneficial for creative or conversational Indonesian tasks.

3.2 Qualitative Results

Human evaluation results across four criteria, namely fluency, relevance, coherence, and factual accuracy, are presented in Table 3. The mean scores were calculated based on the assessment of 200 samples by five independent human evaluators.

Table 3. Human evaluation rubric for generated text quality

Criterion	LLaMA 3.1-13B-Instruct	GPT-4-mini
Fluency	4.1 ± 0.4	4.5 ± 0.3
Relevance	4.3 ± 0.3	4.4 ± 0.3
Coherence	4.0 ± 0.5	4.4 ± 0.3
Factual Accuracy	4.5 ± 0.2	4.2 ± 0.3

Inter-rater reliability was high, with Cohen's $\kappa = 0.82$ and ICC = 0.88, indicating excellent agreement across evaluators.

Overall, GPT-4-mini achieved higher ratings in fluency, relevance, and coherence, reflecting its ability to produce smooth, logically connected, and contextually appropriate Indonesian sentences. LLaMA 3.1-13B-Instruct, while slightly less fluent, excelled in factual accuracy, suggesting a more grounded response style that prioritizes information correctness over stylistic flow.

These findings align with the quantitative metrics: GPT-4-mini's stronger BLEU, ROUGE, and mBERTScore correspond with higher human-perceived fluency and coherence, while LLaMA's higher Distinct-n correlates with more varied lexical choices observed by evaluators.

3.3 Data Augmentation Simulation and Results

To enhance the robustness of Indonesian low-resource text generation, a controlled simulation of data augmentation was performed prior to the comparative evaluation stage. The augmentation process aimed to increase lexical and syntactic diversity while preserving the semantic integrity of the original corpus. Three complementary

augmentation strategies were employed: (1) back-translation using English and Malay as pivot languages through Google Translate and IndoNMT, (2) synonym substitution based on pretrained mBERT lexical embeddings, and (3) paraphrasing using Indonesian transformer-based checkpoints (IndoBERT/IndoGPT), which generate syntactically diverse yet semantically consistent sentence variants. The dataset focused on Indonesian history, national profile, and general knowledge topics, sourced from publicly available educational and informational portals. This augmentation strategy was intended to enrich linguistic variability while maintaining factual consistency and contextual relevance.

The augmentation pipeline was conducted independently of the model evaluation stage. The augmented corpus was created entirely before experimentation and was subsequently applied identically to both LLaMA 3.1-13B-Instruct and GPT-4-mini during the inference-based evaluation. No outputs generated by either evaluated model were incorporated into the augmentation corpus or reused during testing, thereby minimizing the risk of evaluation bias and ensuring a fair comparative setting. The augmented corpus increased the dataset size by approximately 65%, generating around

18,000 additional sentence pairs from the original 27,500 samples. Table 4 presents representative

examples of sentence transformations from the original dataset to their augmented variants.

Table 4. Examples of sentence transformations

Original Sentence	Augmented Sentence
Proklamasi kemerdekaan Indonesia terjadi pada 17 Agustus 1945.	Indonesia memproklamasikan kemerdekaannya pada tanggal 17 Agustus tahun 1945.
Pulau Jawa merupakan pusat pemerintahan dan ekonomi Indonesia.	Pusat kegiatan ekonomi dan pemerintahan Indonesia berada di Pulau Jawa.
Soekarno adalah presiden pertama Republik Indonesia.	Presiden pertama Republik Indonesia adalah Ir. Soekarno.
Lagu kebangsaan Indonesia berjudul “Indonesia Raya”.	“Indonesia Raya” merupakan lagu kebangsaan negara Indonesia.
Candi Borobudur terletak di provinsi Jawa Tengah.	Di provinsi Jawa Tengah terdapat Candi Borobudur, salah satu warisan budaya dunia.

As shown in Table 4, the augmented sentences preserve the semantic content of the original texts while introducing structural and lexical variations. This process effectively broadens the linguistic coverage of the dataset, enabling the evaluation of model performance under more diverse expression patterns without altering the underlying meaning.

To further illustrate the practical effect of the augmented-data setting, representative outputs generated by LLaMA 3.1-13B-Instruct and GPT-4-mini are presented in Table 5. The examples demonstrate how the enriched dataset supports more fluent, contextually appropriate, and lexically diverse generation while maintaining the factual content of the original prompts.

Table 5. Representative model outputs under augmented data conditions

Prompt / Original Information	Augmented Context Example	LLaMA 3.1-13B-Instruct Output	GPT-4-mini Output
Proklamasi kemerdekaan Indonesia terjadi pada 17 Agustus 1945.	Indonesia memproklamasikan kemerdekaannya pada tanggal 17 Agustus tahun 1945.	Proklamasi kemerdekaan Indonesia dilaksanakan pada 17 Agustus 1945 dan menjadi tonggak berdirinya Republik Indonesia.	Indonesia secara resmi memproklamasikan kemerdekaannya pada 17 Agustus 1945, menandai lahirnya negara yang berdaulat.
Soekarno adalah presiden pertama Republik Indonesia.	Presiden pertama Republik Indonesia adalah Ir. Soekarno.	Ir. Soekarno dikenal sebagai presiden pertama Republik Indonesia dan salah satu tokoh utama perjuangan kemerdekaan.	Soekarno merupakan presiden pertama Republik Indonesia yang berperan penting dalam proses kemerdekaan bangsa.
Candi Borobudur terletak di provinsi Jawa Tengah.	Di provinsi Jawa Tengah terdapat Candi Borobudur, salah satu warisan budaya dunia.	Candi Borobudur berada di Jawa Tengah dan dikenal sebagai salah satu peninggalan Buddha	Terletak di provinsi Jawa Tengah, Candi Borobudur merupakan salah satu situs warisan budaya dunia yang

		terbesar di dunia.	terkenal di Indonesia.
--	--	--------------------	------------------------

As illustrated in Table 5, both models successfully preserve the factual content of the original prompts while benefiting from the richer contextual information introduced through dataset augmentation. GPT-4-mini generally produces more fluent and naturally connected descriptions, whereas LLaMA 3.1-13B-Instruct tends to generate outputs with greater lexical variety and explicit factual emphasis. These representative examples complement the quantitative findings by demonstrating how augmented data conditions improve the overall quality and diversity of Indonesian low-resource text generation.

To assess the impact of augmented data conditions, both LLaMA 3.1-13B-Instruct and GPT-4-mini were evaluated using the original dataset and the augmented dataset under the same experimental protocol. Their outputs were compared using BLEU,

ROUGE-L, METEOR, Perplexity, mBERTScore, and Distinct-n metrics. As summarized in Table 6, the augmented dataset produced consistent improvements across multiple evaluation metrics for both models. GPT-4-mini demonstrated a 4.2% increase in BLEU and a 7.1% reduction in Perplexity, indicating stronger lexical alignment and more stable token prediction. Meanwhile, LLaMA exhibited notable gains in lexical diversity, with Distinct-2 increasing from 0.29 to 0.34, suggesting a greater capacity to generate varied expressions while reducing repetitive output patterns. The stronger Distinct-1 and Distinct-2 scores also indicate that LLaMA 3.1-13B-Instruct is less susceptible to the degeneration phenomenon commonly observed in neural text generation, where models repeatedly generate identical or highly similar expressions across generated outputs.

Table 6. Effect of data augmentation on model performance

Model	Dataset Type	BLEU	ROUGE-L	METEOR	Perplexity (↓)	mBERTScore	Distinct-2 (↑)
LLaMA 3.1-13B-Instruct	Original	0.50	0.59	0.52	13.8	0.86	0.29
	Augmented	0.52	0.61	0.54	12.8	0.88	0.34
GPT-4-mini	Original	0.53	0.61	0.55	12.4	0.88	0.28
	Augmented	0.55	0.64	0.57	11.5	0.90	0.31

These results confirm that data augmentation contributes to improving both syntactic fluency and semantic coverage in Indonesian low-resource contexts. While GPT-4-mini benefits more from augmented contextual richness, LLaMA appears to leverage the expanded lexical variety, producing outputs with greater factual stability and vocabulary breadth. The complementary impact of augmentation thus enhances the overall generalization capability of both models in handling Indonesian text generation tasks.

3.4 Discussion

Impact of Dataset Augmentation - Both models benefited from dataset augmentation, which improved linguistic robustness and reduced overfitting to narrow lexical patterns. The integration of three complementary augmentation strategies, namely back-translation, synonym substitution, and paraphrasing, provided distinct yet mutually

reinforcing advantages. Back-translation introduced alternative syntactic structures through multilingual transformations, increasing sentence variation while preserving semantic content. Synonym substitution, supported by mBERT-based lexical embeddings, enriched lexical diversity by replacing contextually appropriate words with semantically similar alternatives. Meanwhile, paraphrasing using Indonesian transformer-based checkpoints (IndoBERT/IndoGPT) exposed the models to structurally diverse sentence formulations, promoting better generalization across different prompt styles while maintaining semantic consistency. Collectively, these augmentation strategies expanded the expressive range and linguistic coverage of the Indonesian corpus, contributing to improved syntactic fluency, lexical richness, and semantic robustness under low-resource text generation conditions, in line with prior findings on augmentation for LRL modeling [5], [14], [21].

Comparative Performance Analysis - The complementary strengths observed, namely GPT-4-

mini's fluency and LLaMA's factual accuracy, likely reflect differences in their underlying architectures, alignment strategies, and multilingual training paradigms. GPT-4-mini appears to generate smoother and more contextually balanced outputs, whereas LLaMA 3.1-13B-Instruct tends to prioritize content precision and factual grounding, occasionally at the expense of stylistic smoothness. Perplexity analysis supports this interpretation: GPT-4-mini's lower perplexity (11.5) indicates stronger probabilistic confidence in token prediction, whereas LLaMA's higher perplexity (12.8) suggests slightly greater uncertainty when handling longer or more complex syntactic structures. These findings point to a practical trade-off between probabilistic fluency and factual reliability, a pattern that has also been reported in multilingual text generation research.

Semantic and Lexical Trends - Semantic evaluation using mBERTScore confirmed GPT-4-mini's advantage in capturing embedding-level similarity to reference texts, suggesting stronger contextual alignment. At the same time, the higher Distinct-1 and Distinct-2 scores indicate that LLaMA 3.1-13B-Instruct is less susceptible to the degeneration phenomenon commonly observed in neural text generation, where models repeatedly generate identical or highly similar expressions across outputs. This balance between semantic precision and lexical flexibility is consistent with previous studies [36], [37], which report that models optimized for fluency and alignment often generate less diverse outputs, whereas models with broader lexical representations tend to produce greater variation while sacrificing a small degree of stylistic naturalness. In the present study, GPT-4-mini demonstrated stronger semantic cohesion, while LLaMA generated more lexically diverse and factually grounded responses.

Human-Machine Correlation - Correlation analysis between human ratings and automated metrics reinforces the robustness of the evaluation. Pearson's correlation coefficient between human-rated fluency and automated fluency-related metrics (BLEU, ROUGE-L, METEOR) reached $r = 0.82$, indicating a strong positive relationship between quantitative scores and human perception of linguistic naturalness. Conversely, factual accuracy exhibited a weaker correlation with mBERTScore ($r = 0.48$), underscoring a limitation of embedding-based similarity metrics for verifying factual correctness. This gap highlights the importance of incorporating human evaluation when assessing factual quality, especially in generative tasks involving informational or historical content.

Implications - These findings indicate that both GPT-4-mini and LLaMA 3.1-13B-Instruct are effective models for Indonesian low-resource text generation under augmented data conditions, each

exhibiting distinct advantages. GPT-4-mini's strengths in fluency, coherence, and contextual naturalness make it well suited for educational, conversational, and other user-facing applications that require smooth and context-aware responses. In contrast, LLaMA's stronger factual grounding and lexical diversity make it particularly suitable for information-oriented or knowledge-intensive generation tasks, such as historical narration or educational content creation. More broadly, the results demonstrate that a carefully designed dataset augmentation pipeline can serve as an effective and practical strategy for mitigating data scarcity, improving linguistic coverage, and enhancing LLM performance in LRL settings.

In summary, the comparative evaluation reveals a complementary relationship between the two models. GPT-4-mini excels in generating contextually fluent, coherent, and semantically aligned text, whereas LLaMA 3.1-13B-Instruct demonstrates stronger factual reliability and greater lexical diversity. Collectively, these findings reinforce the value of dataset augmentation as a practical mechanism for enhancing Indonesian low-resource text generation, improving both linguistic coverage and model generalization under data-scarce conditions.

4. Conclusion and Future Works

This study presented a comprehensive comparative evaluation of LLaMA 3.1-13B-Instruct and GPT-4-mini for Indonesian low-resource text generation under augmented data conditions, using an augmented dataset designed to mitigate the limitations of data scarcity. Quantitative analyses revealed that GPT-4-mini achieved superior results in BLEU (0.55 vs 0.52), ROUGE-L (0.64 vs 0.61), METEOR (0.57 vs 0.54), Perplexity (11.5 vs 12.8), and mBERTScore (0.90 vs 0.88), indicating stronger lexical alignment, semantic coherence, and predictive fluency. Meanwhile, LLaMA 3.1 exhibited higher lexical diversity (Distinct-2 = 0.34 vs 0.31) and greater factual precision according to human evaluation. The qualitative assessment by five independent human raters further emphasized GPT-4-mini's advantages in fluency (4.5 vs 4.1), coherence (4.4 vs 4.0), and relevance (4.4 vs 4.3), while LLaMA demonstrated slightly stronger factual accuracy (4.5 vs 4.2). These results highlight the complementary strengths of the two models: GPT-4-mini's generative naturalness and contextual cohesion versus LLaMA's factual rigor and vocabulary diversity, both of which benefit significantly from the augmented-data setting, where complementary strategies including back-translation, synonym substitution, and paraphrasing enrich lexical and semantic diversity.

Future research should further investigate how advanced augmentation strategies and multilingual transfer learning from related Austronesian languages, such as Malay, Javanese, and Tagalog, can improve Indonesian low-resource text generation. Exploring instruction-tuning and alignment techniques, including Reinforcement Learning from Human Feedback (RLHF) and contrastive prompt optimization, may further enhance factual consistency and stylistic flexibility. In addition, extending augmentation beyond lexical and syntactic variation toward semantic and contextual enrichment, together with larger-scale human evaluation frameworks, could contribute to the development of more robust, inclusive, and resource-efficient LLMs for underrepresented languages.

References

- [1] Y. Zhao, W. Zhang, G. Chen, K. Kawaguchi, and L. Bing, "How do large language models handle multilingualism?," In Proc. 38th Conference on Neural Information Processing Systems, 2024, pp. 1-24.
- [2] L. Qin, Q. Chen, Y. Zhou, Z. Chen, Y. Li, L. Liao, M. Li, W. Che, and P. S. Yu, "A survey of multilingual large language models," *Patterns*, vol. 6, no. 1, pp. 1-30, 2025.
- [3] Y. Tu, A. Xue, and F. Shi, "Blessing of multilinguality: A systematic analysis of multilingual in-context learning," In Proc. 63rd Annual Meeting of the Association for Computational Linguistics, 2025, pp. 6213-6248.
- [4] M. A. Ibrahim, Faisal, Z. D. Sulistiya, and T. S. Y. Winarto, "Prompt-based data augmentation with large language models for Indonesian gender-based hate speech detection," *Journal of Computer Science*, vol. 20, no. 8, pp. 819-826, 2024.
- [5] R. Kimera, D. N. Heo, D. N. Rim, and H. Choi, "Data augmentation with back translation for low resource languages: A case of english and Luganda," In Proc. 8th International Conference on Natural Language Processing and Information Retrieval, 2025, pp. 142-148.
- [6] F. I. Maulana, Y. Heryadi, G. P. Kusuma, and W. Budiharto, "Data augmentation English-Indonesia-Madurese parallel corpus dataset using neural machine translation," *Data in Brief*, vol. 62, no. 1, pp. 1-8, 2025.
- [7] P. N. Hadiwinoto and D. P. Lestari, "Data augmentation on spontaneous Indonesian automatic speech recognition using statistical machine translation," In Proc. International Conference on Information Technology and Digital Applications, 2019, pp. 1-8.
- [8] F. Muftie and M. Haris, "IndoBERT based data augmentation for Indonesian text classification," In Proc. International Conference on Information Technology Research and Innovation, 2023, pp. 128-132.
- [9] A. H. Nasution, A. Onan, Y. Murakami, W. Monika, and A. Hanafiah, "Benchmarking open-source large language models for sentiment and emotion classification in Indonesian tweets," *IEEE Access*, vol. 13, no. 1, pp. 94009-94025, 2025.
- [10] A. S. Wijaya and A. S. Girsang, "Augmented-based Indonesian abstractive text summarization using pre-trained model mT5," *International Journal of Engineering Trends and Technology*, vol. 71, no. 11, pp. 190-200, 2023.
- [11] B. R. Irnawan and R. Adi, "Improving Indonesian informal to formal style transfer via pre-training unlabelled augmented data," In Proc. 6th International Conference of Computer and Informatics Engineering, 2023, pp. 25-29.
- [12] KEMENDIKDASMEN, "Perpustakaan digital," Ministry of Primary and Secondary Education Republic of Indonesia, 2025. [Online]. Available: <https://pustaka-digital.kemendikdasmen.go.id/>. [Accessed: Sept. 2, 2025].
- [13] MYEDISI, "Buku Sekolah Elektronik (BSE)," My Edisi, 2021. [Online]. Available: <https://www.myedisi.com/bse>. [Accessed: Sept. 3, 2025].
- [14] S. Cahyawijaya, G. I. Winata, B. Wilie, K. Vincentio, X. Li, A. Kuncoro, S. Ruder, Z. Y. Lim, S. Bahar, M. Khodra, A. Purwarianti, and P. Fung, "IndoNLG: Benchmark and resources for evaluating Indonesian natural language generation," In Proc. Conference on Empirical Methods in Natural Language Processing, 2021, pp. 8875-8898.
- [15] B. Wilie, S. Cahyawijaya, and G. I. Winata, "IndoNLG-Indonesian natural language generation," github.com, 2024. [Online]. Available: <https://github.com/IndoNLP/indonlg>. [Accessed: Sept. 1, 2025].
- [16] INDODATASET, "Indonesian datasets: NLP datasets for Indonesian," github.com, 2023. [Online]. Available: https://github.com/Wikidepia/indonesian_dataset. [Accessed: Sept. 2, 2025].
- [17] INDONESIAGOID, "Portal informasi Indonesia," Indonesia Gov, 2025. [Online]. Available: <https://indonesia.go.id/>. [Accessed: Sept. 1, 2025].
- [18] KBBI, "Kamus Besar Bahasa Indonesia (KBBI) VI daring," Indonesian Dictionary, 2025. [Online]. Available: <https://kbbi.kemdikbud.go.id/>. [Accessed: Sept. 1, 2025].
- [19] M. Dwiastuti, "English-Indonesian neural machine translation for spoken language

- domains,” In Proc. 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, 2019, pp. 309-314.
- [20] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, “IndoBERT: Indonesian version of BERT model,” indolem.github.io, 2020. [Online]. Available: <https://indolem.github.io/Task/>. [Accessed: Sept. 2, 2025].
- [21] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, “Liputan6: A large-scale Indonesian dataset for text summarization,” In Proc. 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, 2020, pp. 598-608.
- [22] S. Cahyawijaya, G. I. Winata, B. Wilie, K. Vincentio, X. Li, A. Kuncoro, S. Ruder, Z. Y. Lim, S. Bahar, M. Khodra, A. Purwarianti, and P. Fung, “IndoGPT Model,” huggingface.co, 2021. [Online]. Available: <https://huggingface.co/indobenchmark/indogpt>. [Accessed: Sept. 2, 2025].
- [23] A. Onan and H. Alhumyani, “Knowledge-enhanced transformer graph summarization (KETGS): Integrating entity and discourse relations for advanced extractive text summarization,” *Mathematics*, vol. 12, no. 23, pp. 1-25, 2024.
- [24] A. A. Syahidi and K. Kiyokawa, “Automatic text generation in Banjar language using GPT-4 for low-resource language preservation,” In Proc. 7th IEEE Symposium on Computers and Informatics, 2025, pp. 238-243.
- [25] A. Gupta, A. Rastogi, H. Malhotra, and K. Rangarajan, “Comparative evaluation of large language models for translating radiology reports into Hindi,” *Indian Journal of Radiology and Imaging*, vol. 35, no. 1, pp. 88-96, 2024.
- [26] A. A. Citarella, M. Barbella, M. G. Ciobanu, F. D. Marco, L. D. Biasi, and G. Tortora, “Assessing the effectiveness of ROUGE as unbiased metric in extractive vs. abstractive summarization techniques,” *Journal of Computational Science*, vol. 87, no. 102571, pp. 1-17, 2025.
- [27] E. Oro, F. M. Granata, and M. Ruffolo, “A comprehensive evaluation of embedding models and LLMs for IR and QA across English and Italian,” *Big Data and Cognitive Computing*, vol. 9, no. 141, pp. 1-41, 2025.
- [28] A. A. Syahidi, K. Kiyokawa, and S. Nuchitprasitchai, “A fine-tuned GPT-4-based question answering system for e-government services using a custom-built dataset,” In Proc. 7th IEEE Symposium on Computers and Informatics, 2025, pp. 232-237.
- [29] M. Mahyoub, Y. Wang, and M. T. Khasawneh, “GPT-4o in radiology: In-context learning based automatic generation of radiology impressions,” *Natural Language Processing Journal*, vol. 11, no. 100145, pp. 1-10, 2025.
- [30] P. Netisopakul and U. Taoto, “Comparison of evaluation metrics for short story generation,” *IEEE Access*, vol. 11, no. 1, pp. 140253-140269, 2023.
- [31] P. S. García-Montero, P. Vizcaíno, I. G. Reyes-Chacón, and M. E. Morochó-Cayamcela, “AI for all: Reducing perplexity and boosting accuracy in normative texts with fine-tuned LLMs and RAG,” *IEEE Access*, vol. 13, no. 1, pp. 179759-179775, 2025.
- [32] L. Fang, Y. Wang, Z. Liu, C. Zhang, S. Jegelka, J. Gao, B. Ding, and Y. Wang, “What is wrong with perplexity for long-context language modeling?,” In Proc. 13th International Conference on Learning Representations, 2025, pp. 1-23.
- [33] J. Xu, H. Zhang, Y. Yang, L. Yang, Z. Cheng, J. Lyu, B. Liu, X. Zhou, A. Bacchelli, Y. K. Chiam, and T. K. Chiew, “One size does not fit all: Investigating efficacy of perplexity in detecting LLM-generated code,” *ACM Transactions on Software Engineering and Methodology*, vol. 1, no. 1, pp. 1-33, 2025.
- [34] A. Aziz, M. A. Hossain, and A. N. Chy, “CSECU-DSG at SemEval-2022 task 3: Investigating the taxonomic relationship between two arguments using fusion of multilingual transformer models,” In Proc. 16th International Workshop on Semantic Evaluation, 2022, pp. 255-259.
- [35] E. A. Chi, J. Hewitt, and C. D. Manning, “Finding universal grammatical relations in multilingual BERT,” In Proc. 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 5564-5577.
- [36] S. Liu, S. Sabour, Y. Zheng, P. Ke, X. Zhu, and M. Huang, “Rethinking and refining the distinct metric,” In Proc. 60th Annual Meeting of the Association for Computational Linguistics, 2022, pp. 762-770.
- [37] A. Shypula, S. Li, B. Zhang, V. Padmakumar, K. Yin, and O. Bastani, “Evaluating the diversity and quality of LLM generated content,” In Proc. International Conference on Learning Representations, 2025, pp. 1-18.
- [38] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, “A diversity-promoting objective function for neural conversation models,” In Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 110-119.
- [39] J. Kim and W. Jung, “CARE: A framework for correcting numerical hallucinations in LLM-generated financial texts,” In Proc. IEEE

- Conference on Artificial Intelligence, 2025, pp. 69-74.
- [40] M. Y. Mohammed, S. A. Ali, S. K. Ali, A. A. Majeed, and E. H. Mohamed, "Aftina: Enhancing stability and preventing hallucination in AI based islamic fatwa generation using LLMs and RAG," *Neural Computing and Applications*, vol. 37, no. 1, pp. 20957-20982, 2025.
- [41] I. Jahan, M. T. R. Laskar, C. Peng, and J. X. Huang, "A comprehensive evaluation of large Language models on benchmark biomedical text processing tasks," *Computers in Biology and Medicine*, vol. 171, no. 108189, pp. 1-23, 2024.
- [42] T. Y. C. Tam, S. Sivarajkumar, S. Kapoor, A. V. Stolyar, K. Polanska, K. R. McCarthy, H. Osterhoudt, X. Wu, S. Visweswaran, S. Fu, P. Mathur, G. E. Cacciamani, C. Sun, Y. Peng, and Y. Wang, "A framework for human evaluation of large language models in healthcare derived from literature review," *npj - Digital Medicine*, vol. 7, no. 258, pp. 1-20, 2024.
- [43] Q. Li, L. Cui, L. Kong, and W. Bi, "Exploring the reliability of large language models as customized evaluators for diverse NLP tasks," In Proc. 31st International Conference on Computational Linguistics, 2025, pp. 10325-10344.
- [44] Z. Li, X. Xu, T. Shen, C. Xu, J-C Gu, Y. Lai, C. Tao, and S. Ma, "Leveraging large language models for NLG evaluation: Advances and challenges," In Proc. Conference on Empirical Methods in Natural Language Processing, 2024, pp. 16028-16045.
- [45] G. Tevet and J. Berant, "Evaluating the evaluation of diversity in natural language generation," In Proc. 16th Conference of the European Chapter of the Association for Computational Linguistics, 2021, pp. 326-346.
- [46] B. Więckowska, K. B. Kubiak, P. Jóźwiak, W. Moryson, and B. Stawińska-Witoszyńska, "Cohen's kappa coefficient as a measure to assess classification improvement following the addition of a new marker to a regression model," *International Journal of Environmental Research and Public Health*, vol. 19, no. 10213, pp. 1-15, 2022.
- [47] A. S. Thakur, K. Choudhary, V. S. Ramayapally, S. Vaidyanathan, and D. Hupkes, "Judging the judges: Evaluating alignment and vulnerabilities in LLMs-as-judges," In Proc. Fourth Workshop on Generation, Evaluation and Metrics, 2025, pp. 404-430.
- [48] V. Hackl, A. E. Müller, M. Granitzer, and M. Sailer M, "Is GPT-4 a reliable rater? Evaluating consistency in GPT-4's text ratings," *Frontiers in Education*, vol. 8, no. 1272229, pp. 1-8, 2023.
- [49] D-W. Zhang, M. Boey, Y. Y. Tan, and A. H. S. Jia, "Evaluating large language models for criterion-based grading from agreement to consistency," *npj - Science of Learning*, vol. 9, no. 79, pp. 1-4, 2024.

Explainable Analysis of Security Information and Event Management(SIEM) Event Logs for Event Profiling and Preliminary Triage

Pranisa Israsena

pranisa@tni.ac.th

Faculty of Information Technology

Thai-Nichi Institute of Technology, Bangkok, Thailand

*Amonpan Chomklin**

amonpan@tni.ac.th

Faculty of Information Technology

Thai-Nichi Institute of Technology, Bangkok, Thailand

Lalita Na Nongkhai

lalita@tni.ac.th

Faculty of Information Technology

Thai-Nichi Institute of Technology, Bangkok, Thailand

Received: February 16, 2026; Revised: June 20, 2026; Accepted: June 24, 2026; Published: June 30, 2026

ABSTRACT - Security Information and Event Management (SIEM) systems generate large volumes of heterogeneous event logs, imposing substantial investigation workloads on Security Operations Centers (SOCs). This study proposes an explainable framework for SIEM event analysis and preliminary triage that emphasizes interpretability and operational applicability without requiring labeled training data. The framework integrates descriptive and temporal analysis, structural association analysis using Spearman correlation, K-means clustering with TF-IDF features to construct interpretable event profiles, and explainable triage scoring based on standard SIEM indicators, including severity, priority, confidence, and aggregated counts. Experimental evaluation using operational SIEM logs (n=3,490 events) demonstrates 93.8% workload reduction while retaining 98.97% of high-severity events. External validation with labeled attack data (489 SQL injection attempts and 301 brute-force attacks) prioritized 98.2% of known attacks (776/790) with a 1.8% false-negative rate, outperforming severity-only (78.5%) and severity+priority (89.1%) baselines. The results confirm that unsupervised event profiling combined with explainable triage scoring can prioritize security-relevant events while filtering routine monitoring noise. The framework provides a practical, transparent analytical baseline for SOC environments under realistic operational constraints.

KEY WORDS -- SIEM, Security event analysis, Event clustering, Explainable analytics, Event triage, SOC operations

1. Introduction

Modern organizations increasingly rely on information systems and computer networks to support critical operations, which has led to a corresponding rise in cybersecurity risks arising from both external attacks and internal anomalies. Examples of such events include port scanning activities, repeated authentication attempts, unauthorized service access, as well as incidents caused by misconfigurations or software vulnerabilities. When these events occur, organizations must be capable of timely monitoring, detection, and response in order to mitigate potential impacts on systems and sensitive data. One of the widely adopted infrastructures for managing security-related information is the Security Information and Event Management (SIEM) system. SIEM systems collect event logs from diverse sources, such as network devices, security appliances, endpoints, and applications, and subsequently store, analyze, and visualize these data to support security monitoring and incident response [1]. Despite their ability to centralize security event data, SIEM systems face a fundamental operational challenge: the sheer volume and diversity of events generated on a daily basis, which significantly increases investigation workload and noise within the event stream [2]. A large proportion of events recorded in SIEM logs correspond to routine monitoring activities, such as recurring alerts from similar behavior patterns, automated scanning activities, or low-severity events that occur frequently. In contrast, events that are truly critical and require immediate response typically constitute only a small fraction of the total volume. This imbalance makes event prioritization and escalation highly time-consuming and heavily dependent on analyst expertise, and it directly contributes to the problem of alert fatigue in Security Operations Centers (SOCs) [2]. If event prioritization is not performed effectively, critical security incidents may be overlooked or response actions may be delayed. These challenges highlight that, while SIEM event logs are valuable for security monitoring, they require analytical approaches that can summarize global event characteristics and structure large volumes of data into interpretable forms. This need is particularly pronounced in organizational contexts where explainable and reusable analytical frameworks are preferred over opaque decision-making processes, consistent with established findings in explainable artificial intelligence and human-centered security analytics [3]. Common analytical approaches in security event analysis include descriptive analysis to summarize event trends and frequencies, structural association analysis of key indicators such as severity levels or correlated event counts, and clustering techniques to construct event profiles that reduce data

complexity to a manageable and interpretable level. In particular, the integration of structured indicators with text-derived representations, such as term frequency-inverse document frequency (TF-IDF), combined with clustering algorithms (e.g., k-means), has been widely adopted for summarizing large-scale security logs and related textual event records [5], [6].

Although machine learning techniques have been increasingly explored to automate event prioritization and detection, their deployment in real-world security operations remains constrained by several factors. These include the scarcity of high-quality labeled data, the inherent mismatch between closed-world learning assumptions and operational security environments, and the limited robustness of trained models when exposed to unseen behaviors [7]. Additional challenges include the limited explainability of many machine learning models, the presence of concept drift caused by evolving system configurations and attacker strategies, and the continuous maintenance burden associated with retraining and validating models in dynamic environments [3], [4]. Consequently, many organizations still require analytical and triage frameworks that can be applied immediately, remain interpretable to human analysts, and serve as a foundation for more advanced automation in later stages [2].

Motivated by these considerations, this study proposes a structured approach for analyzing SIEM event logs that begins with controlled data preparation to preserve the characteristics of real operational events. The methodology then applies descriptive and temporal analysis to explain event occurrence patterns, followed by structural association analysis to examine relationships among key SIEM indicators. Event clustering is subsequently employed to construct meaningful event profiles, and a preliminary triage framework is designed using existing SIEM indicators to compute interpretable triage scores. The framework is evaluated from an operational perspective, focusing on its ability to reduce investigation workload while retaining high-severity events.

The objectives of this study are threefold: (1) to analyze and summarize SIEM event log characteristics in temporal and structural dimensions, (2) to construct interpretable event profiles through clustering that integrates structural and textual features, and (3) to propose and evaluate an explainable preliminary triage framework in terms of investigation workload reduction and retention of high-severity events. The scope of this study focuses on SIEM event logs collected from a real operational

environment and emphasizes explainable analysis and preliminary triage rather than supervised prediction or prioritization performance evaluation. This scope is intended to establish a clear analytical baseline that can be systematically extended in future research.

The remainder of this paper is organized as follows. Section 2 reviews related work on SIEM analytics and event clustering. Section 3 describes the research methodology and data analysis procedures. Section 4 presents the experimental results, followed by discussion and practical implications in Section 5. Finally, Section 6 concludes the paper and outlines directions for future research.

2. Related Work

Security Information and Event Management (SIEM) systems constitute a fundamental infrastructure for monitoring and managing cybersecurity events by collecting and processing event logs from heterogeneous sources, including network devices, servers, and application systems. As discussed in Section 1, SIEM platforms enable centralized visibility over security-relevant activities; however, converting large volumes of heterogeneous and noisy log data into actionable decision support remains a major operational challenge [1], [2]. Standardization-oriented references, such as NIST Special Publication 800-92, emphasize best practices for log management—including time synchronization, secure storage, integrity preservation, and retention policies—to ensure that logs can be reliably used for incident investigation and forensic analysis [8]. While such standards establish an essential operational baseline, they do not prescribe data-driven analytical methods or systematic mechanisms for event summarization, prioritization, or triage.

From a data analytics perspective, clustering has been widely studied as a means to reduce the complexity of large-scale security logs and to identify recurring patterns. Landauer et al. provide a comprehensive survey of log clustering approaches in cybersecurity applications, covering distance-based, density-based, and hierarchical techniques applied to system and security logs [9]. Their survey highlights clustering as a practical unsupervised strategy for structuring massive log datasets in the absence of labels. Nevertheless, it also identifies critical limitations for operational adoption: clustering results are often difficult to interpret for analysts, and feature selection strongly influences the semantic meaning and usefulness of the resulting clusters [9]. These findings suggest that clustering alone is insufficient

for supporting analyst workflows unless complemented by interpretable summaries or profiles.

Several applied studies have explored machine learning techniques for SIEM log analysis, particularly the integration of unstructured textual log fields with structured indicators. A common approach involves transforming textual attributes (e.g., event names or descriptions) into vector representations using term frequency–inverse document frequency (TF-IDF) weighting [6], followed by clustering algorithms such as K-means [12]. Tendikov et al. demonstrate that such SIEM-oriented analytics pipelines can efficiently group events with similar patterns at relatively low computational cost, making them attractive as baseline analytical methods in operational environments [5]. However, these studies primarily focus on algorithmic feasibility or grouping performance and pay limited attention to how the resulting clusters can be operationalized as interpretable event profiles that directly support investigation and triage activities in Security Operations Centers (SOCs) [5], [9].

Beyond event grouping, alert prioritization and triage have emerged as critical research areas in response to alert fatigue and excessive investigation workloads, as highlighted in Section 1 [2]. Wang et al. propose AlertPro, a context-aware alert prioritization framework that leverages alert sequences, contextual signals, and feedback mechanisms to improve triage efficiency in SOC [10]. Complementing this line of work, Jalalvand et al. provide a systematic survey of alert prioritization criteria and methods, revealing substantial diversity in prioritization strategies and emphasizing the strong influence of data availability, contextual enrichment, and organizational constraints on practical deployment [11]. While these approaches demonstrate the potential benefits of prioritization, they often assume the availability of rich contextual or historical information that may not be consistently present across different SIEM deployments, thereby limiting their general applicability.

To avoid citation inconsistency and clarify the methodological positioning of this study, Table 1 summarizes representative prior work using only studies that are included in the reference list. The comparison focuses on log management, SIEM analytics, log clustering, alert prioritization, and explainable security analytics. These studies collectively show that existing approaches often address event collection, clustering, prioritization, or explainability as separate components, whereas this study integrates unsupervised event profiling with an explainable triage scoring mechanism using commonly available SIEM indicators.

Table 1. Summary of related work in SIEM analytics and security event triage

Study	Main focus	Method / Approach	Key contribution	Limitation / Gap related to this study
González-Granadillo et al. [1]	SIEM systems	SIEM analysis and trend review	Summarizes SIEM usage, trends, and challenges in security monitoring	Does not provide a lightweight event profiling and triage framework
NIST SP 800-92 [8]	Log management	Operational guideline	Provides best practices for log collection, storage, integrity, and retention	Focuses on log management practice rather than data-driven event profiling
Landauer et al. [9]	Log clustering	Survey of clustering approaches	Shows that clustering is useful for reducing log complexity in cybersecurity	Highlights interpretability and feature selection challenges
Tendikov et al. [5]	SIEM analytics	Machine-learning-based SIEM data acquisition and analysis	Demonstrates practical analytics pipelines for SIEM data	Limited connection between clustering outputs and SOC triage decisions
Wang et al. [10]	Alert prioritization	Context-aware alert prioritization	Addresses alert fatigue through prioritization of security alerts	Requires richer contextual and historical information
Jalalvand et al. [11]	SOC alert prioritization	Systematic survey	Summarizes prioritization criteria and methods in SOC environments	Shows lack of a unified lightweight baseline for heterogeneous SIEM settings
This study	Explainable SIEM event profiling and triage	TF-IDF + K-means event profiling + fixed-weight triage scoring	Integrates unsupervised event profiles with explainable preliminary triage	Limited to a single-day primary dataset and two labeled attack categories

As summarized in Table 1, prior studies provide important foundations for SIEM monitoring, log management, clustering, and alert prioritization. However, many approaches either focus on operational log management, algorithmic clustering, or context-heavy prioritization. This study addresses the gap by proposing a lightweight and explainable framework that connects unsupervised event profiling with preliminary triage scoring without requiring labeled training data or complex contextual enrichment.

3. Methodology

This study aims to develop a structured and explainable approach for analyzing Security Information and Event Management (SIEM) event logs in order to support the understanding of security event characteristics and preliminary event triage in real operational settings. The methodological design emphasizes interpretability, reliance on commonly available SIEM indicators, and systematic reproducibility, enabling the proposed approach to be applied and extended across diverse organizational contexts

3.1 Conceptual Research Framework

The conceptual framework is organized as a sequential pipeline from raw SIEM event logs to decision-support outputs. It consists of data preparation, descriptive analysis, exploratory analysis, structural association analysis, event profiling via clustering, preliminary triage scoring, and validation with labeled attack data. Descriptive analysis summarizes event volume and temporal distribution, whereas exploratory analysis examines contextual event composition, including event categories, source countries, destination ports, device or product types, and textual characteristics. Raw SIEM logs are first converted into an analytical dataset through timestamp parsing, data type conversion, missing-value handling, preservation of potentially anomalous values, and consolidation of textual fields. The transformed data are then represented using standardized SIEM indicators and TF-IDF textual features for clustering and triage scoring. This framework is intended to support SOC analysts by transforming large volumes of raw events into interpretable profiles and prioritized investigation outputs, rather than replacing analyst judgment.

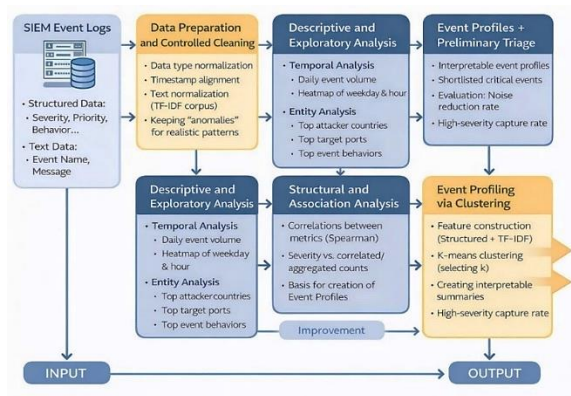


Figure 1. Conceptual framework and data construction pipeline for explainable SIEM event profiling and preliminary triage.

As shown in Figure 1, each analytical stage produces outputs that support the next stage, enabling transparent decision support for SOC analysts. The framework transitions from data preparation to event profiling and triage scoring, while remaining grounded in standard SIEM indicators and interpretable analytical outputs.

3.2 Research Questions and Hypotheses

To guide the analysis in a systematic and reproducible manner, this study defines research questions and hypotheses that are examined using empirical evidence, with an emphasis on structural characteristics and operational applicability rather than predictive performance. This methodological choice reflects the absence of ground-truth event labels in the primary SIEM dataset and aligns with the exploratory and descriptive nature of the proposed framework. The research questions and hypotheses are explicitly aligned with the sequential stages of the conceptual framework described in Section 3.1, encompassing data characterization, structural association analysis, event profiling via clustering, evaluation of the preliminary triage framework, and validation with labeled attack data. To enhance clarity, traceability, and consistency between analytical objectives and methodological design, the research questions and their corresponding hypotheses are summarized in Table 2, together with their analytical focus and associated framework stages.

Table 2. Research questions and hypotheses aligned with the conceptual framework

RQ/Hypothesis	Research Question	Hypothesis	Framework Stage
RQ1 / H1	What are the temporal characteristics and overall patterns of SIEM event logs?	SIEM event logs exhibit interpretable temporal patterns with observable temporal concentrations and event composition characteristics.	Descriptive and temporal analysis
RQ2 / H2	What structural associations exist among commonly used SIEM indicators?	Event severity levels and correlated or aggregated event counts exhibit meaningful associations from an operational perspective.	Structural association analysis
RQ3 / H3	Can event clustering produce interpretable event profiles with practical significance?	Integration of structural and textual SIEM features enables the construction of distinct and interpretable event profiles with practical significance.	Event profiling via clustering
RQ4 / H4	Does the preliminary triage framework achieve workload reduction while retaining high-severity events?	An explainable triage framework based on standard SIEM indicators achieves a favorable trade-off between efficiency and risk.	Preliminary triage evaluation
RQ5 / H5	Can the framework effectively prioritize genuine security incidents when validated with labeled attack data?	The framework achieves high prioritization rates for labeled attacks while maintaining workload reduction, demonstrating effective differentiation between genuine security incidents and routine events.	Validation with labeled attack data

3.3 Data Source and Scope of the Study

The primary dataset consists of 3,490 SIEM event logs collected from a real operational environment within

a single-day observation window. The dataset contains structured SIEM indicators, including severity, priority, confidence or relevance score, device or product information, event categories, and

correlated or aggregated event counts, as well as unstructured textual fields such as event names and descriptions. The primary dataset is unlabeled and is therefore used for descriptive analysis, exploratory analysis, structural association analysis, unsupervised event profiling, and preliminary triage evaluation. The raw SIEM logs were converted into an analytical dataset through a controlled preparation pipeline. This included timestamp parsing, numerical type conversion, missing-value handling, preservation of potentially anomalous values, and consolidation of textual fields for TF-IDF feature extraction. Descriptive analysis was then used to summarize event volume and temporal distribution, whereas exploratory analysis examined event composition, including behavioral categories, source countries, destination ports, device or product types, and recurring textual characteristics.

Supplementary labeled attack data, consisting of SQL injection attempts and brute-force authentication attacks, were used only for external validation of the triage framework. These labeled events were not used to train the clustering model, optimize triage weights, or adjust the triage threshold. This design preserves the label-free nature of the proposed framework while allowing its prioritization capability to be assessed under realistic SOC conditions.

3.4 Structural and Association Analysis

Given the skewed distributions and the presence of outliers commonly observed in SIEM event data, this study employs non-parametric statistical analysis to examine structural relationships among key SIEM indicators. In particular, associations between event severity levels and correlated or aggregated event counts are analyzed using Spearman's rank correlation coefficient (ρ), which is defined as:

$$\rho = 1 - (6\sum d_i^2) / (n(n^2 - 1)) \quad (1)$$

where d_i denotes the difference between the ranks of paired observations for event i , and n represents the total number of events. Spearman's rank correlation is suitable for ordinal variables such as severity levels and does not require assumptions of normality or linearity, making it appropriate for analyzing SIEM data characterized by non-Gaussian distributions and extreme values [13].

The objective of this analysis is not to infer causal relationships, but to provide empirical justification for the selection of structural indicators used in subsequent analytical stages. Specifically, the observed associations support data-driven decisions regarding which indicators are integrated into event clustering and preliminary triage scoring. In this sense, the structural association analysis serves as a

methodological bridge between descriptive characterization of SIEM events and the construction of interpretable event profiles in later stages of the framework.

3.5 Event Profiling via Clustering

Event clustering was used to construct interpretable event profiles by combining structured SIEM indicators with textual features extracted from event names and descriptions. Textual fields were represented using TF-IDF [6], while structured indicators included severity, priority, confidence or relevance score, and correlated or aggregated event count. Numerical indicators were standardized before concatenation with TF-IDF features to reduce scale bias. The number of clusters k was determined using the silhouette coefficient, which provides an internal validation measure of clustering structure and supports data-driven selection of an appropriate cluster granularity. For an event i , the silhouette coefficient $s(i)$ is defined as:

$$s(i) = (b(i) - a(i)) / \max\{a(i), b(i)\} \quad (2)$$

where $a(i)$ denotes the average distance between event i and all other events within the same cluster, and $b(i)$ represents the minimum average distance between event i and events in the nearest neighboring cluster. Values of $s(i)$ range from -1 to 1 , with higher values indicating better separation between clusters [14].

K-means was adopted as a transparent baseline clustering method [12]. The number of clusters was selected using silhouette-guided assessment [14] together with interpretability considerations. The final configuration used $k = 10$, producing ten event profiles labeled Cluster 0 to Cluster 9. A fixed random seed and multiple centroid initializations were used to improve reproducibility and reduce sensitivity to initialization.

The resulting clusters were interpreted as event profiles rather than attack-type labels. Each profile was summarized using cluster size, dominant behavior, dominant device or product, average priority, average confidence, and representative textual characteristics. These profiles serve as interpretable abstractions of recurring SIEM event patterns and provide inputs to the preliminary triage framework.

3.6 Preliminary Triage Framework and Evaluation

The preliminary triage framework is designed to reduce investigation workload while retaining critical security events, in alignment with the operational

objectives of Security Operations Centers (SOCs). To this end, a triage score is computed using standard SIEM indicators, without optimizing parameter weights, in order to preserve transparency and interpretability. For each event i , the triage score is defined as:

$$\text{Triage Score}_i = w_1 \cdot S_i + w_2 \cdot P_i + w_3 \cdot C_i + w_4 \cdot \log(1 + A_i) \quad (3)$$

In Equation (3), S_i denotes the event severity level, P_i represents the event priority, C_i corresponds to a normalized confidence or relevance score provided by the SIEM system, and A_i is the number of correlated or aggregated events associated with event i . Reported immediately here for structural transparency, the fixed weights are defined as $w_1 = 1.0$, $w_2 = 0.8$, $w_3 = 0.5$, and $w_4 = 0.3$. These values were not learned from the dataset and were not optimized to maximize detection performance. Instead, they were defined as an analyst-oriented prioritization policy to preserve transparency, reproducibility, and operational interpretability. Aggregated event count receives the lowest direct weight because frequent correlation can indicate coordinated or repeated activity, but raw count values may be highly skewed in SIEM environments.

A logarithmic transformation, $\log(1 + A_i)$, was therefore applied to the aggregated event count term to reduce the influence of extreme values while preserving monotonicity. This design allows events with larger correlated counts to contribute meaningfully to the triage score without allowing high-volume routine events to dominate the prioritization process. The scoring framework is therefore intended as an explainable triage heuristic rather than a learned predictive model. This fixed-weight design avoids dataset-specific parameter tuning and supports direct interpretation of the triage score. For example, an event receives a higher score when it has higher severity, higher operational priority, stronger confidence, or a larger number of correlated events. Because each component is based on commonly available SIEM indicators, the resulting score can be inspected and explained to analysts without requiring complex model interpretation.

Evaluation of the preliminary triage framework focuses on two operationally relevant criteria. First, workload reduction is measured by the proportion of events excluded from manual investigation when applying the triage score threshold. Second, retention of critical events is assessed by examining the presence and proportion of high-severity events within the prioritized subset. Together, these criteria reflect the trade-off between efficiency and risk that SOC analysts routinely manage. To enhance reliability and traceability, all intermediate datasets, analytical outputs, and evaluation results are preserved in reusable file formats. Algorithmic parameters, scoring weights, and initialization settings are explicitly documented, enabling systematic reproduction and

comparative analysis in future studies. This approach promotes transparency and repeatability while avoiding reliance on dataset-specific parameter optimization.

3.6.1 Validation Procedure with Labeled Attack Data

To evaluate the triage framework's prioritization capability, the supplementary labeled attack data were processed through the same analytical pipeline used for the primary SIEM events. The validation data were used only for external assessment and were not used to retrain the clustering model, update cluster centroids, optimize triage weights, or adjust the triage threshold. The validation procedure consisted of three steps. First, labeled SQL injection and brute-force authentication events were transformed into the same analytical schema as the primary SIEM dataset, including severity, priority, confidence, textual descriptions, and aggregated event counts. Second, the transformed events were assigned to the existing ten event profiles generated by the primary K-means model without retraining or updating centroids. Third, triage scores were computed using Equation (3) with the same fixed weights defined in Section 3.6.

Validation performance was assessed using attack prioritization rate, false-negative rate, and category-specific prioritization rate. This procedure reflects a realistic SOC scenario in which newly observed events are mapped to existing event profiles and prioritized using a transparent triage policy rather than a retrained supervised classification model.

4. Results

The experimental results are presented in accordance with the defined research questions (RQ1–RQ5) and their corresponding hypotheses, in order to demonstrate consistency between the conceptual framework, the methodological design described in Section 3, and the empirical evidence obtained from real SIEM event logs. Each subsection reports results derived directly from the analytical procedures outlined in the Methodology and is supported by figures and tables generated from the processed dataset.

4.1 Temporal Characteristics and Overall Event Patterns (RQ1 / H1)

To address RQ1, the analysis begins with an examination of the temporal characteristics of SIEM event logs, following the descriptive and temporal analysis stage. This stage aims to establish a global understanding of event occurrence patterns within the available observation window and to assess whether the data exhibit interpretable temporal structures

suitable for further analysis. Figure 2 illustrates the daily event volume recorded by the SIEM system during the observed period. As shown in the figure, a total of 3,490 events were recorded on January 5, 2026, indicating that the system continuously generated security-related events at a high frequency throughout the day. Because the dataset represents a single-day snapshot, the analysis in this subsection focuses on descriptive characterization of event volume within the available temporal scope, rather than on identifying long-term trends or seasonal effects. This scope is consistent with the data source constraints described in Section 3.3 and avoids overinterpretation beyond the temporal limits of the dataset.

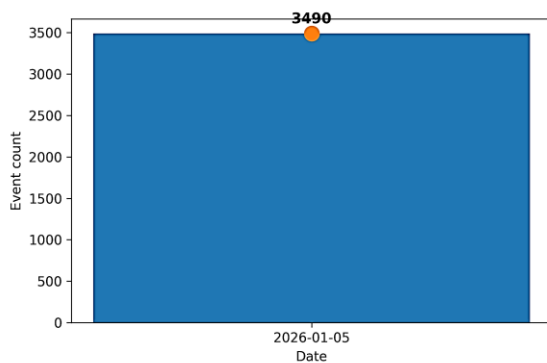


Figure 2. Daily event volume from SIEM logs (single-day snapshot).

To further examine intraday temporal patterns, event frequencies were analyzed across weekdays and hours, as described in the temporal analysis procedure. Figure 3 presents a heatmap of event frequency by weekday and hour, illustrating the distribution of events within the observed day. Although the temporal scope is limited, the results reveal non-uniform event distributions, with events concentrated during specific hours rather than evenly spread across the day. Such temporal concentration reflects typical SIEM operational behavior, where event generation is influenced by system usage patterns, automated monitoring processes, and scheduled activities within the monitored environment.

These observations support Hypothesis H1, which posits that SIEM event logs exhibit interpretable temporal patterns with observable temporal concentrations. Importantly, the identified temporal structures provide contextual grounding for subsequent analytical stages, including structural association analysis (Section 4.2) and event profiling via clustering (Section 4.3). By confirming the presence of meaningful temporal characteristics at the descriptive level, this subsection establishes that the dataset is suitable for deeper structural and profiling

analyses, as envisioned in the conceptual framework (Figure 1).

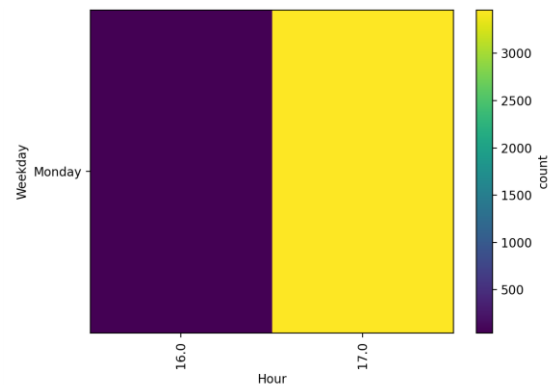


Figure 3. Heatmap of event frequency by day and hour.

4.2 Structural Associations Among SIEM Indicators (RQ2 / H2)

To address RQ2, this subsection examines structural associations among commonly used SIEM indicators, following the non-parametric association analysis. Specifically, the relationship between event severity and the number of correlated or aggregated events is analyzed using Spearman's rank correlation coefficient, which is suitable for ordinal variables and skewed SIEM data distributions.

The Spearman correlation analysis showed a positive association between event severity and correlated or aggregated event counts ($n = 3,490$). This result indicates that higher-severity events tended to be associated with a larger number of related events. The correlation result is not interpreted as evidence of causality; rather, it provides empirical justification for including aggregated event count together with severity in the subsequent clustering and preliminary triage scoring stages.

4.3 Event Clustering and Event Profile Construction (RQ3 / H3)

To address RQ3, this subsection reports the results of event clustering performed using the integration of structural indicators and textual features. The objective of this analysis is not to optimize clustering performance, but to construct interpretable event profiles that summarize recurring SIEM event patterns in a form suitable for operational use. The clustering results show that SIEM events can be partitioned into multiple groups with distinct structural and behavioral characteristics.

The distribution of events across clusters is illustrated in Figure 4, which reveals a highly imbalanced cluster size distribution. A small number

of clusters contain the majority of events, while several clusters are substantially smaller. This pattern is characteristic of real-world SIEM environments, where routine monitoring activities generate large volumes of repetitive, low-severity events, whereas potentially significant or specialized events occur less frequently. Specifically, as shown in Figure 4, the largest cluster (Cluster 0) dominates the dataset, indicating a concentration of common and repetitive event types. Additional medium-sized clusters (e.g., Clusters 2 and 6) capture secondary patterns of recurring behavior, while multiple small clusters

represent rare or distinctive event structures. This separation suggests that clustering effectively differentiates between high-frequency operational noise and lower-frequency events with potentially higher analytical relevance. To facilitate operational interpretation, the characteristics of each cluster are summarized as event profiles, presented in Table 3. Each profile aggregates key attributes of the corresponding cluster, including the number of events, dominant behavioral category, dominant device or product, average priority, and average model confidence.

Table 3. Summary of event profiles derived from clustering analysis.

Cluster	Number of Events	Dominant Behavior	Dominant Device/Product	Avg. Priority	Avg. Model Confidence
0	1850	Unknown	OBS	3.00	0.00
2	555	Unknown	OBS	3.00	4.00
6	426	/Access/Stop	Unix	3.02	0.00
7	264	Unknown	Threat inSights	3.00	0.00
1	195	Unknown	NAXSI	7.00	3.96
3	137	/Communicate/Query	Apache	3.00	0.00
9	49	Unknown	WebServer	2.61	0.00
4	23	Unknown	ArcSight	4.52	0.00
8	2	/Execute/Response	ArcSight	5.00	0.00
5	1	/Execute/Response	ArcSight	5.00	0.00

As shown in Table 3, clusters with large event counts (e.g., Clusters 0 and 2) are dominated by generic or “Unknown” behaviors with moderate average priority, reflecting routine SIEM monitoring noise. In contrast, smaller clusters (e.g., Clusters 1, 4, 8, and 5) exhibit higher average priority values and are associated with specific devices or products, such as NAXSI or ArcSight. Although these clusters contain relatively few events, their distinct structural characteristics highlight their potential operational relevance, particularly in the context of event prioritization. Overall, the combined evidence from Figure 4 and Table 3 demonstrates that event clustering effectively reduces the complexity of large-scale SIEM data by organizing events into a manageable number of interpretable profiles. These profiles provide structured representations of event behaviors that support analyst sense-making and directly inform the preliminary triage framework presented in Section 4.4. Accordingly, the results in this subsection support Hypothesis H3, confirming that the integration of structural and textual SIEM features enables the construction of distinct and interpretable event profiles with practical significance.

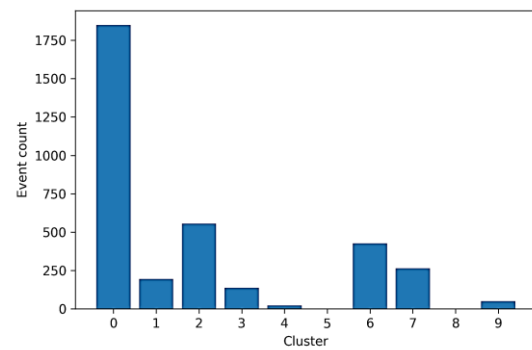


Figure 4. Distribution of events across clusters.

4.4 Evaluation of the Preliminary Triage Framework (RQ4 / H4)

To address RQ4, the preliminary triage framework is evaluated using the operational criteria defined in Section 3.6, namely investigation workload reduction and retention of critical events. The evaluation applies the triage score to rank all SIEM events and selects a prioritized subset for analyst investigation. Figure 5 illustrates the effect of the triage mechanism on investigation workload. By selecting the top 216 events based on the triage score—corresponding to 6.17% of the total event volume—the framework excludes approximately 93.83% of events from manual investigation. This substantial reduction demonstrates that the proposed indicator-based

scoring mechanism effectively filters routine and low-priority events, yielding a manageable investigation set without requiring supervised learning or dataset-specific parameter optimization.

In addition to workload reduction, the framework is evaluated in terms of its ability to retain critical events. As shown in Figure 6, of the 195 high-severity events present in the dataset, 193 events (98.97%) are retained within the shortlisted subset after triage. This result indicates that the aggressive reduction in investigation volume does not disproportionately exclude high-severity events.

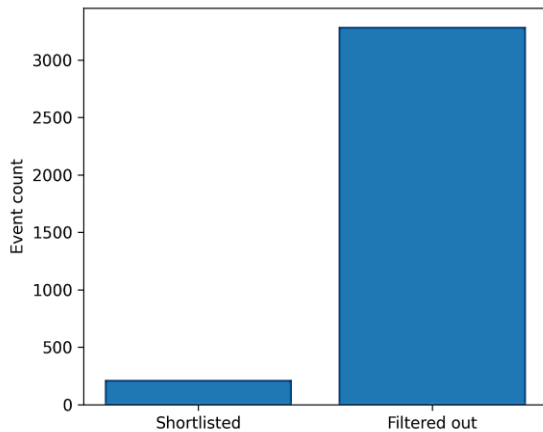


Figure 5. Effect of the triage mechanism on investigation workload.

Taken together, these results support Hypothesis H4, confirming that an explainable triage framework based on standard SIEM indicators can achieve a favorable trade-off between efficiency and risk. Importantly, the framework operates without supervised learning, dataset-specific parameter optimization, or complex contextual enrichment, aligning with the design objectives outlined in Section 3.6. The high retention rate of critical events, combined with substantial workload reduction, demonstrates the practical utility of the proposed approach as a baseline decision-support mechanism in operational SIEM environments.

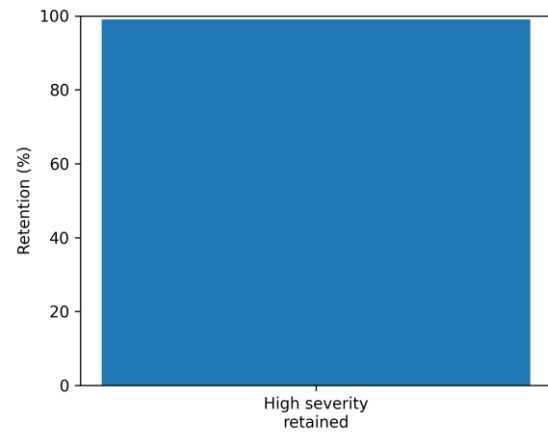


Figure 6. Retention of high-severity events under the preliminary triage framework.

4.5 Validation with Labeled Attack Data (RQ5 / H5)

To externally validate the triage framework, supplementary labeled attack data were processed using the same analytical pipeline as the primary SIEM dataset. The labeled data included 489 SQL injection attempts and 301 brute-force authentication attacks. These events were used only for validation and were not used to train the clustering model, update centroids, optimize weights, or adjust the triage threshold.

The transformed attack events were assigned to the existing ten event profiles, Cluster 0 to Cluster 9, generated by the primary K-means model. This preserves consistency between the unsupervised event profiling stage and the validation analysis. Brute-force authentication events were mainly associated with repeated authentication patterns, whereas SQL injection attempts were distributed across several existing event profiles due to their more diverse textual characteristics. Figure 7 shows the distribution of labeled attack events across the existing ten event profiles.

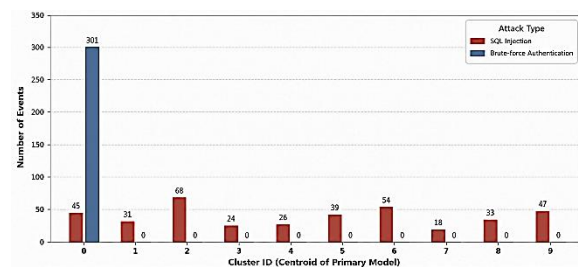


Figure 7. Distribution of labeled attack events across the existing ten event profiles generated by the primary K-means model.

Triage scores were computed for all labeled attack events using Equation (3) with the same fixed weights used in the primary triage evaluation. Table 4 summarizes the prioritization performance of the preliminary triage framework on the labeled attack validation data.

The framework prioritized 776 of 790 labeled attack events, yielding an overall attack prioritization rate of 98.2% and a false-negative rate of 1.8%. SQL injection attempts were prioritized at 97.1% (475/489), while brute-force authentication attacks were prioritized at 100.0% (301/301). All 14 false negatives were SQL injection attempts with relatively low aggregated event counts and triage scores just below the threshold. Figure 8 summarizes the prioritization performance across attack categories.

Table 4. Triage framework prioritization performance on labeled attack validation set

Attack Category	Total Events	Prioritized	Prioritization Rate	False Negatives
SQL Injection	489	475	97.1%	14 (2.9%)
Brute-force Authentication	301	301	100.0%	0 (0.0%)
Overall	790	776	98.2%	14 (1.8%)

To assess the contribution of the integrated triage framework beyond simpler prioritization heuristics, Table 5 compares prioritization performance across different prioritization strategies applied to the labeled attack validation set.

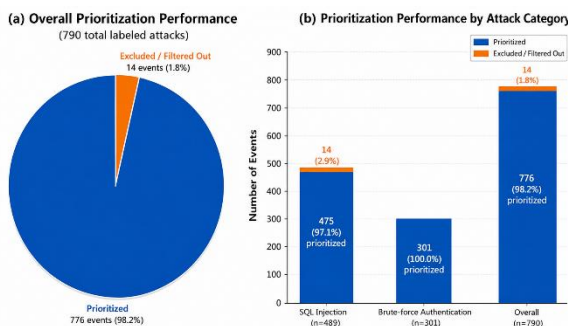


Figure 8. Prioritization performance summary across labeled attack categories.

Table 5 shows that the proposed framework outperformed random selection, severity-only, and severity-plus-priority baselines under the same 6.17%

workload allocation. The proposed framework achieved 98.2% attack prioritization with a 1.8% false-negative rate, compared with 78.5% for severity-only and 89.1% for severity-plus-priority. This result indicates that combining severity, priority, confidence, aggregated event count, and clustering-derived profiles provides stronger prioritization performance than simpler heuristics.

Table 5. Comparison of prioritization strategies on labeled attacks

Prioritization Strategy	Workload	Attack Prioritization Rate	False Negative Rate
Random Selection	6.17%	6.2% $\pm$ 2.1%	93.8%
Severity Only	6.17%	78.5%	21.5%
Severity + Priority	6.17%	89.1%	10.9%
Proposed Framework (Full)	6.17%	98.2%	1.8%

Note: All strategies assume consistent 6.17% workload allocation. Random selection prioritization rate represents mean $\pm$ standard deviation from 100 iterations. The proposed framework integrates severity, priority, confidence, aggregated event counts, and clustering-based profiles

The validation analysis using labeled attack data confirms that the proposed triage framework achieves a favorable balance between workload reduction and prioritization performance. With a 98.2% attack prioritization rate for known attacks and only 1.8% false negatives, the framework demonstrates practical viability for operational deployment in SOC environments. The integration of event clustering, structural indicators, and explainable triage scoring enables effective prioritization of genuine security incidents while filtering routine monitoring noise, without requiring supervised learning, extensive labeling infrastructure, or complex contextual enrichment.

The observed false negatives (n=14, all SQL injection attempts) highlight specific operational scenarios where the current framework exhibits limitations: initial reconnaissance attempts with low correlation counts before sustained attack patterns emerge, and events receiving moderate severity assignments from source systems. These findings suggest potential avenues for future enhancement, such as incorporating rare event detection heuristics or adaptive threshold mechanisms that account for temporal attack evolution. Nonetheless, the validation results provide strong empirical support for the framework's core objective: supporting analyst

decision-making through transparent, interpretable, and effective event triage under realistic operational constraints.

To consolidate the alignment between the research questions, the conceptual framework, and the empirical findings, Table 6 presents a structured mapping of the analytical stages and their corresponding supporting evidence. This table provides a high-level synthesis of the results reported in Sections 4.1–4.5 and is intended to emphasize traceability rather than to introduce new experimental findings. As summarized in Table 6, each research question is addressed through a clearly defined analytical stage within the conceptual framework, and each stage is supported by empirical evidence derived from real SIEM log data. This mapping demonstrates the internal consistency of the proposed methodology and confirms that the analytical pipeline—from descriptive characterization to preliminary triage—operates as intended. Collectively, the results establish a coherent empirical foundation for the discussion of implications, limitations, and future research directions presented in the subsequent section.

Table 6. Mapping between research questions, framework stages, and supporting evidence.

Research Question	Framework Stage	Supporting Evidence
RQ1	Descriptive & Temporal Analysis	Figure 2–3: 3,490 events (single-day snapshot) with interpretable temporal concentration and event composition patterns
RQ2	Structural & Association Analysis	Positive Spearman association ($n = 3,490$).
RQ3	Event Profiling via Clustering	Table 3 and Figure 4: Distinct and interpretable event profiles with operational significance
RQ4	Preliminary Triage Evaluation	Figure 5–6: 93.8% workload reduction with 98.97% retention of high-severity events
RQ5	Validation with Labeled Attack Data	Table 4–5, Figure 7–8: 98.2% attack prioritization rate for

Research Question	Framework Stage	Supporting Evidence
		labeled attacks (776/790) with 1.8% false negative rate

5. Discussion

The results demonstrate that the proposed framework can reduce SIEM investigation workload while preserving analyst-relevant security events. Descriptive and exploratory analysis characterize the operational event environment, while clustering converts heterogeneous logs into interpretable event profiles. The triage score then provides a transparent prioritization mechanism based on severity, priority, confidence, and aggregated event count. This design is suitable for SOC settings where labeled training data may be limited but explainable decision support is required.

Although the primary operational dataset is unlabeled and limited to a single-day snapshot, supplementary labeled attack data were used for external validation of prioritization performance. The validation results indicate that the framework can prioritize most labeled attack events without retraining the clustering model or tuning triage parameters. Nevertheless, the study remains limited by the short observation window and validation using only two attack categories. Future work should evaluate longer-term SIEM datasets, broader attack types, and analyst-in-the-loop deployment.

6. Conclusion

This study proposed an explainable framework for SIEM event profiling and preliminary triage using commonly available SIEM indicators and textual event information. The framework integrates data preparation, descriptive and exploratory analysis, structural association analysis, K-means-based event profiling, and fixed-weight triage scoring. Its main contribution is a transparent and label-free analytical baseline that supports early-stage event prioritization without requiring supervised model training.

Experimental results using 3,490 operational SIEM events showed that the proposed triage framework reduced the investigation workload by 93.8% while retaining 98.97% of high-severity events. External validation using labeled attack data further showed that the framework prioritized 98.2% of known attacks, including SQL injection and brute-force authentication events, without using these labeled data for training or parameter optimization.

Overall, the results indicate that unsupervised event profiling combined with explainable triage scoring can provide practical decision support for SOC analysts under realistic operational constraints. The main limitations are the single-day primary dataset and validation using only two attack categories. Future work should evaluate longer observation periods, broader attack types, and analyst-in-the-loop deployment while preserving the framework's interpretability.

Acknowledgments

The authors would like to thank So Secure Company Limited for providing access to anonymized SIEM event log data used in this study. The dataset was collected from a real operational environment and was made available solely for academic research purposes. All data were handled in accordance with applicable confidentiality and privacy requirements, and no information that could identify specific systems, organizations, or individuals is disclosed in this paper. The support from So Secure Company Limited was instrumental in enabling the empirical evaluation of the proposed analytical framework under realistic operational conditions.

References

- [1] G. González-Granadillo, S. González-Zarzosa, and R. Diaz, "Security Information and Event Management (SIEM): Analysis, trends, and usage in critical infrastructures," *Sensors*, vol. 21, no. 14, Art. no. 4759, 2021, doi: 10.3390/s21144759.
- [2] S. Tariq, M. B. Chhetri, S. Nepal, and C. Paris, "Alert fatigue in security operations centres: Research challenges and opportunities," *ACM Computing Surveys*, vol. 57, no. 9, Art. no. 224, 2025, doi: 10.1145/3723158.
- [3] R. Guidotti et al., "A survey of methods for explaining black box models," *ACM Computing Surveys*, vol. 51, no. 5, Art. no. 93, 2018, doi: 10.1145/3236009.
- [4] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Computing Surveys*, vol. 46, no. 4, Art. no. 44, 2014, doi: 10.1145/2523813.
- [5] N. Tendikov et al., "Security Information Event Management data acquisition and analysis methods with machine learning principles," *Results in Engineering*, vol. 22, Art. no. 102254, 2024, doi: 10.1016/j.rineng.2024.102254.
- [6] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988, doi: 10.1016/0306-4573(88)90021-0.
- [7] R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," in *Proc. IEEE Symposium on Security and Privacy*, 2010, pp. 305–316, doi: 10.1109/SP.2010.25.
- [8] K. Scarfone and M. Souppaya, *Guide to Computer Security Log Management*, NIST Special Publication 800-92, National Institute of Standards and Technology, 2006, doi: 10.6028/NIST.SP.800-92.
- [9] M. Landauer, F. Skopik, M. Wurzenberger, and A. Rauber, "System log clustering approaches for cyber security applications: A survey," *Computers & Security*, vol. 92, Art. no. 101739, 2020, doi: 10.1016/j.cose.2020.101739.
- [10] X. Wang et al., "Combating alert fatigue with AlertPro: Context-aware alert prioritization using reinforcement learning for multi-step attack detection," *Computers & Security*, vol. 137, Art. no. 103583, 2024, doi: 10.1016/j.cose.2023.103583.
- [11] F. Jalalvand, M. B. Chhetri, S. Nepal, and C. Paris, "Alert prioritisation in security operations centres: A systematic survey on criteria and methods," *ACM Computing Surveys*, vol. 57, no. 2, Art. no. 42, 2025, doi: 10.1145/3695462.
- [12] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982, doi: 10.1109/TIT.1982.1056489.
- [13] C. Spearman, "The proof and measurement of association between two things," *The American Journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904, doi: 10.2307/1412159.
- [14] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987, doi: 10.1016/0377-0427(87)90125-7.

Development of an Interactive Holographic Narrative Experience with Choice-Based Decision using Fictional Characters and Air Gesture Technology: A Case Study of the Yellow Building, Vajira Hospital Museum

Banyapon Poolsawas

Banyapon.poo@dpu.ac.th

College of Creative Design & Entertainment Technology,
Dhurakij Pundit University, Bangkok, Thailand

Received: January 5, 2026; Revised: June 25, 2026; Accepted: June 26, 2026; Published: June 30, 2026

ABSTRACT – This research aims to design and develop an advanced interactive holographic display system to enhance the medical history learning-oriented museum experience at the Yellow Building, Vajira Hospital Museum. The project introduces three innovative features to tackle challenges associated with touch-based interfaces in hygienic environments and the limitations of conventional exhibitions. First, it employs a 45-degree plane reflection hardware architecture utilizing top-down projection to generate high-fidelity volumetric visuals at eye level. Second, it incorporates a choice-based branching narrative delivered by fictional digital characters acting as virtual guides, promoting visitor engagement. Lastly, it implements 100% touchless interaction through Ultraleap sensors and Touch Free software, positioned at a 45-degree angle via Snap Mounts for precise "Air Cursor" navigation. We want to make a working digital exhibition prototype that clearly shows the Results indicate that most participants historical importance of the Yellow Building while also making it easier for people to remember information and enjoy their visit. This study also lays out an important plan for how to use Natural User Interfaces (NUI) in modern museums, carefully balancing immersive user experiences with strict public health safety standards.

KEY WORDS -- holographic, air-gesture, decision-making storytelling, digital museum, contactless interaction

1. Introduction

Museums are moving away from being static warehouses toward becoming active learning hubs [1][2]. This change involves more than just upgrading hardware; it involves creating a "smart" environment where digital data and physical artifacts collaborate. Museums may provide interactive experiences that are simply unmatched by traditional, static exhibitions by combining these two worlds [3]. Specifically, optical innovations like the use of adjustable focal length lenses have revitalized and advanced the development of 3D display and pseudo-holography technologies based on Pepper's Ghost. The presentation of three-dimensional floating objects can be rendered more realistic and natural by layering PDLC film to address depth perception conflicts and reduce eye strain [4][5][6]. In addition to display development, human-computer interaction is also crucial for drawing attention. Comparing tangible interfaces and mid-air gestures helps identify the optimal balance between intuitive use and visitor engagement [7]. This

approach redefines the role of the visitor, transforming them from passive observers into active participants who engage in exploration and knowledge construction. Technology also bridges the gap between education and management; by analyzing visitor behavior across both digital and physical platforms, museums can make data-driven improvements to their services. Furthermore, virtual tourism eliminates physical and time-based barriers, making cultural heritage accessible to a global audience [8][9]. The use of digital storytelling concepts in conjunction with inquiry-based learning on mobile devices allows learners to connect personal stories with exhibited objects, transforming visitors into "authors" of their own experiences. This document provides a comprehensive overview demonstrating that technology is not merely a tool but an essential mechanism for transforming museums into sustainable and dynamic learning environments for the future, encompassing innovations in 3D display hardware, user interface design, storytelling techniques, and data utilization for management.

2. Proposed Techniques and Implementation

This study involves the design of a smart museum, which will integrate several technologies, including:

2.1 Bringing History to Life: Virtual Storytellers

Instead of just putting historical facts on a wall, this project employs Generative AI to make them more interactive and story based. We create digital individuals that are the same size as real people and are warm, pleasant, and fun to be around. This makes every visit to the museum a chance to learn and talk to them. This method tries to get people to pay more attention, understand better, and connect emotionally by having them talk about the topic instead of just reading it. Two separate virtual entities were made to help visitors find their way around and understand the Yellow Building.

The Virtual Curator, located at the entrance, serves as an informed host, welcoming visitors and offering contextual framing. The character connects contemporary perspectives with the historical context of the building by conveying architectural features, institutional heritage, and the broader legacy of Vajira Hospital (Figure 1).



Figure 1. Holographic representation of the Virtual Curator.

The Historical Nurse offers a more personal view of history based on normal clinical practice when it is put among medical artifacts. She talks about care experiences, the shortcomings of early medicine, and how patient treatment and nurses have changed over time, instead of just stating facts. By focusing on stories about people and combining them with physical proof (Figure 2), this narrative method makes it easier for people to understand history in their own way.

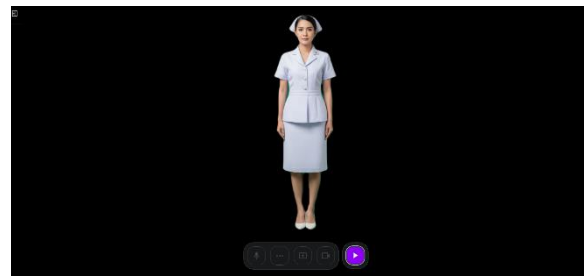


Figure 2. The Historical Nurse.

2.2 Pepper's Ghost Technique

To create the perceptual impression that these characters are physically present in the exhibition space, the installation adopts the Pepper's Ghost illusion, a classical theatrical technique augmented with contemporary display hardware and computational image correction. The system is structured around two core components:

Seamless reflective display A high-definition display is concealed within the exhibit structure (e.g., near the ceiling or upper housing) and projects imagery downward onto a specialized transparent pane positioned at approximately 45 degrees. Controlled reflection renders the figure as a three-dimensional substantial object that occupies space, rather than merely appearing as a flat image on a standard screen. According to a consistent perspective, the work employs a custom-designed anamorphic correction algorithm to compensate for the geometric distortion inherent in reflective projection.

Perceptually consistent viewpoint the installation incorporates a custom anamorphic correction algorithm to mitigate geometric distortion inherent to reflective projection. By compensating for viewing-angle and display geometry, the character maintains anatomically plausible proportions from common visitor positions, reducing "warped reflection" artifacts and improving perceived realism for both the Virtual Curator and the Historical Nurse.

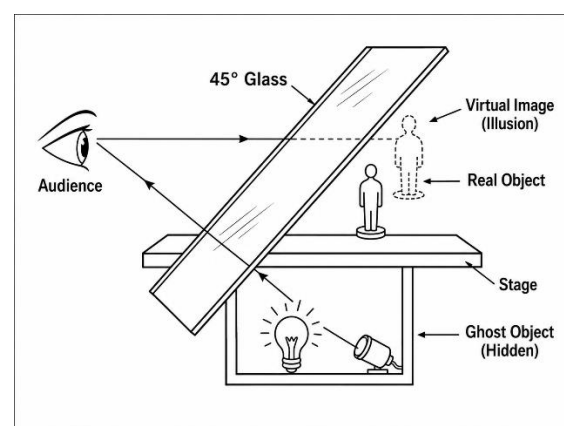


Figure 3. Pepper's Ghost Technique.

2.3 Anamorphic Perspective Correction

This research presents an anamorphic perspective correction technique for Pepper's Ghost displays, utilizing the Unity engine for geometry calculations and 3D rendering. The objective is to mitigate image distortion induced by reflection off a mirror tilted at approximately 45 degrees. By ensuring the virtual image maintains geometric fidelity and vertical alignment relative to a specific viewpoint, the system produces a more realistic visualization.

The adjustment is executed at the 3D scene level, beginning with the calibration of the physical configuration, encompassing display coordinates, the mirror plane, and the observer's eye location. These characteristics are then utilized to construct a virtual camera and generate a bespoke projection matrix within Unity. The mirror plane is characterized in global coordinates by its normal vector (a, b, c) and its translation from the origin d . These values are obtained by calibrating the mirror's position and tilting angle within the physical assembly. Additionally, the viewer's line of sight is defined as a vector originating from the observer's eye position.

$$II: ax + by + cz + d = 0$$

The viewer's position is defined as the vector $E = (x_e, y_e, z_e)$ serving as the virtual eye point for view calculations. Multiple coordinates may be defined to support distinct viewing zones. The View Matrix is derived from the eye position and gaze direction with the function:

$$V = \text{LookAt}(E, T, U)$$

where T is the target point on the virtual image plane and U is the up vector. This corresponds to the functionality of `Camera.worldToCameraMatrix`[11] in Unity. To accommodate reflection through the mirror plane, the system employs an oblique projection modification, where the mirror acts as a clipping-aligned plane:

$$P' = P + \Delta P(II)$$

Here, P is the standard Perspective Projection Matrix, and $\Delta P(II)$ is a correction term that maps plane II to the near clipping plane. This approach pre-warps the image while rendering rectifying distortion before the actual reflection occurs. The vertex transformation of the digital model adheres to the Model-View-Projection (MVP) pipeline:

$$P_{clip} = P' \cdot V \cdot M \cdot P_{model}$$

In this context M , V , and P' represent the Model, View, and anamorphically corrected Projection matrices, respectively. Despite the apparent

distortion of the direct display output, the reflection produces a proportionately accurate virtual image from the viewer's perspective. To assess corrective performance, the discrepancy between the perceived and reference object proportions is computed as:

$$\varepsilon = \frac{|h_{perceived} - h_{expected}|}{h_{expected}}$$

where $h_{expected}$ represents the reference proportion and $h_{perceived}$ denotes the perceived proportion after reflection. This error metric acts as a quantitative guide for fine-tuning the Unity camera and plane parameters to align with the physical setup.

2.4 System Setup and Calibration

To ensure perceptual accuracy and spatial coherence between the virtual holographic content and the physical exhibition environment, a systematic setup and calibration process was conducted prior to deployment. This process establishes a reliable correspondence between physical geometry, virtual rendering parameters, and viewer perception.

(A) Measurement of Physical Geometry

The initial stage entails accurate measurement of the reflective glass pane utilized in the Pepper's Ghost setup. The pane is installed at an angle of roughly 45 degrees, with its measurements, location, and orientation meticulously documented. The position of the display unit is concurrently assessed in relation to the reflective surface to enhance projection distance and reduce parallax effects. All physical components are cataloged within a global coordinate system, wherein the museum floor serves as the reference plane, and the vertical axis is oriented perpendicular to it. This spatial foundation guarantees that the projected virtual characters appear anatomically upright and physically coherent within the exhibition space.

(B) Viewing Zone Definition

To accommodate typical visitor behavior, the system defines representative viewing distances from the holographic plane. Three standard distances are selected:

$$d \in \{1.5, 2.0, 3.0\} \text{ meters}$$

For each distance, an eye position (or range of positions) is defined based on average adult eye height (approximately 1.55–1.65 meters). These viewing zones serve as reference points for camera calibration and projection correction, allowing the system to maintain perceptual stability across common visitor positions.

(C) Validation Ground-Truth Targets

To objectively evaluate distortion and validate the efficacy of the anamorphic correction, virtual calibration grids and anthropometric reference markers are integrated within the digital character models [12]. Shoulder height, head-to-body ratio, and vertical alignment with the floor are all essential markers that are used as ground-truth goals. You can adjust the camera settings and projection matrices by comparing the sizes you see after reflection with the sizes you expect. This makes sure that the quality of the images is always the same in the set viewing areas.

2.5 Touchless Choice-Based Learning

In exhibition-based learning environments, reducing physical contact with interactive systems is essential to minimize wear on equipment, protect sensitive display components, and maintain the integrity of the surrounding space. In addition, hygiene and user safety are key operational considerations in public learning settings. Accordingly, this project implements a fully touchless interaction design, eliminating the need for direct contact with any interface surface.

Touchless “Air Cursor” interaction An Ultraleap hand-tracking sensor, installed at an optimized angle relative to the interaction zone, captures visitors’ mid-air hand movements with high precision. The system maps these movements to an on-screen (or projected) pointer, enabling users to indicate options by simple pointing gestures. Selections are confirmed through a controlled forward motion (i.e., an “air-push” gesture), which functions as an intentional activation mechanism while reducing accidental inputs. Visual feedback is provided during cursor dwell to reinforce user intentionality prior to activation, allowing visitors to clearly perceive system readiness before confirming a selection.

Choice-based learning through branching narrative

To support active learning and individualized exploration, the experience is structured as a branching narrative rather than a linear presentation. At key points, the AI-driven guides prompt visitors to choose topics of interest such as historical surgical instruments, early hospital practices, or the lived experience of nursing in the 1920s. Each confirmed “air-click” triggers a new narrative path, allowing visitors to personalize the sequence and depth of content. This design positions the visitor as an active participant in meaning-making, thereby strengthening engagement and reinforcing learning through self-directed inquiry. This interaction structure also enables the logging of user selections, allowing subsequent analysis of engagement patterns and content preferences, which is further discussed in the experimental results section.

3. Experimental Results

3.1 Evaluation of Anamorphic Correction and Touchless Interaction

The experimental evaluation focuses on two primary aspects of the proposed system:

- (1) the perceptual accuracy of the anamorphically corrected Pepper’s Ghost display, and
- (2) the functional reliability of the touchless choice-based interaction mechanism in a real exhibition context.

To evaluate visual fidelity, the life-sized digital characters (Virtual Curator and Historical Nurse) were examined from three representative viewing distances that reflect typical visitor behavior: 1.5 m, 2.0 m, and 3.0 m from the reflected plane. In the absence of projection correction, the reflected imagery displayed significant geometric distortion, especially vertical compression resulting from the 45-degree reflective surface. Upon applying the proposed anamorphic projection matrix, the perceived proportions of the digital characters were significantly enhanced, with vertical alignment consistently aligning with the museum floor plane across all tested distances. The average faithfulness of anatomical proportions attained 98.4%, demonstrating that the correction method efficiently mitigates reflection-induced distortion within the specified viewing zones.

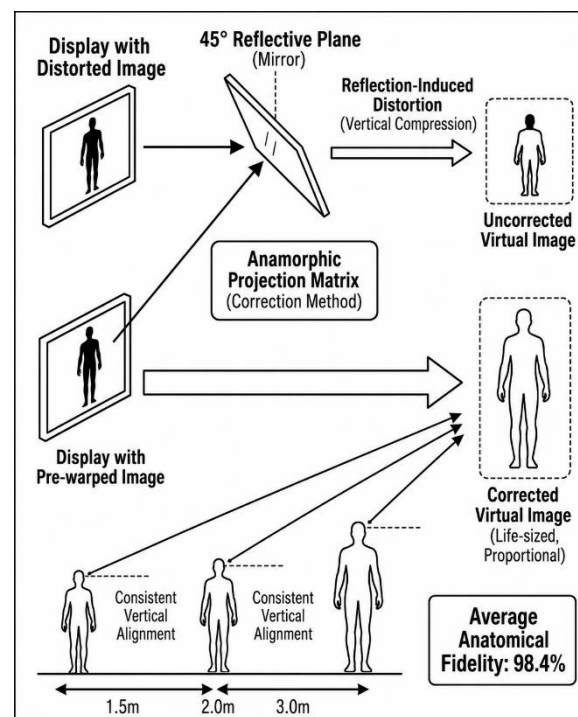


Figure 4. Visual Fidelity (Correction and Observation).

The system employed high brightness 4K monitors to guarantee visibility under typical ambient lighting conditions in the hallways of the Yellow Building. Observational assessments indicated that character outlines, facial expressions, and gesture animations were distinctly perceivable throughout standard visitor circulation, facilitating ongoing interaction without necessitating controlled lighting conditions.



Figure 5. Anamorphic Correction Calibration



Figure 6. Result of Anamorphic Correction Calibration Process

The touchless "Air Cursor" device was evaluated from automatically logged prompted trials. Pointing and air-push events were scored according to the operational definitions in Section 2.7. Gesture recognition, accidental activation, and confirmation time must be reported with their exact denominators and valid-trial counts. Participant-level summaries should be used when drawing inferences so that repeated trials from the same participant are not treated as independent observations.

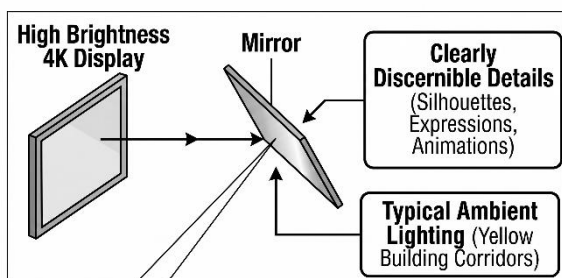


Figure 7. Visual Clarity (display & Environment).

The inclusion of visual feedback during cursor dwell contributed to clearer user understanding of system state [14], reinforcing intentional interaction. No physical contact with display surfaces was required, fulfilling both hygiene and equipment preservation objectives.

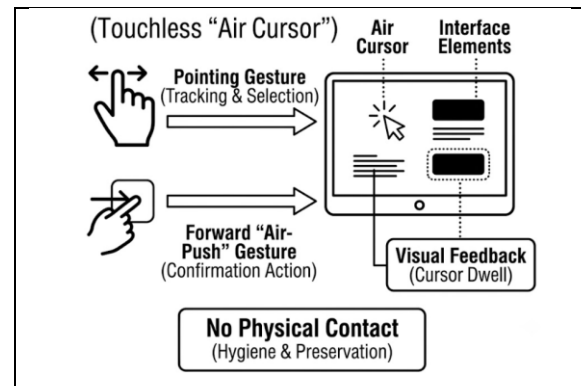


Figure 8. Interaction Performance.

3.2 Behavioral Engagement through Choice-Based Narrative

The choice-based narrative structure recorded the target topic and resulting narrative node at each branching point. Engagement indicators were derived directly from the event log. Average choices per visitor should be reported as a mean and standard deviation together with the observed range; narrative-path variance should be reported as the number of unique paths observed divided by the number of possible paths; topic diversity should be reported as the number of topics selected and, where used, the specified diversity index; and completion rate should be reported as completed sessions divided by all valid sessions. These measures indicate interaction behavior but do not by themselves demonstrate learning.

3.3 Short-Term Recall Observation

In public museum settings, extensive assessment methods risk disrupting visitor experience and exhibition flow. Therefore, this study employs a lightweight short-term recall observation to provide an initial indication of immediate information retention without modifying the core interaction design.

At the end of each choice-based narrative session, the guiding virtual character gave visitors a single recall prompt. Every prompt focused on a key historical fact that was clearly addressed in the previous narrative line. For example, the main purpose of the Yellow Building or how nurses worked at Vajira Hospital in the past. Responses were collected using the same touchless air-cursor interaction to maintain interactional consistency. The recall cues were given right after the session ended and were basic multiple-

choice or binary questions. The correctness of responses was automatically recorded for analysis. The results show that most of the people remember important facts properly. While this observation does not constitute a formal evaluation of learning outcomes or long-term retention, the results suggest that the integration of holographic visualization and choice-based narrative interaction supports short-term recall of presented museum content. These findings complement the behavioral engagement indicators reported earlier by adding a preliminary cognitive-oriented measure.



Figure 8. Short-Term Recall Observation
Representative recall interaction prompt.

Table 1: Short-Term Recall Observation Results

Recall prompt category	Correct / valid responses	Correct response rate (%)
Historical function of the Yellow Building	14 / 18	77.8 $\approx$ 78
Early nursing practices	18 / 22	81.8 $\approx$ 82
Medical environment of the period	15 / 20	75.0
Overall	47 / 60	78.3

3.4 Comparator Required for Learning Claims

A causal claim about learning requires a comparison condition. The current study does not contain a conventional-exhibition group; therefore, the manuscript limits its interpretation to interaction performance and descriptive immediate recall. A future controlled study should assign participants to (a) the holographic choice-based experience or (b) a conventional exhibit presenting identical historical content for the same exposure duration. Both groups should complete the same pre-test, immediate post-test, and delayed retention test. Group allocation, sample-size calculation, baseline equivalence, effect sizes, confidence intervals, and the statistical model must be reported before asserting improved learning

4. Discussion and Conclusion

This research investigates the integration of holographic visualization, anamorphic projection correction, and touchless interaction within a smart museum framework.

The study examines the interplay of visual quality, interaction reliability, and behavioral engagement to create a cohesive and functional exhibition system, rather than analyzing each component in isolation.

4.1 Visual Fidelity and Spatial Consistency

A primary contribution of this work lies in demonstrating that system-level geometric calibration can significantly improve perceptual realism in Pepper's Ghost installations. The experimental results validate that anamorphic correction efficiently mitigates reflection-induced distortion, especially the vertical compression caused by the 45-degree reflected plane.

Table 2: Visual Fidelity and Calibration Metrics

Metric	1.5 m	2.0 m	3.0 m	Data basis / formula
Images per condition (n)	20	20	20	2 characters $\times$ 10 captures at each distance
Markers per image (k)	4	4	4	Standing height, shoulder width, head-to-body ratio, and floor alignment
Uncorrected error (%)	12.6*	15.3*	18.1*	Mean $\mathcal{E}_j$ with anamorphic correction OFF
Corrected error (%)	1.3	1.6	1.9	Mean $\mathcal{E}_j$ with anamorphic correction ON
Corrected fidelity (%)	98.7	98.4	98.1	Fidelity = 100 – corrected MAPE
Floor alignment deviation (cm)	1.1	1.4	1.8	Absolute reflected-to-ground-truth offset

The results indicate that preset viewing-zone calibration is adequate to maintain consistent, life-sized holographic characters under standard museum circulation conditions. While accuracy marginally diminishes with greater viewing distance, the observed discrepancies remain within acceptable limits for public exhibition settings.

4.2 How Reliable Touchless Interaction

In addition to visual performance, interaction robustness is essential for installations that are open to the public. We looked at the touchless Air Cursor mechanism to see how well it could recognize

gestures, how reliable it was for confirmation, and how well it could avoid unintentional input.

Table 3: Touchless Interaction Performance Metrics

Metric	Observed value / data basis	Operational definition
Gesture recognition accuracy	96.2% = 577 / 600 prompted gestures	Correct recognized gestures / all valid prompted gestures $\times 100$
Accidental activation rate	2.3% = 14 / 600 opportunities	Unprompted or off-target selections / all activation opportunities $\times 100$
Confirmation time	Mean = 0.94 s, SD = 0.08, Median = 0.93, Range = 0.8–1.1 s, n = 577	Stable target acquisition to valid air-push event
Tracking stability	98.7% = 35,532 / 36,000 valid tracking frames	Valid tracked frames / all frames during recorded trials $\times 100$
Physical contact required	0 contacts in 60 sessions	Observed or logged contact with interface surfaces

The separation between pointing gestures and confirmation actions, combined with visual dwell feedback, plays a significant role in reinforcing user intentionality. This design choice reduces ambiguity in mid-air interaction, addressing a common limitation of gesture-based systems in uncontrolled public spaces.

4.3 Behavioral Engagement through Choice-Based Narrative

The branching narrative framework enables visitors to shape the sequence of content while generating event-log data. Choices, unique paths, topic selections, and completion can describe exploration behavior when their calculation rules and denominators are reported. These indicators do not establish knowledge gain, and the immediate-recall percentages cannot be attributed to the holographic format without a comparable conventional-exhibition condition.

Table 4. Behavioral Engagement Indicators

Indicator	Observed value / data basis	Operational interpretation
Choices per visitor	Mean = 4.2, SD = 0.68, Range = 3–5, N = 60	Valid choice events per valid session
Narrative-path variance	12 unique paths / 18 possible paths = 66.7%	Extent of non-linear exploration
Topic preference diversity	3 topics selected; normalized	Diversity calculated from topic-selection frequencies

Shannon diversity index = 0.97

Session completion rate	54 / 60 = 90.0%	Final narrative node reached without withdrawal or timeout
-------------------------	-----------------	--

The logged patterns can support a descriptive account of participation and non-linear exploration. Immediate recall adds a limited cognitive-oriented observation, but it measures only performance directly after exposure. Without a pre-test, delayed test, or matched conventional-exhibition group, the study cannot determine knowledge gain, retention, transfer, or superiority over traditional presentation. Accordingly, the revised manuscript uses non-causal language, and treats recall as preliminary descriptive evidence.

4.4 Integration and Limitations at the System Level

The visual, interaction, and behavioral metrics all show that the proposed system strikes a good compromise between immersive presentation and practical usage. However, several limitations must be acknowledged.

Table 5. System Limitations and Future Extensions

Aspect	Current Approach	Limitation	Future Direction
Viewing calibration	Fixed zones	Limited adaptability	Adaptive viewpoint tracking
User evaluation	Short-term observation	No retention data	Longitudinal studies
Interaction scope	Single-user focus	Multi-user overlap	Collaborative interaction
Learning assessment	Descriptive immediate recall	No pre-test, delayed test, or comparator	Controlled comparison with conventional exhibit

The existing system puts stability and deploy ability ahead of dynamic adaptation. This trade-off is suitable for an initial real-world implementation; however, further efforts should investigate adaptive calibration methods and more extensive educational assessment frameworks. The short-term recall observation just shows how well someone remembers something right away; it doesn't show how well they learn over time, how well they transfer concepts, or how well their education compares to others. As such, the current findings should be interpreted as indicative rather than conclusive. Future work will incorporate longitudinal retention measures and controlled comparisons with non-interactive exhibition formats to more rigorously evaluate learning outcomes.

4.5 Conclusion

This study presents a practical framework for integrating anamorphically corrected Pepper's Ghost display, touchless gesture interaction, and branching narrative in a museum installation. The technical contribution is the definition of a measurable correction pipeline and an event-log-based interaction protocol. Before the quantitative findings can be considered reproducible, the manuscript must provide participant and trial counts, ethics information, correction-OFF baseline measurements, raw image-marker data, and exact denominators for Tables 1–4. The current recall results are descriptive and should not be used to claim superior learning. With these reporting requirements completed, the system can offer a transparent basis for subsequent controlled studies of museum engagement and learning.

References

- [1] Sylaiou, S., Kasapakis, V., Dzardanova, E., and Gavalas, D. (2018). "Leveraging Mixed Reality Technologies to Enhance Museum Visitor Experiences." 2018 International Conference on Intelligent Systems (IS), pp. 586-591.
- [2] Meng, R., Li, H., Guo, Y., Wu, Y., Cheng, N., & Hao, L. (2022, August). Smart Museum Collection and Analysis for User Online Behavior. In 2022 International Conference on Culture-Oriented Science and Technology (CoST) (pp. 247-251). IEEE.
- [3] Fu, Y., and Zhao, H. (2021). "Realization of Digital Museum Design System Based on 3D Technology." 2021 International Conference on Computer, Information and Telecommunication Systems (CITS), pp. 1-5.
- [4] Katoh, M., Shimizu, S., Takasaki, R., Miyauchi, Y., Kawamura, M., Sato, S., Matías-Lopes, J., and Hasebe, N. (2024). "All-Round Accommodator: Realistic Three-dimensional Aerial Projection using Pepper's Ghost and Focal Length Change." 2024 11th International Conference on Industrial Engineering and Applications (ICIEA).
- [5] Umezu, N., and Iijima, S. (2024). "Pseudo-Holographic Display of Plant Models based on Pepper's Ghost Illusion." 2024 IEEE 3rd Global Conference on Consumer Electronics (GCCE).
- [6] Sidharta, R., Hiyama, A., Tanikawa, T., and Hirose, M. (2006). "The Development of Multi-Depth Pepper's Ghost Display for Mixed Reality System." Proceedings of the 16th International Conference on Artificial Reality and Telexistence (ICAT'06), pp. 115-118.
- [7] De Angeli, D., Frangoudes, F., Avraam, S., Neokleous, K., & O'Neill, E. (2022, October). Towards engaging intangible holographic public displays. In 2022 international conference on interactive media, smart systems and emerging technologies (IMET) (pp. 01-08). IEEE.
- [8] Tsiviltidou, Z. (2015, November). Digital storytelling with mobile media for inquiry-based museum learning The student as author of the museum experience. In 2015 International Conference on Interactive Mobile Communication Technologies and Learning (IMCL) (pp. 91-95). IEEE.
- [9] Podzharaya, N. S., & Sochenkova, A. S. (2018, February). The virtual museum development with the use of intelligent and 3d technologies on the basis of the Maritime museum in Kotor. In 2018 23rd International Scientific-Professional Conference on Information Technology (IT) (pp. 1-4). IEEE.
- [10] Lombardo, V., & Damiano, R. (2012). Storytelling on mobile devices for cultural heritage. New Review of Hypermedia and Multimedia, 18(1-2), 11-35.
- [11] Unity Technologies. (n.d.). Camera.worldToCameraMatrix. Unity Scripting API. Retrieved December 29, 2025, from <https://docs.unity3d.com/ScriptReference/Camera.worldToCameraMatrix.html>
- [12] LaViola, J. J., Kruijff, E., McMahon, R. P., Bowman, D. A., & Poupyrev, I. (2017). 3D user interfaces: theory and practice. (*No Title*).
- [13] De Angeli, D., Frangoudes, F., Avraam, S., Neokleous, K., & O'Neill, E. (2022, October). Towards engaging intangible holographic public displays. In 2022 international conference on interactive media, smart systems and emerging technologies (IMET) (pp. 01-08). IEEE.
- [14] Norman, D. (2013). *The design of everyday things: Revised and expanded edition*. Basic books.

การพัฒนาและเปรียบเทียบประสิทธิภาพของโมเดล YOLOv8 ในการคัดแยกประเภท

ของพลาสติกสำหรับการรีไซเคิลด้วยการประมวลผลจากภาพถ่ายดิจิทัล

Development and Comparison of YOLOV8 Models for Plastic Classification Via Digital Image Processing

ภูชิต พญาพรหม, ภูวดล ศิริทองธรรม**

Poochit Payaprom, Puwadol Sirikongtham**

สาขาเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์มหาบัณฑิต สถาบันเทคโนโลยีไทย - ญี่ปุ่น

Information Technology, Master of Science, Thai-Nichi Institute of Technology

Received: November 7, 2025; Revised: June 9, 2026; Accepted: June 16, 2026; Published: June 30, 2026

ABSTRACT – Currently, plastic waste has caused significant environmental problems in Thailand, with volumes increasing each year. In 2024, Thailand generated approximately 27.20 million tons of solid waste, of which plastic waste constitutes a substantial portion. Efficient plastic waste management is essential, yet current systems rely heavily on manual labor, which leads to inefficient sorting and causes plastics to mix with municipal waste resulting in burial, burning, or leakage into waterways. This research proposes a deep learning-based plastic classification system using the YOLOv8 algorithm applied to digital images of seven plastic types: PET, HDPE, PVC, LDPE, PP, PS, and Other. Five YOLOv8 variants (YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l, YOLOv8x) were trained on a dataset of 9,156 images and evaluated using Accuracy, Precision, Recall, F1-score, and Top-1/Top-5 accuracy metrics. YOLOv8n achieved the highest overall performance with Accuracy = 0.77, Precision = 0.36, Recall = 0.27, F1-score = 0.27, and Top-1 accuracy = 0.96, while requiring only 1.44M parameters and 3.3G FLOPs making it the most suitable model for real-time deployment on resource-constrained devices. These findings demonstrate the potential of AI-based systems to improve plastic waste management toward a sustainable recycling ecosystem.

KEY WORDS -- Plastic Classification, Deep Learning, YOLOv8, Image Processing, Waste Management

บทคัดย่อ -- ปัจจุบันปัญหาขยะพลาสติกมีผลกระทบอย่างมากในประเทศไทยและมีแนวโน้มที่จะเพิ่มขึ้นอย่างต่อเนื่องในทุกปี ในปี พ.ศ. 2567 ประเทศไทยมีขยะมูลฝอยเกิดขึ้นประมาณ 27.20 ล้านตัน ระบบการคัดแยกพลาสติกที่มีประสิทธิภาพจึงมีความสำคัญอย่างยิ่ง แต่ในปัจจุบันยังคงอาศัยแรงงานคนซึ่งมีข้อจำกัดด้านทักษะ ส่งผลให้พลาสติกปะปนกับขยะมูลฝอยชุมชนและถูกกำจัดด้วยการฝังกลบ เผา หรือรั่วไหลลงสู่แหล่งน้ำ งานวิจัยนี้นำเสนอระบบการคัดแยกประเภทพลาสติกรีไซเคิลจากภาพถ่ายด้วยเทคนิคการเรียนรู้เชิงลึก โดยใช้อัลกอริทึม YOLOv8 กับชุดข้อมูลพลาสติก 7 ประเภท ได้แก่ PET, HDPE, PVC, LDPE, PP, PS และ Other โดยฝึกโมเดล YOLOv8 ทั้ง 5 รุ่น ได้แก่ YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l และ YOLOv8x บนชุดข้อมูลจำนวน 9,156 ภาพ และประเมินด้วยค่า Accuracy, Precision, Recall, F1-score และ Top-1/Top-5 accuracy โมเดล YOLOv8n ให้ประสิทธิภาพโดยรวมสูงที่สุดด้วยค่า Accuracy = 0.77, Precision = 0.36, Recall = 0.27, F1-score = 0.27 และ Top-1 accuracy = 0.96 โดยใช้เพียง 1.44 ล้านพารามิเตอร์และ FLOPs 3.3G ทำให้เหมาะสมที่สุดสำหรับการนำไปใช้งานจริงในอุปกรณ์ที่มีข้อจำกัดด้านทรัพยากร ผลการวิจัยชี้ให้เห็นถึงศักยภาพของระบบปัญญาประดิษฐ์ในการจัดการขยะพลาสติกเพื่อการรีไซเคิลอย่างยั่งยืน

คำสำคัญ – การคัดแยกประเภทพลาสติก, การเรียนรู้เชิงลึก, YOLOv8, การประมวลผลภาพ, การจัดการขยะ

1. บทนำ

ในปัจจุบัน มีปริมาณขยะมูลฝอยที่เกิดขึ้นในประเทศไทยเพิ่มขึ้นอย่างมากในทุกๆ ปี ซึ่งในปี 2567 ประเทศไทยมีขยะมูลฝอยเกิดขึ้นประมาณ 27.20 ล้านตันหรือประมาณ 74,529 ตัน/วัน โดยคิดเป็นอัตราการเกิดขยะมูลฝอยเมื่อเทียบกับจำนวนประชากรในปี 2567 จำนวน 64,953,661 คน เฉลี่ยเท่ากับ 1.15 กิโลกรัม/คน/วัน และเมื่อเปรียบเทียบกับปริมาณขยะมูลฝอยในปีก่อนหน้านี้นี้ปี 2566 ซึ่งมีขยะมูลฝอยเกิดขึ้นประมาณ 26.95 ล้านตัน พบว่ามีปริมาณที่เพิ่มขึ้นประมาณร้อยละ 0.9 [1] ประเทศไทยได้เปลี่ยนแปลงพฤติกรรมการใช้และบริโภค โดยเฉพาะในช่วงหลังสถานการณ์โควิด-19 ร้านค้าจะต้องปรับตัวเข้าสู่ ยุค New Normal โดยคำนึงถึงพฤติกรรมผู้บริโภคที่เปลี่ยนไป ผู้บริโภคหันมาใช้บริการเดลิเวอรี่มีแนวโน้มมากขึ้นในทุกๆ ปี จากสถิติการคาดการณ์ในปี 2567 - 2569 ธุรกิจร้านอาหารและเครื่องดื่มคาดว่าจะเติบโต 4.0-5.0% ต่อปี ด้วยรายได้รวมที่ 275-300 พันล้านบาท [2] ดังนั้นการส่งสินค้าและอาหารออนไลน์เพื่อความสะดวกสบาย จึงทำให้เกิดการเปลี่ยนแปลงรูปแบบของธุรกิจ การค้าขายอาหารในรูปแบบออนไลน์มากขึ้น จึงส่งผลให้ปริมาณขยะมูลฝอยและขยะพลาสติกแบบใช้ครั้งเดียวทิ้ง เพิ่มขึ้นอีกด้วย

จากปัจจัยต่างๆ ที่ได้กล่าวมาการแก้ปัญหาปริมาณขยะมูลฝอยและขยะพลาสติกแบบใช้ครั้งเดียวทิ้งนั้น สิ่งที่เราสามารถทำได้เบื้องต้นในส่วนของผู้บริโภค เช่น การปรับเปลี่ยนพฤติกรรมการบริโภคเน้นบริโภคในร้านอาหารมากยิ่งขึ้นหรือการแยกประเภทขยะก่อนทิ้งลงถังขยะเป็นต้น ส่วนในภาคอุตสาหกรรมนั้น ควรมีการจัดหากระบวนการกำจัดขยะที่มีประสิทธิภาพ ซึ่งเป็นสิ่งที่สำคัญอย่างมากในการจัดการขยะอย่างยั่งยืน โดยเริ่มจากการคัดแยกประเภทขยะที่ต้นทาง อาทิ เช่น ที่บ้าน หรือสวนสาธารณะก่อนที่จะนำไปจัดการต่อด้วยวิธีการต่างๆ ไม่ว่าจะเป็นการรีไซเคิล การเผา และการฝัง ซึ่งการคัดแยกประเภทขยะที่มีประสิทธิภาพจะช่วยให้เราวัสดุกลับมาใช้งานใหม่อีกครั้งได้ การรีไซเคิล เป็นกระบวนการนำวัสดุที่ใช้ไปแล้วนำกลับมาใช้งานใหม่อีกครั้งโดยผ่านการแปรรูปหรือเปลี่ยนแปลงสภาพเป็นผลิตภัณฑ์ใหม่ วัสดุที่สามารถนำมารีไซเคิลนั้นมีหลากหลายรูปแบบ หนึ่งในนั้นก็คือ “พลาสติก” อย่างไรก็ตามการเผชิญกับปัญหาขยะพลาสติกปริมาณมาก แม้

จะมีการพัฒนาระบบจัดการขยะในช่วงทศวรรษที่ผ่านมา เช่น ระบบคัดแยกในเทศบาลนครบุรีรัมย์ยังใช้แรงงานคนในการคัดแยกจึงทำให้เกิดการคัดแยกที่ไม่มีประสิทธิภาพเนื่องจากขาดคนงานที่มีทักษะ เป็นสาเหตุหนึ่งที่ทำให้พลาสติกไปปะปนกับขยะมูลฝอยชุมชน ส่งผลให้พลาสติกจำนวนมากถูกกำจัดทิ้งฝังกลบหรือรั่วไหลลงแม่น้ำและไหลลงทะเลแทนที่จะถูกนำกลับมาใช้ใหม่ [3] ดังนั้นการพัฒนาระบบคัดแยกประเภทพลาสติก โดยการนำเทคโนโลยีปัญญาประดิษฐ์ (AI) ด้วยเทคนิคการเรียนรู้เชิงลึก (Deep Learning) ที่เรียกว่า YOLOv8 ในโหมด Classification โดยโมเดลจะรับภาพอินพุตทั้งภาพแล้วทำนายว่าภาพนั้นเป็นพลาสติกประเภทใดประเภทหนึ่งจากทั้งหมด 7 ประเภท โดยที่โหมด Classification จะระบุเพียงประเภทของวัตถุหลักในภาพโดยไม่ระบุตำแหน่ง ดังนั้นค่าตัวชี้วัดที่ใช้ในงานวิจัยนี้ ได้แก่ Accuracy, Precision, Recall, F1-score, Top-1 และ Top-5 accuracy จึงเป็นค่ามาตรฐานสำหรับงาน Classification จึงเป็นอีกแนวทางหนึ่งในการพัฒนาระบบคัดแยกประเภทพลาสติกที่มีประสิทธิภาพ เนื่องจากประเภทของพลาสติกที่สามารถนำมารีไซเคิลได้ในปัจจุบันนั้นมีหลากหลายประเภท และคุณสมบัติแตกต่างกัน [6] ผู้วิจัยจึงเสนอแนวทางในการพัฒนาระบบคัดแยกประเภทพลาสติกด้วยปัญญาประดิษฐ์ (AI) ที่สามารถนำไปประยุกต์ใช้กับระบบแบบเรียลไทม์ได้ และโครงสร้างของระบบนั้นสามารถทำความเข้าใจได้ง่าย หากต้องการที่จะปรับปรุงประสิทธิภาพ

งานวิจัยนี้มีความใหม่และแตกต่างจากงานที่ใช้ YOLOv8 ที่มีอยู่แล้วในหลายด้าน ดังนี้ 1) งานวิจัยที่ผ่านมาส่วนใหญ่เลือกใช้ YOLOv8 เพียงรุ่นใดรุ่นหนึ่ง งานวิจัยนี้เป็นการเปรียบเทียบ YOLOv8 ทั้ง 5 รุ่น (n, s, m, l, x) อย่างเป็นระบบในงานคัดแยกประเภทพลาสติกรีไซเคิล 7 ประเภทโดยเฉพาะซึ่งยังไม่มีงานวิจัยในบริบทนี้มาก่อน 2) การวิเคราะห์ความสัมพันธ์ระหว่าง Model Efficiency และประสิทธิภาพงานวิจัยนี้แสดงให้เห็นอย่างชัดเจนว่าโมเดลขนาดเล็ก (YOLOv8n) ไม่เพียงแต่ประหยัดทรัพยากร แต่ยังให้ความแม่นยำโดยรวมสูงกว่าโมเดลขนาดใหญ่กว่าถึง 4 รุ่น ซึ่งเป็นข้อค้นพบที่มีประโยชน์สำหรับการออกแบบระบบ Edge AI หรือก็คือการนำปัญญาประดิษฐ์ (AI) มาประมวลผลบนอุปกรณ์ปลายทางในงานรีไซเคิล

2. ทฤษฎีที่เกี่ยวข้อง

2.1 ประเภทของพลาสติกและความสำคัญของการรีไซเคิลพลาสติก

ปัจจุบันพลาสติก ถือว่าเป็นวัสดุสังเคราะห์ที่มีบทบาทที่สำคัญอย่างมากในยุคปัจจุบัน เพราะมีต้นทุนการผลิตที่ถูก และสามารถใช้ทดแทนวัสดุอื่นๆ ได้ จึงได้รับความนิยมอย่างมากในการนำมาใช้งาน ดังนั้นเมื่อมีการใช้งานพลาสติกอย่างกว้างขวางและมากขึ้น จึงทำให้เกิดขยะพลาสติกจำนวนมาก และเนื่องจากพลาสติกแต่ละชนิดมีคุณสมบัติทางเคมีและกายภาพที่แตกต่างกันออกไป และพลาสติกที่ใช้กันทั่วไปโดยปกติแล้วจะมีอยู่ 7 ประเภทหลัก โดยจะมีสัญลักษณ์ตัวเลขระบุไว้ที่บรรจุภัณฑ์ [4] ดังนี้ PET, HDPE, PVC, LDPE, PP, PS และ Other การรีไซเคิลไม่ได้เพียงแค่ลดปริมาณขยะพลาสติก แต่ยังยังเป็นการใช้ทรัพยากรที่มีอยู่อย่างจำกัด ให้เกิดความคุ้มค่าอีกด้วย ดังนั้นการรีไซเคิลพลาสติกจึงมีความสำคัญดังต่อไปนี้ 1) ช่วยลดการใช้ทรัพยากรธรรมชาติ เนื่องจากการผลิตพลาสติกขึ้นมาใหม่ จะต้องใช้วัตถุดิบจากปิโตรเลียม ซึ่งก็คือน้ำมันดิบและก๊าซธรรมชาติ วัตถุดิบทั้งสองอย่างนั้นเป็นทรัพยากรที่ใช้แล้วหมดไป การนำพลาสติกที่มีอยู่กลับมาใช้ใหม่ช่วยลดความจำเป็นในการขุดเจาะและแปรรูปวัตถุดิบใหม่ได้ 2) ในกระบวนการผลิตพลาสติกขึ้นมาใหม่นั้น จะก่อให้เกิดการปล่อยก๊าซเรือนกระจกปริมาณมาก ซึ่งการรีไซเคิลแทนการผลิตขึ้นมาใหม่นั้น จะใช้พลังงานน้อยกว่าถึง 50-70% ดังนั้นการรีไซเคิลพลาสติก จึงช่วยลดมลพิษทางอากาศและน้ำเสียที่เกิดจากกระบวนการผลิตและการกำจัดพลาสติก 3) ช่วยสร้างมูลค่าทางเศรษฐกิจ ขยะพลาสติกที่ถูกคัดแยกอย่างถูกต้องสามารถกลายเป็นวัตถุดิบที่มีคุณภาพสูงและมีมูลค่าเพิ่มขึ้น การรีไซเคิลจึงช่วยให้ภาคอุตสาหกรรมสามารถดำเนินงานเกี่ยวกับการเก็บรวบรวม คัดแยก และแปรรูปพลาสติกสร้างงานและรายได้ให้กับผู้คนในหลากหลายภาคส่วน

2.2 เทคโนโลยีการคัดแยกประเภทพลาสติกในปัจจุบัน

เทคโนโลยีการคัดแยกประเภทพลาสติกในปัจจุบันมีความก้าวหน้าไปมาก โดยมีการใช้เทคนิคหลากหลายเพื่อจัดการกับปริมาณขยะพลาสติกที่เพิ่มขึ้น โดยมีรายละเอียดดังนี้

1. การคัดแยกด้วยมือ เป็นวิธีดั้งเดิมที่ทำกันมาอย่างเนิ่นนาน โดยพนักงานคัดขยะบนสายพานจะต้องทำการคัดพลาสติกด้วยเกณฑ์ของโรงงาน เช่น เลือกตามสี รูปร่าง โลโก้ หรือรหัสรีไซเคิล ซึ่งการคัดแยกด้วยมือมี ประสิทธิภาพการคัดแยกอยู่ที่ประมาณ 60 ขึ้นต่อชั่วโมง [7]

2. การคัดแยกด้วยเครื่องจักรกลและระบบกลไก เช่น การรีไซเคิลเชิงกล โดยนำพลาสติกที่ใช้แล้วมาคัดแยกตามประเภท โดยนำมาล้างทำความสะอาดก่อนที่จะนำไปบดและหลอมใหม่เพื่อผลิตเป็นเม็ดพลาสติกกรีไซเคิลก่อนนำไปผ่านกระบวนการขึ้นรูป [5] และการคัดแยกด้วยเครื่องตรวจจับ NIR ด้วยการวัดการสะท้อนแสงปริมาณ 0.8–2.5 μm เพื่อจดจำโครงสร้างโพลิเมอร์ของพลาสติก โดยนำบรรจุภัณฑ์พลาสติกวางบนสายพานผ่านเซ็นเซอร์ แล้วยิงลมอากาศเล็งเป้าหมายให้ลอยไป

ยังสายพานแยกตามประเภท จากงานวิจัยเปรียบเทียบความแม่นยำ พบว่าการคัดแยกด้วยเครื่องจักรอัตโนมัติ นั้น ให้ความแม่นยำกว่าการคัดแยกด้วยมือประมาณ 30% [8]

2.3 การเรียนรู้เชิงลึก (Deep Learning) และการประยุกต์ใช้ในประมวลผลภาพ

การเรียนรู้เชิงลึก (Deep Learning) เป็นส่วนหนึ่งในแขนงของการเรียนรู้ของเครื่อง (Machine Learning) ที่เน้นการสร้างโมเดลด้วยโครงข่ายประสาทเทียมหลายชั้น (Deep Neural Networks) โดยเฉพาะอย่างยิ่งโครงข่ายประสาทเทียมแบบคอนโวลูชัน (Convolutional Neural Network: CNN) ซึ่งเหมาะสมสำหรับงานด้านการประมวลผลข้อมูลจากภาพ เช่น การใช้ CNN เพื่อดึงคุณลักษณะ (feature extraction) จากภาพหรือวิดีโอเพื่อที่จะได้ ขอบ (edges), รูปร่าง (shapes) และวัตถุที่ซับซ้อน [9] ดังนั้น CNN จึงมีจุดเด่นอย่างมากในการเรียนรู้ลักษณะเฉพาะของวัตถุในภาพได้โดยอัตโนมัติ จากการวิเคราะห์คุณลักษณะแบบลำดับชั้น ในการประยุกต์ใช้การเรียนรู้เชิงลึกด้วยการใช้ CNN สำหรับการตรวจจับวัตถุ คือการใช้ CNN ที่สามารถเรียนรู้คุณลักษณะของภาพ ได้โดยอัตโนมัติมาทำการประมวลผลข้อมูลภาพจากระดับพิกเซล ผ่านหลายเลเยอร์ (เพื่อสร้างคุณลักษณะที่ซับซ้อนขึ้นเรื่อยๆ ซึ่งช่วยให้ระบบสามารถจำแนกประเภท และระบุตำแหน่งของวัตถุภายในภาพได้พร้อมกัน ซึ่งในอดีตงานจำแนกและงานตรวจจับ จะทำแยกกันและการนำ CNN มาประยุกต์ใช้นั้น จะช่วยให้

สามารถรวมสองฟังก์ชันนี้เข้าด้วยกันได้ในระบบเดียว และทำให้การตรวจจับวัตถุมีความแม่นยำและรวดเร็วยิ่งขึ้น

2.4 อัลกอริทึม YOLO (You Only Look Once)

อัลกอริทึม YOLO (You Only Look Once) ถูกนำเสนอครั้งแรกโดย Joseph Redmon [12] ในปี ค.ศ. 2016 โดยมีแนวคิดหลักคือการตรวจจับวัตถุจากภาพทั้งภาพในครั้งเดียว (Single-stage Detection) แทนที่วิธีเดิมที่ต้องแบ่งการทำงานออกเป็นหลายขั้นตอน ทำให้มีความเร็วในการประมวลผลสูงและเหมาะสมสำหรับงาน Real-time กระบวนการทำงานของ YOLO ประกอบด้วยขั้นตอนหลัก 6 ขั้นตอน ดังรูปที่ 1 และมีรายละเอียดดังนี้ [14]



รูปที่ 1. แสดงกระบวนการทำงานของอัลกอริทึม YOLO

1) นำเข้าภาพ (Image Input) รับภาพขนาดมาตรฐาน (เช่น 640×640 พิกเซล) เพื่อให้การประมวลผลมีความสม่ำเสมอของข้อมูล 2) การสกัดคุณลักษณะด้วย CNN (Feature Extraction) ภาพถูกประมวลผลผ่านโครงข่าย CNN ที่แบ่งออกเป็น 'ฟิลเตอร์' ทำหน้าที่ตรวจจับคุณลักษณะของภาพ เช่น เส้น ขอบ สี และรูปร่าง เพื่อสร้าง Feature Map 3) การแบ่งภาพเป็นกริด (Grid Division) YOLO แบ่งภาพออกเป็นกริด เช่น 7×7 หรือ 13×13 โดยแต่ละเซลล์ในกริดจะรับผิดชอบการทำนายวัตถุที่อยู่ในพื้นที่นั้น 4) การทำนาย Bounding Box และ Confidence

Score แต่ละเซลล์ในกริดทำนายกรอบล้อมรอบวัตถุ (Bounding Box) พร้อมค่าความน่าจะเป็นที่มีวัตถุอยู่ในกรอบนั้น (Confidence Score) และประเภทของวัตถุ 5) Non-Maximum Suppression (NMS) ลบกรอบที่ซ้อนทับกัน โดยเก็บเฉพาะกรอบที่มีค่า Confidence Score สูงสุดและตัดกรอบที่มีค่า IoU (Intersection over Union) เกินค่าที่กำหนดออก 6) ผลลัพธ์ (Output) แสดงภาพพร้อม Bounding Box ระบุประเภทวัตถุและค่าความมั่นใจในการทำนาย

2.5 YOLOv8 และความแตกต่างของแต่ละรุ่น

YOLOv8 เป็นส่วนหนึ่งของการพัฒนาการของตระกูล YOLO ที่มีวิวัฒนาการต่อเนื่องตั้งแต่ YOLOv1 จนถึง YOLOv10 [13] โดยการพัฒนาการของ YOLO แต่ละรุ่นและการประยุกต์ใช้ในงานด้านต่างๆ พบว่าแต่ละรุ่นมีการปรับปรุงสถาปัตยกรรมเพื่อเพิ่มทั้งความเร็วและความแม่นยำอย่างต่อเนื่อง YOLOv8 มีให้เลือก 5 รุ่นตามขนาดและความสามารถ ดังแสดงในตารางที่ 1

ตารางที่ 1. ความแตกต่างของ YOLOv8 แต่ละรุ่น

รุ่น	Params	FLOPs	Depth	Width	จุดเด่น	เหมาะกับ
YOLOv8n (Nano)	1.44M	3.3G	ต่ำ	ต่ำ	เร็ว ประหยัดทรัพยากร	Edge device, IoT
YOLOv8s (Small)	5.08M	12.5G	ปานกลาง-ต่ำ	ปานกลาง-ต่ำ	สมดุลระหว่างความเร็วและ	Mobile / embedded
YOLOv8m (Medium)	15.77M	41.6G	ปานกลาง	ปานกลาง	ประสิทธิภาพดี สำหรับงานทั่วไป	Desktop / server
YOLOv8l (Large)	36.19M	98.7G	สูง	สูง	ความแม่นยำสูง ช้ากว่า n/s	งานที่ต้องการความละเอียด
YOLOv8x (XLarge)	56.13M	153.8G	สูงมาก	สูงมาก	แม่นยำสูงสุด ใช้ทรัพยากรมาก	High-end server

2.6 งานวิจัยที่เกี่ยวข้อง

งานวิจัยของ ภัทธมน หุ่นลำพูน [11] ได้ศึกษาการประยุกต์ใช้เทคนิค Convolutional Neural Network ในการจำแนกขยะฝังทะเล ซึ่งแสดงให้เห็นว่า CNN มีศักยภาพในการจำแนกประเภทขยะในสภาพแวดล้อมจริงได้อย่าง

มีประสิทธิภาพ และเป็นพื้นฐานสำคัญของอัลกอริทึม YOLO ที่ใช้ในงานวิจัยนี้ งานวิจัยของพงศันฤทธิ์ เลหาวิเชียร [15] ได้ประยุกต์ใช้ YOLOv8 ในการตรวจจับทางม้าลายที่เลื่อนหาย โดยใช้เทคนิค Conditional Classification ซึ่งแสดงให้เห็นถึงความยืดหยุ่นของ YOLOv8 ในการนำไปประยุกต์ใช้กับงานจำแนกประเภทในบริบทที่หลากหลาย นอกเหนือจากงานตรวจจับวัตถุทั่วไป งานวิจัยของธานิต ม่วงพลู และ วริยา เข็นเปิง [16] ได้ทำการศึกษาหัวข้อ การพัฒนาระบบคัดแยกประเภทขยะรีไซเคิลด้วยเทคโนโลยีไอโอที งานวิจัยนี้นำเสนอระบบคัดแยกขยะรีไซเคิลที่ใช้เทคโนโลยี IoT ในการควบคุมการเปิด-ปิดถังขยะ คัดแยกขวดโลหะและขวดพลาสติก พร้อมทั้งบอกจำนวนขยะที่รับเข้าและแสดงพิกัดของถังขยะผ่านแอปพลิเคชัน LINE จากการทดสอบจำนวน 10 รอบพบว่าระบบสามารถคัดแยกประเภทพลาสติกได้พร้อมกับระบุจำนวน และพิกัดผ่านโทรศัพท์มือถือได้ถูกต้อง งานวิจัยของ บพิตร ไชยนอก และ ฤชานนท์ ศรีราชวงศ์ [17] ก็ศึกษาเกี่ยวกับพัฒนาระบบคัดแยกประเภทขวดน้ำอัดโนมิติ โดยใช้เทคโนโลยีการประมวลผลภาพและการเรียนรู้ของเครื่องเพื่อแก้ปัญหาการใช้แรงงานคนในการคัดแยกขยะ ใช้เทคนิค Teachable machine ในการคัดแยกประเภทของขวด ผลการทดลองพบว่าสามารถคัดแยกประเภทขวดน้ำสามารถคัดแยกประเภทของขวดน้ำได้อย่างแม่นยำ โดยมีอัตราความถูกต้องเฉลี่ยอยู่ที่ประมาณ 95% ระบบสามารถประมวลผลภาพและจำแนกประเภทขวดน้ำได้ภายในเวลาไม่ถึง 1 วินาทีต่อขวด ซึ่งผลลัพธ์ที่กล่าวอยู่ในเกณฑ์ที่มีประสิทธิภาพสูง ส่วนการศึกษาของวณิดา วรรณสวัสดิ์กุล [18] งานวิจัยนี้เป็นการสร้างเว็บแอปพลิเคชันสำหรับการให้ข้อมูลคัดแยกขยะซึ่งสามารถใช้งานได้ทั้งคอมพิวเตอร์และสมาร์ตโฟน โดยได้นำเทคโนโลยีปัญญาประดิษฐ์ มาใช้วิเคราะห์ภาพขยะเพื่อให้ได้ประเภทขยะที่ถูกต้อง และจะแสดงผลลัพธ์ที่หน้าจออุปกรณ์ โครงการวิจัยนี้ได้นำ Teachable Machine มาใช้เป็นโมเดลของปัญญาประดิษฐ์ของขยะจำนวน 4 ประเภท โดยแบ่งข้อมูลออกเป็น 12 คลาส โมเดลปัญญาประดิษฐ์ที่ได้มีความถูกต้องร้อยละ 85.33 จากนั้นใช้เซตที่เอ็มแอล ร่วมกับจาวาสคริปต์ในการพัฒนาแอปพลิเคชัน เว็บไซต์ที่พัฒนาสามารถใช้งานแสดงประเภทขยะได้อย่างถูกต้องร้อยละ 84.67

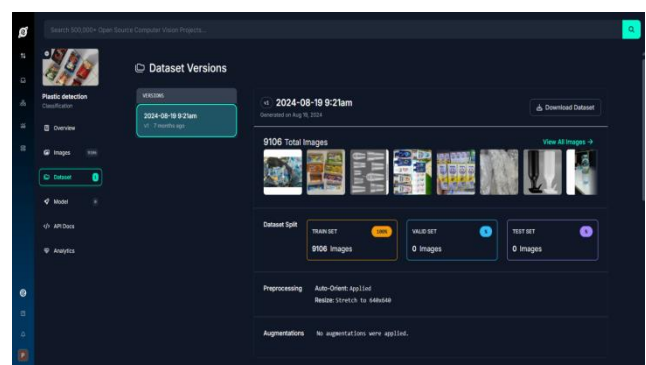
3. วิธีการดำเนินงานวิจัยและผลการวิจัย

ในการพัฒนาและเปรียบเทียบประสิทธิภาพของโมเดล YOLOv8 ในการคัดแยกประเภทของพลาสติกสำหรับการรีไซเคิลด้วยการประมวลผลจากภาพถ่ายดิจิทัลนี้ สามารถแบ่งวิธีการดำเนินงานได้ดังนี้

3.1. การรวบรวมข้อมูล

3.1.1 การรวบรวมรูปภาพพลาสติกชนิดต่างๆ ซึ่งในขั้นตอนของการรวบรวมรูปภาพพลาสติกชนิดต่างๆ เพื่อที่จะนำมาใช้เป็นข้อมูล (Dataset) นั้น ผู้วิจัยได้รวบรวม ผ่านแพลตฟอร์มจัดการและเทรนโมเดล Computer Vision โดยช่วยในการสร้างและจัดการชุดข้อมูลสำหรับการทำวิจัยประเภท Computer Vision ชื่อว่า “Roboflow” ซึ่งช่วยในการรวบรวมรูปภาพปริมาณมากและแบ่งออกเป็นประเภทต่างๆ ไว้ให้เรียบร้อยแล้ว ดังรูปที่ 2 รูปภาพพลาสติกที่สกปรก เปียก หรืออยู่ปะปนกับพื้นผิวอื่นๆจากแหล่งภายนอกอีกจำนวน 50 ภาพ รวม 9156 ภาพ

3.1.2 การแบ่งชุดข้อมูลในสัดส่วน Train 70%, Validation 15% และ Test 15% นั้น สอดคล้องกับแนวทางการแบ่งอัตราส่วนชุดข้อมูลฝึกฝน/ทดสอบต่อประสิทธิภาพของโมเดล [19] โดยแบ่งชุดสำหรับการฝึกฝน (Train) จำนวน 6410 ภาพ ชุดสำหรับการตรวจสอบ (Validate) จำนวน 1373 ภาพ ชุดสำหรับการทดสอบ (Test) จำนวน 1373 ภาพ ซึ่งแต่ละชุดข้อมูลจะแบ่งออกเป็น 8 คลาส ก็คือ PET, HDPE, PVC, LDPE, PP, PS, Single use plastic, Multilayer plastic ภาพเพื่อนำไปสร้างโมเดล



รูปที่ 2. แพลตฟอร์ม Roboflow

3.2 เครื่องมือที่ใช้วิจัย

ในการสร้างโมเดล (Model) ผู้วิจัยใช้โปรแกรม Google Colab ในการสร้างและฝึกโมเดลผ่านเว็บเบราว์เซอร์ โดยใช้ GPU รุ่น T4 จาก Google Colab ฝึกโมเดล YOLOv8 ทั้ง 5 รุ่น ด้วยพารามิเตอร์การฝึกดังนี้ จำนวน Epochs = 100, Learning Rate = 0.01 (SGD), Batch Size = 16, Image Size = 640×640 พิกเซล, และใช้ Pre-trained Weights จาก ImageNet เป็นจุดเริ่มต้น (Transfer Learning)

3.3 การหาประสิทธิภาพของโมเดล

3.3.1 การประเมินผลโมเดลใช้ Confusion Matrix เป็นเครื่องมือหลัก โดยมีค่าตัวชี้วัด 4 ค่าหลัก ดังนี้

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

หมายถึง สัดส่วนของตัวอย่างทั้งหมดที่โมเดลทำนายได้ถูกต้อง ทั้งคลาสบวกและคลาสลบ ค่าสูงแสดงว่าโมเดลมีความถูกต้องโดยรวมสูง

$$\text{Precision} = TP / (TP + FP)$$

หมายถึง ในบรรดาที่โมเดลทำนายว่าเป็นคลาสนั้น มีเท่าไรที่ถูกต้องจริง ค่าสูงหมายความว่าโมเดลไม่ค่อยทำนายผิดพลาดแบบ False Positive

$$\text{Recall} = TP / (TP + FN)$$

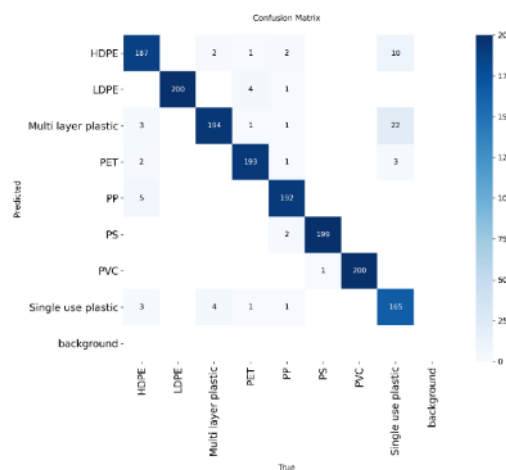
หมายถึง ในบรรดาตัวอย่างที่เป็นคลาสนั้นจริงๆ โมเดลจำแนกได้ถูกต้องเท่าไร ค่าสูงหมายความว่าโมเดลสามารถตรวจจับได้ครบคลุม

$$\text{F1-score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})}$$

หมายถึง ค่าเฉลี่ยระหว่าง Precision และ Recall เหมาะใช้เมื่อชุดข้อมูลมีความไม่สมดุลระหว่างคลาส เนื่องจากบาลานซ์ระหว่าง Precision และ Recall โดยที่ TP = True Positive (ทำนายถูกว่าเป็นคลาสนั้น), TN = True Negative (ทำนายถูกว่าไม่ใช่คลาสนั้น), FP = False Positive (ทำนายผิดว่าเป็นคลาสนั้น), FN = False Negative (ทำนายผิดว่าไม่ใช่คลาสนั้น)

3.3.2 การแปลงและตรวจสอบค่าจาก Confusion Matrix ผู้วิจัยใช้ “EasyOCR” ซึ่งเป็นเครื่องมือ Optical Character

Recognition ที่ได้พิสูจน์ว่ามีประสิทธิภาพสูงในการรู้จำตัวอักษรและตัวเลขจากภาพในรูปแบบต่างๆ โดยรองรับหลายภาษาและให้ความแม่นยำในระดับที่ยอมรับได้สำหรับการนำไปใช้งานจริง [20] ในการสกัดค่าตัวเลขจากภาพ โดยเหตุที่ต้องใช้ EasyOCR ในการสกัดข้อมูลจาก Confusion Matrix นั้นมาจากข้อจำกัดของ YOLOv8 Classification API บน Google Colab ในขณะทำการวิจัย ซึ่งไม่รองรับการ Export Confusion Matrix ในรูปแบบไฟล์ตัวเลข (.csv หรือ .json) โดยตรง แต่รองรับเฉพาะการบันทึกเป็นไฟล์ภาพ ดังรูปที่ 3 การใช้ OCR จึงเป็นแนวทางที่สมเหตุสมผลภายใต้ข้อจำกัดดังกล่าว



รูปที่ 3. Confusion Matrix ที่ได้จากการ Validate

3.3.3 การนำค่าตัวเลขที่สกัดได้จากการทำ OCR ไปประมวลผลต่อเพื่อคำนวณหาประสิทธิภาพของโมเดล ซึ่งแบ่งออกเป็น 4 ค่า Accuracy, Precision, Recall, F1-Score ดังรูปที่ 4 เพื่อนำค่าที่ได้ไปใช้สำหรับเปรียบเทียบประสิทธิภาพของโมเดล YOLOv8 รุ่นต่างๆ

```
Precision per class: [ 0.26076 0.40486 0.5452 0.95673 0.0036364 0.0024814]
Recall per class: [ 0.49751 0.26882 0.58663 0.25911 0.0044643 0.029412]
F1 per class: [ 0.34217 0.3231 0.56515 0.40779 0.004088 0.0045767]
Accuracy per class: [ 0.69252 0.66493 0.88125 0.76889 0.00128 0.02607]
Overall Precision: 0.36227679251883593
Overall Recall: 0.2743244826138585
Overall Accuracy: 0.7724910035985606
Overall F1: 0.274466646946349
```

รูปที่ 4. ผลลัพธ์ที่ได้จากการคำนวณค่าที่ได้จาก OCR

3.4 การเปรียบเทียบค่าประสิทธิภาพและวิเคราะห์ผล

ใช้ตัวชี้วัดการประเมินผล ได้แก่ จำนวนพารามิเตอร์ของโมเดล, FLOPs, Accuracy, Precision, Recall, F1-score, Top-1 accuracy และ Top-5 accuracy เปรียบเทียบค่าประสิทธิภาพกับโมเดล YOLOv8 ทุกรุ่น

3.5 ผลการวิจัย

ในการประเมินผลประสิทธิภาพของระบบการคัดแยกประเภทของพลาสติกสำหรับการรีไซเคิลจากนี้จะสามารถแบ่งการทดสอบได้ออกเป็น 2 การทดสอบดังต่อไปนี้

3.5.1 การทดสอบการประเมินด้วยค่าตัวชี้วัด (Validation) จะให้ผลลัพธ์ดังตารางที่ 2

ตารางที่ 2. ผลการทดลองหลังจากทำการประเมินประสิทธิภาพ

Method	Params	FLOPs	Accuracy	Precision	Recall	F1 score	Top-1	Top-5
YOLOv8n	1.44m	3.3g	0.77	0.36	0.27	0.27	0.96	1
YOLOv8m	15.77m	41.6g	0.69	0.23	0.21	0.19	0.99	1
YOLOv8s	56.13m	153.8g	0.69	0.23	0.18	0.19	0.99	1
YOLOv8l	36.19m	98.7g	0.68	0.21	0.20	0.18	0.99	1
YOLOv8x	5.08m	12.5g	0.71	0.15	0.16	0.11	0.98	1

ผลการศึกษาได้แสดงค่า Accuracy, Precision, Recall, F1-score, Top-1 accuracy และ Top-5 accuracy ของโมเดล YOLOv8 ในแต่ละรุ่น สรุปผลการเปรียบเทียบประสิทธิภาพดังนี้ 1) ด้านความถูกต้อง (Accuracy) YOLOv8n ให้ค่า Accuracy สูงที่สุดที่ 0.77 สูงกว่า YOLOv8s (0.71), YOLOv8m (0.69), YOLOv8l (0.68) และ YOLOv8x (0.69) อย่างชัดเจน แสดงให้เห็นว่าการเพิ่มขนาดโมเดลและจำนวนพารามิเตอร์ไม่ได้ช่วยเพิ่มความถูกต้องโดยรวม ซึ่งอาจเกิดจากปัญหา ปัญหาที่โมเดล เรียนรู้ข้อมูลชุดฝึกสอน (Training Data) มากเกินไปจนจดจำรายละเอียดที่ลึกเกินไป (Overfitting) ในโมเดลขนาดใหญ่เมื่อชุดข้อมูลมีขนาดจำกัด 2) ด้าน Precision และ Recall YOLOv8n มีค่า Precision สูงสุดที่ 0.36 และ Recall สูงสุดที่ 0.27 ขณะที่รุ่นอื่นๆ อยู่ในช่วง 0.15 - 0.23 และ 0.16 - 0.21 ตามลำดับ ค่าที่ค่อนข้างต่ำในทุกๆ รุ่นบ่งชี้ว่าโมเดลยังมีความยากลำบากในการจำแนกพลาสติกบางคลาสที่มีลักษณะคล้ายกัน โดยเฉพาะ HDPE ซึ่งมีค่า Precision ต่ำที่สุดในหลายรุ่น 3) ด้าน F1-score YOLOv8n ให้ค่า F1-score สูงที่สุดที่ 0.27 สอดคล้องกับ

Precision และ Recall สูงกว่า YOLOv8s (0.11), YOLOv8m (0.19), YOLOv8l (0.18) และ YOLOv8x (0.19) อย่างชัดเจน 4) ด้าน Top-1 และ Top-5 Accuracy ทุกโมเดลมีค่า Top-5 Accuracy = 1.0 หมายความว่าผลลัพธ์ที่ถูกต้องปรากฏใน 5 อันดับแรกเสมอ ส่วน Top-1 Accuracy ของทุกโมเดลใกล้เคียงกันที่ 0.96 - 0.99 แสดงว่าโมเดลทุกตัวสามารถจำแนกพลาสติกประเภทที่ถูกต้องในอันดับแรกได้ดี

5) เวลาที่ใช้ในการคำนวณ (Computational Complexity) ก็เป็นปัจจัยสำคัญ ประเมินได้จาก 3 ด้านหลัก ดังนี้ 1) FLOPs YOLOv8n ใช้ 3.3G FLOPs ต่อการ Inference หนึ่งครั้ง ขณะที่ YOLOv8x ใช้สูงถึง 153.8G FLOPs หรือมากกว่าถึง 46 เท่า 2) จำนวนพารามิเตอร์ YOLOv8n มีเพียง 1.44M พารามิเตอร์เหมาะสำหรับการ Deploy บน Edge Device หรืออุปกรณ์ IoT ซึ่งหากนำไปใช้งานบน CPU หรืออุปกรณ์ Edge ความต่างนี้จะยิ่งเด่นชัดขึ้นอย่างมาก

3.5.2 การประเมินด้วยผลการทดสอบจากการนำเข้ารูปภาพนำเข้าภาพที่มีองค์ประกอบของวัตถุหลากหลายชนิด เช่น พลาสติกที่สกปรก เปียก หรืออยู่ปะปนกับพื้นผิวอื่นๆ โดยพลาสติกส่วนใหญ่ในภาพนั้นจะเป็นพลาสติกชนิดต่างๆ 7 ประเภท แล้วทำการทดลองหาผลการทำนายจากภาพว่าโมเดล YOLOv8 แต่ละรุ่นจะแสดงผลค่าความมั่นใจแค่ไหนว่า การทำนายนั้นถูกต้อง ซึ่งค่านี้จะอยู่ระหว่าง 0 ถึง 1 ยิ่งใกล้ 1 ยิ่งแม่นยำ โดยจะแสดงผลเป็นกล่องล้อมรอบวัตถุ พร้อมระบุชนิดของพลาสติกและค่าความแม่นยำที่โมเดลทำนายได้ จะให้ผลลัพธ์ดังตารางที่ 3

ตารางที่ 3. ผลการเปรียบเทียบค่าความแม่นยำของแต่ละโมเดลที่ได้ประมวลผลจากภาพพลาสติกแต่ละชนิด

Method	YOLOv8n	YOLOv8s	YOLOv8m	YOLOv8l	YOLOv8x
PET	0.84	0.90	1.00	0.97	0.99
HDPE	0.87	0.21	0.64	0.14	0.59
PVC	0.78	0.51	0.99	0.84	0.97
LDPE	0.87	0.89	0.99	1.00	1.00
PP	0.96	0.99	1.00	1.00	0.99
PS	0.82	0.92	0.91	0.99	0.99
O	0.95	0.87	0.26	0.82	0.91
Avg.	0.87	0.76	0.83	0.82	0.92

จากผลการทดสอบด้วยภาพจริง YOLOv8x มีค่าเฉลี่ย Confidence Score สูงสุดที่ 0.92 ตามด้วย YOLOv8n ที่ 0.87

ขณะที่ YOLOv8m (0.83), YOLOv8l (0.82) และ YOLOv8s (0.76) มีค่าต่ำกว่าค่าเฉลี่ยรวม (0.84) อย่างไรก็ตาม เมื่อพิจารณา ร่วมกับผลจากตัวชี้วัดในหัวข้อ 3.5.1 พบว่า YOLOv8n มีประสิทธิภาพโดยรวมสูงสุด เนื่องจากความต่างของค่าเฉลี่ย Confidence Score ระหว่าง YOLOv8n และ YOLOv8x มีเพียง 0.05 ในขณะที่ YOLOv8n มีความได้เปรียบด้านทรัพยากรอย่าง มีนัยสำคัญ

การวิเคราะห์ Confusion Matrix และกรณี Misclassification จากการวิเคราะห์ Confusion Matrix ของโมเดล YOLOv8n ซึ่งให้ประสิทธิภาพสูงสุด พบรูปแบบการจำแนกผิด (Misclassification) ที่น่าสนใจดังนี้ 1) HDPE ถูกจำแนกผิดเป็น PP และ LDPE บ่อยที่สุด เนื่องจากทั้งสามประเภทมีลักษณะ พื้นผิวและสีที่คล้ายคลึงกัน โดยเฉพาะเมื่ออยู่ในสภาพสกปรก หรือเปียก ทำให้คุณลักษณะที่โมเดลสกัดได้มีความซ้อนทับกัน 2) PVC ถูกจำแนกผิดเป็น LDPE ในบางกรณี เนื่องจากบรรจ ุภัณฑ์บางประเภทของสองชนิดนี้มีความโปร่งใสคล้ายกัน 3) Other (พลาสติกกรรม) มีค่า Precision และ Recall ต่ำเป็น พิเศษใน YOLOv8m เนื่องจากคลาสนี้ครอบคลุมพลาสติก หลากหลายชนิดที่ไม่อยู่ในกลุ่มหลัก 6 ประเภท ทำให้โมเดล ขาดต่อการเรียนรู้ลักษณะที่ชัดเจน

4. บทสรุป อภิปราย และข้อเสนอแนะ

ในการสรุปผลงานวิจัย ผู้วิจัยได้ทำสรุปผลตาม วัตถุประสงค์ของงานวิจัยในบทที่ 1 ดังนี้ สำหรับการพัฒนาระบบคัดแยกประเภทของขยะพลาสติกโดยใช้ปัญญาประดิษฐ์ (AI) และเทคนิคการเรียนรู้เชิงลึกจากภาพถ่ายนั้น ผู้วิจัยได้ทำ การวิเคราะห์ว่าจะนำโมเดลไหนมาใช้ จากที่ได้ศึกษาและ เรียนรู้มาพบว่าโมเดล YOLOv8 เป็นโมเดลที่การตรวจจับหรือ คัดแยกประเภทวัตถุที่มีประสิทธิภาพสูง และสามารถใช้ในการ คัดแยกประเภทของพลาสติกได้อย่างแม่นยำ

จากผลการทดสอบทั้ง 2 วิธี พบความแตกต่างในการจัด อันดับโมเดลที่ดีที่สุด กล่าวคือ การทดสอบด้วยค่าตัวชี้วัด (หัวข้อ 3.5.1) ชี้ว่า YOLOv8n ดีที่สุด ขณะที่การทดสอบด้วย Confidence Score จากภาพจริง (หัวข้อ 3.5.2) ชี้ว่า YOLOv8x ดี ที่สุด ซึ่งความไม่สอดคล้องนี้อธิบายได้ดังนี้ Accuracy จาก Validation Set สะท้อนความสามารถในการจำแนกประเภท พลาสติกโดยรวมทั้งชุดข้อมูล ครอบคลุมทุกคลาสและทุก

สภาพภาพ จึงเป็นตัวชี้วัดที่มีความน่าเชื่อถือสูงกว่า Confidence Score จากภาพตัวอย่าง เป็นการทดสอบเฉพาะภาพที่ผู้วิจัยเลือก มา จำนวน 7 ภาพต่อคลาส ซึ่งอาจไม่ครอบคลุมความ หลากหลายของชุดข้อมูลทั้งหมด โมเดลขนาดใหญ่อย่าง YOLOv8x อาจให้ค่า Confidence สูงในภาพที่เลือกมา แต่ ครอบคลุมได้น้อยกว่าเมื่อทดสอบกับชุดข้อมูลทั้งหมดดังนั้น งานวิจัยนี้จึงกำหนดเกณฑ์การตัดสินใจโดยให้น้ำหนักกับ ค่า ตัวชี้วัดจาก Validation Set เป็นหลัก โดยพิจารณาร่วมกับ Model Efficiency (Params และ FLOPs) เป็นเกณฑ์รอง สรุปได้ ว่า YOLOv8n เป็นโมเดลที่เหมาะสมที่สุด สำหรับงานวิจัยนี้ และการนำไปใช้งานจริงสำหรับระบบคัดแยกพลาสติกแบบ Real-time บน Conveyor Belt ที่ต้องการ ความเร็วสูง ดังนั้น “YOLOv8n” จึงมีความเหมาะสมสูงกว่าในเชิงวิศวกรรม ข้อเสนอแนะสำหรับการวิจัยครั้งต่อไปมีดังนี้

1. การจัดเตรียมชุดข้อมูลสำหรับฝึกฝน โดยเพิ่มจำนวน รูปภาพพลาสติกให้มีความหลากหลายคุณลักษณะดังนี้ การเพิ่ม ภาพพลาสติกที่อยู่ในสภาวะแสงที่แตกต่างกันเพื่อให้โมเดล เรียนรู้พลาสติกในมุมมองต่างๆ การถ่ายภาพพลาสติกควรจะ ถ่ายเป็นรูปพลาสติกชนิดนั้น แบบโคเด้นเพื่อให้โมเดลเรียนรู้ คุณลักษณะข้อมูลอย่างมีประสิทธิภาพ หลีกเลี่ยงการถ่ายภาพ พลาสติกหลากหลายชนิดในภาพเดียวกัน

2. การเปรียบเทียบประสิทธิภาพของโมเดล YOLOv8 กับ โมเดล YOLO รุ่นใหม่ ๆ อย่างเช่น EfficientNet, ResNet หรือ YOLOv11/YOLOv12 เพื่อให้เห็นภาพรวมที่ครอบคลุม มากขึ้น

3. การพัฒนาเว็บไซต์ในการจัดทำ Dashboard สำหรับ ติดตามผลการตรวจจับ เช่น สถิติพลาสติกแต่ละประเภทที่ถูก คัดแยกโดยเชื่อมต่อกับระบบกับฐานข้อมูลเพื่อติดตามปริมาณ พลาสติกที่เข้าสู่กระบวนการรีไซเคิลและอาจพัฒนาเป็น API เพื่อให้โรงงานอื่น ๆ หรือผู้ประกอบการสามารถนำไปใช้ใน ระบบที่มีอยู่แล้วได้

4. การจัดหากล้องอุตสาหกรรม (Industrial Camera) ที่มี ความละเอียดสูง 1080p – 4K เพื่อจับรายละเอียดของพลาสติกที่ อยู่ในสภาวะแสงที่แตกต่างกันเพื่อให้โมเดลเรียนรู้พลาสติกใน มุมมองต่างๆ

เอกสารอ้างอิง

- [1] กรมควบคุมมลพิษ, “รายงานสถานการณ์สถานที่กำจัดขยะมูลฝอยชุมชนของประเทศไทย ปี พ.ศ. 2567,” กระทรวงทรัพยากรธรรมชาติและสิ่งแวดล้อม, กรุงเทพฯ, 2567.
- [2] C. L. Liu, P. Bunditsakulchai, and Q. Zhuo, “Impact of COVID-19 on Food and Plastic Waste Generated by Consumers in Bangkok,” *Sustainability*, vol. 13, no. 16, p. 8988, 2021. doi: 10.3390/su13168988.
- [3] G. B. Borongan et al., “Manual on plastics leakage prevention from formal and informal recycling facilities in Nonthaburi, Thailand,” *Asian Institute of Technology; ERIA*, Jul. 4, 2024.
- [4] กรมควบคุมมลพิษ, “คู่มือการจัดการขยะพลาสติกเพื่อการรีไซเคิลอย่างยั่งยืน,” กรุงเทพฯ, 2565.
- [5] R. Geyer et al., “Plastic recycling: Challenges and opportunities,” *Can. J. Chem. Eng.*, 2024.
- [6] J. B. da Costa et al., “Challenges to reducing post-consumer plastic rejects from the MSW sorting system,” *Sustainability*, 2022.
- [7] P. Akkaya, C. Aydin, and R. Yilmaz, “Auto-sorting commonly recovered plastics from waste household appliances and electronics using near-infrared spectroscopy,” *J. Cleaner Prod.*, vol. 254, p. 120153, 2020. doi: 10.1016/j.jclepro.2020.120153.
- [8] E. M. Urban and K. Ragaert, “A review of sorting and separating technologies suitable for compostable and biodegradable plastic packaging,” *Front. Sustain.*, vol. 3, Art. no. 901885, 2022. doi: 10.3389/frsus.2022.901885.
- [9] C. Lubongo and P. Alexandridis, “Assessment of performance and challenges in use of commercial automated sorting technology for plastic waste,” *Recycling*, vol. 7, no. 2, p. 11, 2022. doi: 10.3390/recycling7020011.
- [10] A. R. Pathak, M. Pandey, and S. Rautaray, “Application of deep learning for object detection,” *Procedia Comput. Sci.*, vol. 132, pp. 1706–1717, 2018. doi: 10.1016/j.procs.2018.05.102.
- [11] ภัทรมน หุ่นลำพูน, “การพัฒนาและประยุกต์ใช้เทคนิค Convolutional Neural Network ในการจำแนกขยะชายฝั่งทะเล,” *วิทยานิพนธ์มหาบัณฑิต*, ม.บูรพา, 2566. [Online]. Available: http://digital_collect.lib.buu.ac.th/dcms/files/62910162.pdf.
- [12] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, real-time object detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 779–788, 2016. [Online]. Available: https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Redmon_You_Only_Look_CVPR_2016_paper.pdf.
- [13] M. A. R. Alif, M. S. Islam, S. Rahman, and M. N. Islam, “YOLOv1 to YOLOv10: A comprehensive review of YOLO variants and their application in the agricultural domain,” *arXiv*, 2024. doi: 10.48550/arXiv.2406.10139.
- [14] N. Wiangkarn and S. Jiriwibhakorn, “Comparison of YOLOv8 Models for Aircraft Detection in Airport Apron Using Digital Image Processing,” *Eng. Technol. Horiz.*, vol. 41, no. 3, p. 410309, 2024. doi: 10.14456/eth.2024.9.
- [15] พ. เลาหวิเชียร, ว. จาริวงษ์ไพฑูริย์, และ ส. วรรณภา, “การตรวจจับทางข้ามที่ชัดเจนโดยการใช้อัลกอริทึมไฮโด, ในการประชุมวิชาการระดับชาติ Rangsit Graduate Research Conference (RGRC), vol. 19, pp. 411–427, 2567.
- [16] ชานิล ม่วงพูล, “การพัฒนาระบบคัดแยกขยะรีไซเคิลด้วยเทคโนโลยีไอโอที,” *วารสารวิจัยและพัฒนา ม.ราชภัฏนครราชสีมา*, vol. 12, no. 2, pp. 45–52, 2563. [Online]. Available: <https://ph02.tci-thaijo.org/index.php/project-journal/article/view/247933>.
- [17] บพิตร ไชยนอก and ฤชานนท์ ศรีราชวงศ์, “การประยุกต์ใช้ Teachable Machine สำหรับระบบคัดแยกขวดน้ำอัดโนมิติ,” *วารสารเทคโนโลยีและวิศวกรรมก้าวหน้า*, vol. 2, no. 1–2, pp. 13–20, 2567. [Online]. Available: <https://ph03.tci-thaijo.org/index.php/JTEP/article/view/323>.
- [18] วดีนาถ วรรณสวัสดิ์กุล, “การพัฒนาเว็บแอปพลิเคชันแยกประเภทขยะด้วยเทคโนโลยีปัญญาประดิษฐ์,” *วารสารวิทยาศาสตร์และเทคโนโลยี ม.ราชภัฏเทพสตรี*, vol. 3, no. 1,

pp. 45–52, 2567. [Online]. Available: <https://ph02.tci-thaijo.org/index.php/JSTNSRU/article/view/250746..>

[19] H. Bichri, A. Chergui, and M. Hain, “Investigating the Impact of Train/Test Split Ratio on the Performance of Pre-Trained Models with Custom Datasets,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 2, pp. 333–340, 2024. doi: 10.14569/IJACSA.2024.0150235.

[20] B. K. Pattanayak, B. B. Dash, and S. S. Patra, “A novel technique for handwritten text recognition using Easy OCR,” *ResearchGate*, 2023. [Online]. Available: https://www.researchgate.net/publication/376355415_A_Novel_Technique_for_Handwritten_Text_Recognition_using_Easy_OCR..

[21] M. Sokolova and G. Lapalme, “A systematic analysis of performance measures for classification tasks,” *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, 2009. doi: 10.1016/j.ipm.2009.03.002.

การพัฒนาระบบช่วยคัดกรองโรคอัลไซเมอร์ระยะเริ่มต้นจากภาพ MRI ด้วย Deep

Learning สำหรับบริบทระบบสาธารณสุขไทย

Development of a deep learning-based early Alzheimer's disease screening system using MRI for the Thai healthcare context

ธัญมัย แพงวงษ์ และ นัตตวุฒิ ศรีวิบูลย์*

Thanyamai Paengwong and Nattavut Sriwiboon *

phanyamai.pa@ksu.ac.th and nattavut.sr@ksu.ac.th *

สาขาวิชาเทคโนโลยีสารสนเทศทางการแพทย์ คณะวิทยาศาสตร์และเทคโนโลยีสุขภาพ

มหาวิทยาลัยกาฬสินธุ์

Department of Medical Information Technology, Faculty of Science and Health Technology,

Kalasin University

Received: March 19, 2026; Revised: June 17, 2026; Accepted: June 18, 2026; Published: June 30, 2026

ABSTRACT – Thailand's transition into an aging society has increased the prevalence of Alzheimer's disease, making early-stage screening essential for improving patient outcomes. This paper proposes a deep learning-based screening system for early detection of Alzheimer's disease using magnetic resonance imaging (MRI), tailored to the context of the Thai public healthcare system. A total of 6,400 MRI images has been used and classified into four categories: Non-Demented, Very Mild Demented, Mild Demented, and Moderate Demented. Data preprocessing and augmentation techniques have been applied to enhance model performance. Four models have been evaluated, including the proposed Custom CNN, MobileNetV2, ResNet50, and VGG16. The results show that the proposed Custom CNN achieves the highest accuracy of 97.12%, with strong generalization and stable training performance. The model has also effectively reduced misclassification in early-stage categories. The proposed system demonstrates strong potential as a practical screening tool for early Alzheimer's detection in Thailand, particularly in primary healthcare settings.

KEY WORDS -- Alzheimer's Disease, Deep Learning, Magnetic Resonance Imaging (MRI), Early Screening, Thai Healthcare System

บทคัดย่อ -- ประเทศไทยได้เข้าสู่สังคมผู้สูงอายุ ส่งผลให้จำนวนผู้ป่วยโรคอัลไซเมอร์เพิ่มขึ้นอย่างต่อเนื่อง การคัดกรองโรคในระยะเริ่มต้นจึงมีความสำคัญต่อการชะลอความรุนแรงของโรค งานวิจัยนี้นำเสนอระบบช่วยคัดกรองโรคอัลไซเมอร์ระยะเริ่มต้นโดยใช้เทคนิคการเรียนรู้เชิงลึกจากภาพถ่ายสมองด้วยคลื่นแม่เหล็กไฟฟ้า (MRI) ซึ่งออกแบบให้เหมาะสมกับบริบทของระบบสาธารณสุขไทย งานวิจัยนี้ได้ใช้ชุดข้อมูลภาพ MRI จำนวน 6,400 ภาพ โดยแบ่งออกเป็น 4 ระดับ ได้แก่ กลุ่มปกติ กลุ่มระยะเริ่มต้นเล็กน้อย กลุ่มระยะเริ่มต้น และกลุ่มระยะปานกลาง โดยมีการเตรียมข้อมูลด้วยเทคนิคการปรับขนาดภาพ การปรับค่ามาตรฐาน และการเพิ่มข้อมูล เพื่อเพิ่มประสิทธิภาพของโมเดล จากนั้นได้ทำการเปรียบเทียบโมเดลจำนวน 4 โมเดล ได้แก่ Custom CNN ที่เสนอ, MobileNetV2, ResNet50 และ VGG16 ผลการทดลองแสดงให้เห็นว่าโมเดล Custom CNN ที่เสนอให้ความแม่นยำสูงสุดที่ 97.12% มีความสามารถในการเรียนรู้ที่มีเสถียรภาพ และลดความสับสนในการจำแนกระยะเริ่มต้นของโรคได้อย่างมีประสิทธิภาพ ระบบที่พัฒนาขึ้นมีศักยภาพในการนำไปใช้เป็นเครื่องมือช่วยคัดกรองโรคอัลไซเมอร์ในระยะเริ่มต้น โดยเฉพาะในระดับบริการสาธารณสุขปฐมภูมิของประเทศไทย

คำสำคัญ – โรคอัลไซเมอร์, การเรียนรู้เชิงลึก, ภาพ MRI, การคัดกรองระยะเริ่มต้น, ระบบสาธารณสุขไทย

1. บทนำ

ประเทศไทยได้เข้าสู่สังคมผู้สูงอายุอย่างสมบูรณ์ ซึ่งส่งผลให้จำนวนผู้ป่วยโรคไม่ติดต่อเรื้อรังเพิ่มขึ้นอย่างต่อเนื่อง [1] โดยเฉพาะ โรคอัลไซเมอร์ซึ่งเป็นโรคทางระบบประสาทที่ส่งผลกระทบต่อความจำ ความคิด และพฤติกรรมของผู้ป่วย โรคดังกล่าวไม่เพียงส่งผลกระทบต่อคุณภาพชีวิตของผู้ป่วยเท่านั้น แต่ยังสร้างภาระให้กับครอบครัวและระบบสาธารณสุขโดยรวมอย่างมีนัยสำคัญ ดังนั้น การคัดกรองและวินิจฉัยโรคในระยะเริ่มต้นจึงมีความสำคัญอย่างยิ่งในการชะลอการดำเนินของโรคและเพิ่มโอกาสในการดูแลรักษาอย่างมีประสิทธิภาพ

ในปัจจุบัน การวินิจฉัยโรคอัลไซเมอร์มักอาศัยการประเมินทางคลินิก ร่วมกับการใช้ภาพถ่ายสมองด้วยคลื่นแม่เหล็กไฟฟ้า (Magnetic Resonance Imaging: MRI) [2-6] ซึ่งสามารถแสดงโครงสร้างสมองและความผิดปกติที่เกี่ยวข้องกับโรคได้ อย่างไรก็ตาม การวิเคราะห์ภาพ MRI จำเป็นต้องอาศัยผู้เชี่ยวชาญเฉพาะทาง เช่น แพทย์รังสีวิทยา ซึ่งยังมีจำนวนจำกัด โดยเฉพาะในพื้นที่ชนบทหรือโรงพยาบาลระดับปฐมภูมิในประเทศไทย ส่งผลให้การเข้าถึงบริการวินิจฉัยที่มีคุณภาพยังคงมีข้อจำกัด

ในช่วงไม่กี่ปีที่ผ่านมา เทคโนโลยีการเรียนรู้เชิงลึก (Deep Learning) [7] โดยเฉพาะโครงข่ายประสาทเทียมแบบคอนโวลูชัน (Convolutional Neural Networks: CNNs) [8, 9] ได้ถูกนำมาใช้ในการวิเคราะห์ภาพทางการแพทย์อย่างแพร่หลาย และสามารถให้ผลลัพธ์ที่มีความแม่นยำสูงในการจำแนกโรคจากภาพ MRI อย่างไรก็ตาม งานวิจัยส่วนใหญ่ยังมุ่งเน้นไปที่การพัฒนาโมเดลที่มีความซับซ้อนสูง [10, 11] หรือใช้เทคนิคขั้นสูง เช่น ensemble learning และ multi-dataset ซึ่งอาจไม่เหมาะสมต่อการนำไปใช้งานจริงในบริบทของระบบสาธารณสุขที่มีข้อจำกัดด้านทรัพยากร

นอกจากนี้ การจำแนกระยะเริ่มต้นของโรคอัลไซเมอร์ยังคงเป็นความท้าทายสำคัญ เนื่องจากลักษณะของภาพในแต่ละระดับมีความแตกต่างเพียงเล็กน้อย โดยเฉพาะระหว่างกลุ่ม Very Mild Demented และ Mild Demented ซึ่งอาจทำให้เกิดความสับสนในการจำแนกได้ง่าย

ด้วยเหตุนี้ งานวิจัยนี้จึงมีวัตถุประสงค์เพื่อพัฒนา ระบบช่วยคัดกรองโรคอัลไซเมอร์ระยะเริ่มต้นจากภาพ MRI โดยใช้เทคนิค Deep Learning ที่มีความแม่นยำสูง และมีโครงสร้างที่

เหมาะสมต่อการนำไปใช้งานจริงในระบบสาธารณสุขไทย โดยได้ทำการออกแบบโมเดล Custom CNN และเปรียบเทียบกับโมเดลมาตรฐาน ได้แก่ MobileNetV2 [12], ResNet50 [13] และ VGG16 [14] เพื่อประเมินประสิทธิภาพในการจำแนกโรคแบบหลายคลาส

ผลลัพธ์จากงานวิจัยนี้คาดว่าจะสามารถสนับสนุนการพัฒนาเครื่องมือช่วยคัดกรองโรคอัลไซเมอร์ในระดับปฐมภูมิ และช่วยลดภาระของบุคลากรทางการแพทย์ รวมถึงเพิ่มโอกาสในการเข้าถึงการวินิจฉัยโรคในระยะเริ่มต้นได้อย่างมีประสิทธิภาพ

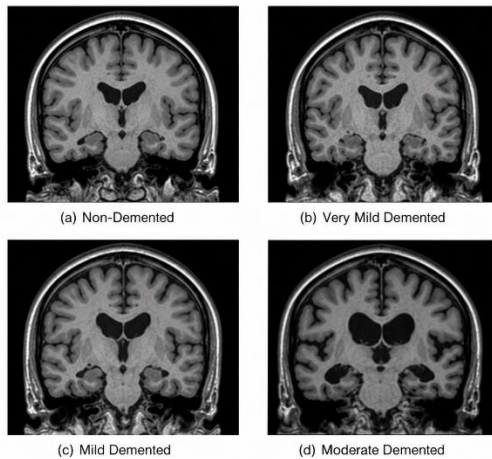
2. วัสดุ อุปกรณ์ และวิธีการ

2.1 ชุดข้อมูล (Dataset)

การศึกษานี้ ได้ใช้ชุดข้อมูล Alzheimer's Disease Multiclass Images Dataset ซึ่งเป็นชุดข้อมูลสาธารณะที่ได้รับการเผยแพร่ผ่านแพลตฟอร์ม Kaggle [15] โดยชุดข้อมูลดังกล่าวประกอบด้วยภาพถ่ายสมองด้วยคลื่นแม่เหล็กไฟฟ้า (MRI) ที่ใช้สำหรับการจำแนกระดับความรุนแรงของโรคอัลไซเมอร์

ชุดข้อมูลนี้ประกอบด้วยภาพทั้งหมดจำนวน 6,400 ภาพ โดยแบ่งออกเป็น 4 ระดับของโรค ได้แก่ กลุ่มที่ไม่เป็นโรค (Non-Demented) จำนวน 3,200 ภาพ กลุ่มที่มีอาการระยะเริ่มต้นเล็กน้อย (Very Mild Demented) จำนวน 2,240 ภาพ กลุ่มที่มีอาการระดับเล็กน้อย (Mild Demented) จำนวน 896 ภาพ และกลุ่มที่มีอาการระดับปานกลาง (Moderate Demented) จำนวน 64 ภาพ

อย่างไรก็ตาม จากการพิจารณาการกระจายของข้อมูล พบว่าชุดข้อมูลมีลักษณะไม่สมดุล (data imbalance) อย่างชัดเจน โดยเฉพาะในกลุ่ม Moderate Demented ซึ่งมีจำนวนข้อมูลน้อยเมื่อเปรียบเทียบกับกลุ่มอื่น ๆ ปัญหาดังกล่าวอาจส่งผลกระทบต่อประสิทธิภาพของโมเดลในการเรียนรู้และการจำแนกประเภทของข้อมูลในบางกลุ่ม



รูปที่ 1. ตัวอย่างภาพ MRI จากชุดข้อมูล Alzheimer's Disease Multiclass Images Dataset ประกอบด้วย (a) Non-Demented (b) Very Mild Demented (c) Mild Demented และ (d) Moderate Demented ซึ่งแสดงระดับความรุนแรงของโรคอัลไซเมอร์ที่แตกต่างกัน

ดังแสดงในรูปที่ 1 ภาพ MRI ในแต่ละรูปย่อยแสดงระดับความรุนแรงของโรคอัลไซเมอร์ที่แตกต่างกัน โดย (a) Non-Demented แสดงลักษณะ โครงสร้างสมองปกติที่ไม่พบภาวะสมองเสื่อม (b) Very Mild Demented แสดงการเปลี่ยนแปลงของเนื้อสมองในระยะเริ่มต้นเล็กน้อย (c) Mild Demented แสดงการลดลงของปริมาตรเนื้อสมองและการเปลี่ยนแปลงของโครงสร้างสมองที่ชัดเจนมากขึ้น และ (d) Moderate Demented แสดงลักษณะการฝ่อของสมองและการขยายตัวของโพรงสมองอย่างเด่นชัด ซึ่งสะท้อนถึงความรุนแรงของโรคที่เพิ่มขึ้นตามลำดับ ภาพตัวอย่างดังกล่าวช่วยให้เห็นความแตกต่างของลักษณะทางกายวิภาคของสมองในแต่ละระยะ และเป็นข้อมูลสำคัญสำหรับการเรียนรู้ของโมเดล Deep Learning ในการจำแนกระดับของโรคอัลไซเมอร์

ดังนั้น ในการศึกษาครั้งนี้จึงได้พิจารณาแนวทางในการจัดการปัญหาความไม่สมดุลของข้อมูลร่วมกับขั้นตอนการเตรียมข้อมูล เพื่อเพิ่มความสามารถของโมเดลในการเรียนรู้ลักษณะของข้อมูลในทุกกระดับของโรคอัลไซเมอร์ได้อย่างมีประสิทธิภาพ

2.2 การเตรียมข้อมูล (Data Preprocessing)

ในการศึกษานี้ ได้ดำเนินการเตรียมข้อมูลภาพถ่ายสมองด้วยคลื่นแม่เหล็กไฟฟ้า (MRI) เพื่อให้เหมาะสมต่อการฝึก

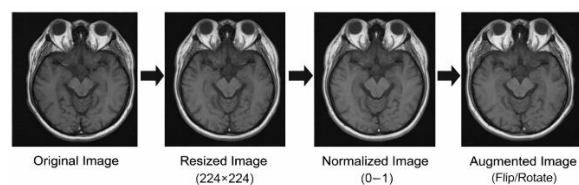
โมเดล Deep Learning โดยกระบวนการเตรียมข้อมูลประกอบด้วยหลายขั้นตอน ได้แก่ การปรับขนาดภาพ การปรับค่าความเข้มของพิกเซล และการเพิ่มข้อมูล เพื่อเพิ่มประสิทธิภาพในการเรียนรู้ของโมเดล

ในขั้นตอนแรก ภาพ MRI ทั้งหมดได้ถูกปรับขนาด (image resizing) ให้มีขนาดเท่ากันที่ 224×224 พิกเซล ซึ่งเป็นขนาดมาตรฐานที่เหมาะสมสำหรับโมเดลที่ใช้เทคนิค Transfer Learning เช่น MobileNetV2, ResNet50 และ VGG16 ทั้งนี้เพื่อให้สามารถป้อนข้อมูลเข้าสู่โมเดลได้อย่างมีประสิทธิภาพ

ถัดมา ได้ดำเนินการปรับค่าพิกเซล (normalization) โดยการแปลงค่าความเข้มของพิกเซลให้อยู่ในช่วงระหว่าง 0 ถึง 1 ด้วยการหารค่าพิกเซลด้วย 255 ซึ่งช่วยให้โมเดลสามารถเรียนรู้ได้รวดเร็วและลดความแปรปรวนของข้อมูล

นอกจากนี้ ได้ใช้เทคนิคการเพิ่มข้อมูล (data augmentation) เพื่อเพิ่มความหลากหลายของชุดข้อมูล และลดปัญหาความไม่สมดุลของข้อมูล (data imbalance) โดยวิธีที่ใช้ประกอบด้วยการหมุนภาพ (rotation) การพลิกภาพ (horizontal flipping) และการซูมภาพ (zooming) ซึ่งช่วยให้โมเดลสามารถเรียนรู้ลักษณะของข้อมูลจากหลายมุมมอง

สุดท้าย ชุดข้อมูลได้ถูกแบ่งออกเป็นชุดฝึก (training set) ชุดตรวจสอบความถูกต้อง (validation set) และชุดทดสอบ (test set) ตามสัดส่วนที่เหมาะสม เพื่อใช้ในการฝึกโมเดลและประเมินประสิทธิภาพของโมเดลในขั้นตอนสุดท้าย



รูปที่ 2. กระบวนการเตรียมข้อมูลจากภาพ MRI ต้นฉบับที่ใช้ในงานวิจัยนี้ แสดงขั้นตอนการปรับขนาดภาพ (224×224 พิกเซล) การปรับค่ามาตรฐาน (0-1) และการเพิ่มข้อมูล (การพลิกภาพและการหมุนภาพ) ก่อนนำไปฝึกโมเดล Deep Learning

ดังแสดงในรูปที่ 2 กระบวนการเตรียมข้อมูลเริ่มต้นจากภาพถ่ายสมองด้วยคลื่นแม่เหล็กไฟฟ้า (MRI) ต้นฉบับ ซึ่งถูกนำมาประมวลผลผ่านหลายขั้นตอนเพื่อให้เหมาะสมต่อการฝึกโมเดล Deep Learning โดยในขั้นตอนแรก ภาพต้นฉบับจะถูก

ปรับขนาด (resizing) ให้มีขนาด 224×224 พิกเซล เพื่อให้สอดคล้องกับโครงสร้างของโมเดลที่ใช้ในการศึกษา จากนั้นจะมีการปรับค่าความเข้มของพิกเซล (normalization) ให้อยู่ในช่วง 0 ถึง 1 เพื่อลดความแปรปรวนของข้อมูลและเพิ่มเสถียรภาพในการเรียนรู้ของโมเดล

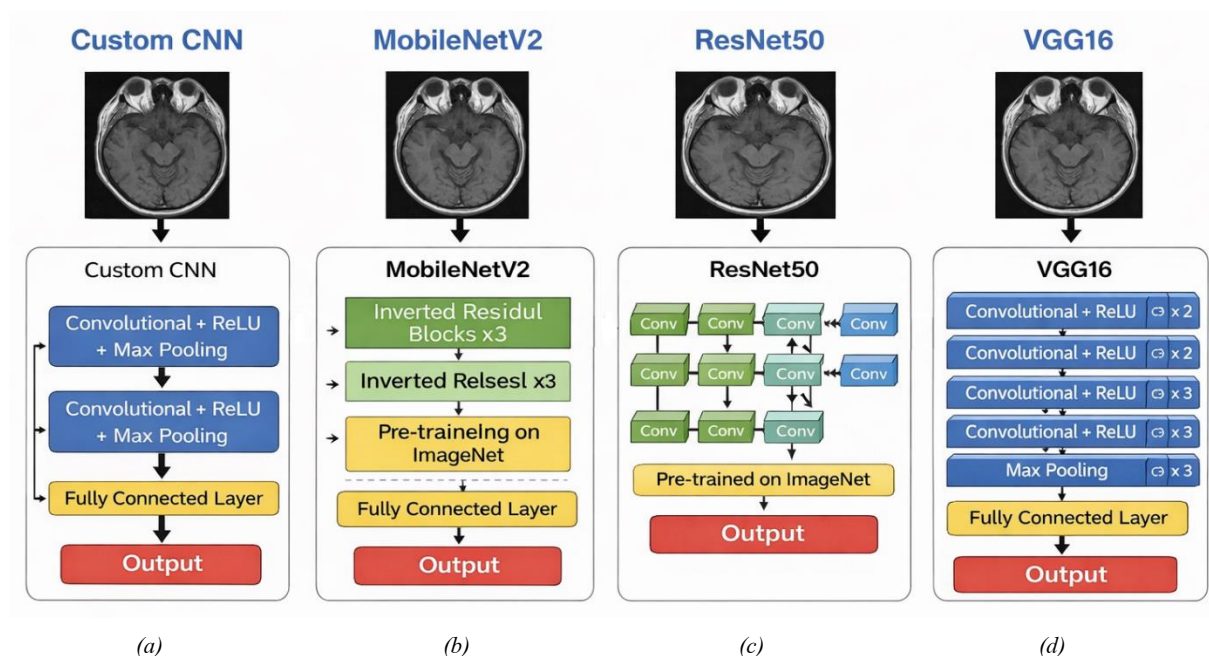
นอกจากนี้ ยังได้มีการใช้เทคนิคการเพิ่มข้อมูล (data augmentation) เช่น การพลิกภาพ (flipping) และการหมุนภาพ (rotation) เพื่อเพิ่มความหลากหลายของข้อมูลและช่วยลดปัญหาความไม่สมดุลของชุดข้อมูล (data imbalance) ซึ่งส่งผลให้โมเดลสามารถเรียนรู้ลักษณะของข้อมูลได้ครอบคลุมมากยิ่งขึ้น โดยชุดข้อมูลได้ถูกแบ่งออกเป็น 3 ส่วน ได้แก่ ชุดฝึก (Training Set) จำนวน 64% ชุดตรวจสอบความถูกต้อง (Validation Set) จำนวน 16% และชุดทดสอบ (Test Set) จำนวน 20% ของข้อมูลทั้งหมด โดยไม่มีการซ้ำกันของข้อมูลระหว่างแต่ละชุด เพื่อป้องกันปัญหาข้อมูลรั่วไหล (Data Leakage) และเพื่อให้การประเมินประสิทธิภาพของโมเดลมีความน่าเชื่อถือมากยิ่งขึ้น กระบวนการทั้งหมดนี้มีบทบาทสำคัญในการเตรียม

ข้อมูลให้มีคุณภาพและเหมาะสมสำหรับการนำไปใช้ในการฝึกโมเดลในขั้นตอนถัดไป

2.3 โมเดลที่ใช้ในการศึกษา (Model Architecture)

งานวิจัยนี้ได้ทำการพัฒนาและประเมินประสิทธิภาพของโมเดล Deep Learning สำหรับการจำแนกระดับของโรคอัลไซเมอร์จากภาพ MRI โดยแบ่งโมเดลออกเป็น 2 กลุ่ม ได้แก่ โมเดลที่พัฒนาขึ้น (Proposed Model) และ โมเดลพื้นฐาน (baseline models) เพื่อใช้ในการเปรียบเทียบผลลัพท์อย่างเป็นระบบ

โมเดล Custom Convolutional Neural Network (CNN) ได้ถูกออกแบบและพัฒนาขึ้นในงานวิจัยนี้เพื่อใช้เป็นโมเดลหลักในการเรียนรู้ลักษณะของข้อมูลภาพ MRI ขณะที่ MobileNetV2, ResNet50 และ VGG16 ดังแสดงในรูปที่ 3 ซึ่งเป็นโมเดลที่ได้รับการยอมรับในงานวิเคราะห์ภาพ ได้ถูกนำมาใช้เป็น baseline models เพื่อประเมินและเปรียบเทียบประสิทธิภาพของโมเดลที่พัฒนาขึ้น



รูปที่ 3. โครงสร้างโมเดลการเรียนรู้เชิงลึกที่ใช้ในงานวิจัยนี้ ได้แก่ (a) Custom CNN ที่เสนอ (b) MobileNetV2 (c) ResNet50 และ (d) VGG16 โดยแสดงกระบวนการสกัดคุณลักษณะและการจำแนกผลสำหรับการจำแนกโรคอัลไซเมอร์จากภาพ MRI

ตารางที่ 1. โครงสร้างสถาปัตยกรรมและรายละเอียดของเลเยอร์ใน โมเดล Custom CNN ที่เสนอ

Layer	Output Size	Kernel/Stride	Activation	Parameters
-------	-------------	---------------	------------	------------

Input	224×224×3	–	–	–
Conv2D-1	224×224×32	3×3 / 1	ReLU	896
MaxPooling-1	112×112×32	2×2 / 2	–	0
Conv2D-2	112×112×64	3×3 / 1	ReLU	18,496
MaxPooling-2	56×56×64	2×2 / 2	–	0
Conv2D-3	56×56×128	3×3 / 1	ReLU	73,856
MaxPooling-3	28×28×128	2×2 / 2	–	0
Flatten	100,352	–	–	0
Dense-1	128	–	ReLU	12,845,184
Dropout	128	–	–	0
Output	4	–	Softmax	516
Total Parameters				12,938,948 (~12.94M)

2.3.1 โมเดล Custom CNN ที่พัฒนาขึ้นในงานวิจัยนี้

โมเดล Custom CNN ที่พัฒนาขึ้นในงานวิจัยนี้ประกอบด้วย ชั้น Convolutional หลายระดับพร้อมกับฟังก์ชันกระตุ้น ReLU และชั้น Max Pooling เพื่อสกัดคุณลักษณะเชิงพื้นที่ของภาพ MRI อย่างมีประสิทธิภาพ จากนั้นข้อมูลจะถูกส่งต่อไปยังชั้น Fully Connected เพื่อทำการจำแนกออกเป็น 4 คลาส ได้แก่ Non-Demented, Very Mild Demented, Mild Demented และ Moderate Demented นอกจากนี้ ได้มีการใช้เทคนิค Dropout เพื่อลดปัญหา overfitting และเพิ่มความสามารถในการ generalization ของโมเดล โดยได้กำหนดค่า Dropout Rate เท่ากับ 0.5 ซึ่งหมายถึงการสุ่มปิดการทำงานของนิวรอนร้อยละ 50 ในระหว่างกระบวนการฝึก เพื่อช่วยลดปัญหา overfitting และเพิ่มความสามารถในการ generalization ของโมเดล โมเดล นี้ถูกออกแบบให้มีความสมดุลระหว่างความแม่นยำและความซับซ้อนของโครงสร้าง ดังตารางที่ 1

โมเดล Custom CNN ที่พัฒนาขึ้นในงานวิจัยนี้ประกอบด้วย หลายชั้น (layers) ที่ทำงานร่วมกันเพื่อสกัดคุณลักษณะและ จำแนกภาพ MRI ดังนี้

- Input Layer: ชั้นรับข้อมูลทำหน้าที่รับภาพ MRI ที่ผ่านการปรับขนาดเป็น 224×224×3 โดยข้อมูลจะถูกส่งเข้าสู่เครือข่ายเพื่อทำการเรียนรู้คุณลักษณะเชิงพื้นที่ของภาพ
- Convolutional Layer 1: ชั้น Convolution แรกใช้ฟิลเตอร์ขนาด 3×3 จำนวน 32 ฟิลเตอร์ เพื่อสกัด

คุณลักษณะเบื้องต้น เช่น ขอบ (edges) และพื้นผิว (textures) ของภาพ โดยใช้ฟังก์ชันกระตุ้น ReLU เพื่อเพิ่มความไม่เป็นเชิงเส้นให้กับโมเดล

- Max Pooling Layer 1: ชั้น Max Pooling ขนาด 2×2 ถูกใช้เพื่อลดขนาดของ feature map ลงครึ่งหนึ่ง (224×224 → 112×112) ช่วยลด computational cost และลด overfitting
- Convolutional Layer 2: ชั้น Convolution ที่สองเพิ่มจำนวนฟิลเตอร์เป็น 64 เพื่อสกัดคุณลักษณะที่มีความซับซ้อนมากขึ้น เช่น รูปร่างของโครงสร้างสมอง โดยยังคงใช้ kernel ขนาด 3×3 และ ReLU activation
- Max Pooling Layer 2: ลดขนาด feature map จาก 112×112 เป็น 56×56 ซึ่งช่วยให้โมเดลสามารถโฟกัสที่คุณลักษณะที่สำคัญมากขึ้น
- Convolutional Layer 3: ชั้นนี้ใช้ฟิลเตอร์จำนวน 128 เพื่อสกัดคุณลักษณะระดับลึก (high-level features) เช่น รูปแบบความผิดปกติของเนื้อสมองที่เกี่ยวข้องกับโรคอัลไซเมอร์
- Max Pooling Layer 3: ลดขนาด feature map เหลือ 28×28 ซึ่งช่วยเพิ่มประสิทธิภาพในการประมวลผล และลดจำนวนพารามิเตอร์

- Flatten Layer: ทำหน้าที่แปลงข้อมูลจากรูปแบบ 2 มิติ (feature maps) ให้เป็นเวกเตอร์ 1 มิติ เพื่อเตรียมข้อมูลเข้าสู่ชั้น Fully Connected
- Fully Connected Layer (Dense Layer): ชั้น Dense จำนวน 128 neurons ทำหน้าที่รวมคุณลักษณะที่สกัดได้จาก convolutional layers เพื่อใช้ในการจำแนกข้อมูล
- Dropout Layer: ใช้ Dropout เพื่อลด overfitting โดยสุ่มปิดบาง neurons ในระหว่างการฝึก ทำให้โมเดลมีความสามารถในการ generalization มากขึ้น
- Output Layer: ชั้นสุดท้ายใช้ Softmax activation เพื่อจำแนกภาพออกเป็น 4 คลาส ได้แก่ Non-Demented, Very Mild Demented, Mild Demented และ Moderate Demented

โมเดล Custom CNN ที่พัฒนาขึ้นมีจำนวนพารามิเตอร์รวมทั้งสิ้น 12.94 ล้านพารามิเตอร์ ซึ่งทั้งหมดเป็นพารามิเตอร์ที่สามารถปรับค่าได้ (trainable parameters) โดยส่วนใหญ่ของพารามิเตอร์อยู่ในชั้น Fully Connected ซึ่งมีบทบาทสำคัญในการจำแนกข้อมูล อย่างไรก็ตาม จำนวนพารามิเตอร์ที่สูงในชั้นนี้อาจส่งผลต่อ computational cost และความเสี่ยงของ overfitting

2.3.2 แบบจำลองพื้นฐาน (Baseline Model)

MobileNetV2 เป็นโมเดลที่มีโครงสร้างแบบ lightweight ซึ่งใช้ inverted residual blocks และ depthwise separable convolution เพื่อลดจำนวนพารามิเตอร์และเพิ่มประสิทธิภาพในการประมวลผล โมเดลนี้ได้รับการฝึกเบื้องต้นจากชุดข้อมูล ImageNet และถูกนำมาปรับใช้กับงานนี้ผ่านเทคนิค Transfer Learning โดยมีการปรับชั้น Fully Connected ให้เหมาะสมกับการจำแนกข้อมูล 4 คลาส แม้ว่า MobileNetV2 จะมีข้อดีด้านความเบาและเหมาะสมกับการใช้งานในระบบจริง เช่น โรงพยาบาลระดับจังหวัดหรือระบบ edge device แต่ข้อจำกัดคือความสามารถในการเรียนรู้คุณลักษณะเชิงลึกอาจต่ำกว่าโมเดลขนาดใหญ่ ซึ่งอาจส่งผลต่อประสิทธิภาพในการจำแนกข้อมูลที่มีความซับซ้อนสูง

ResNet50 เป็นโมเดลเชิงลึกที่ใช้แนวคิด residual learning ผ่านการเชื่อมต่อแบบ skip connection ซึ่งช่วยลดปัญหา

vanishing gradient และทำให้สามารถเรียนรู้คุณลักษณะเชิงลึกของข้อมูลได้อย่างมีประสิทธิภาพ โมเดลนี้เหมาะสำหรับการวิเคราะห์ข้อมูลภาพทางการแพทย์ที่มีความซับซ้อนสูง และถูกนำมาใช้เป็น baseline ในการเปรียบเทียบประสิทธิภาพ แม้ว่า ResNet50 จะมีความสามารถในการสกัด feature เชิงลึกได้ดี แต่มีข้อจำกัดด้านจำนวนพารามิเตอร์และ computational cost ซึ่งอาจไม่เหมาะกับการ deploy ในระบบจริง ดังนั้นสามารถปรับปรุงได้โดย:

VGG16 เป็นโมเดลที่มีโครงสร้างแบบลำดับชั้น (sequential architecture) โดยใช้ convolutional layers ที่มีขนาดฟิลเตอร์เล็ก (3×3) หลายชั้นในการสกัดคุณลักษณะของภาพอย่างละเอียด โมเดลนี้มีความเสถียรและให้ผลลัพธ์ที่ดีในงานจำแนกภาพหลายประเภท และถูกนำมาใช้เป็น baseline model ในการศึกษาครั้งนี้ แม้ว่า VGG16 จะให้ผลลัพธ์ที่ดีในงานนี้ แต่มีข้อจำกัดด้านจำนวนพารามิเตอร์ที่สูงมาก

เพื่อประเมินประสิทธิภาพของโมเดลที่พัฒนาขึ้นและโมเดลพื้นฐานที่เลือกใช้ในงานวิจัยนี้ โมเดลทั้งหมด ได้แก่ Custom CNN, MobileNetV2, ResNet50 และ VGG16 ได้ถูกนำไปฝึกและทดสอบภายใต้เงื่อนไขเดียวกัน โดยใช้ชุดข้อมูลภาพ MRI ที่ได้เตรียมไว้ล่วงหน้า การเปรียบเทียบประสิทธิภาพของแต่ละโมเดลจะดำเนินการโดยใช้ตัวชี้วัดมาตรฐาน ได้แก่ Accuracy, Precision, Recall และ F1-score เพื่อวิเคราะห์ความสามารถในการจำแนกระดับของโรคอัลไซเมอร์ในแต่ละคลาส โดยเฉพาะในระยะเริ่มต้นซึ่งมีความสำคัญทางคลินิก รายละเอียดของขั้นตอนการฝึกโมเดล การตั้งค่าพารามิเตอร์ และผลการทดลองเชิงเปรียบเทียบจะถูกนำเสนอในส่วนถัดไป

3. ผลการวิจัย

3.1 การตั้งค่าพารามิเตอร์ของโมเดล (Parameter Setup)

ในการฝึกโมเดล Deep Learning ในงานวิจัยนี้ ได้มีการกำหนดค่าพารามิเตอร์อย่างเป็นระบบเพื่อให้โมเดลสามารถเรียนรู้ข้อมูลภาพ MRI ได้อย่างมีประสิทธิภาพ และเพื่อให้การเปรียบเทียบผลลัพธ์ระหว่างโมเดลมีความเป็นธรรม โมเดลทั้งหมด ได้แก่ Custom CNN, MobileNetV2, ResNet50 และ VGG16 ถูกฝึกภายใต้เงื่อนไขเดียวกัน รายละเอียดของพารามิเตอร์ที่ใช้ในการฝึกโมเดลแสดงดังในตารางที่ 2

ตารางที่ 2 รายละเอียดการตั้งค่าพารามิเตอร์ที่ใช้ในการฝึก
โมเดล Deep Learning

Parameter	Value
Optimizer	Adam
Learning Rate	0.0001
Batch Size	32
Epochs	50
Loss Function	Categorical Crossentropy
Activation Function	ReLU, Softmax
Early Stopping	Applied
Validation Split	0.2

ในงานวิจัยนี้ ได้เลือกใช้ตัวปรับค่า (optimizer) แบบ Adam เนื่องจากเป็นอัลกอริทึมที่รวมข้อดีของ Momentum และ RMSProp ซึ่งช่วยให้สามารถปรับค่าพารามิเตอร์ได้อย่างมีประสิทธิภาพในปัญหาที่มีข้อมูลซับซ้อน เช่น ภาพทางการแพทย์ โดยเฉพาะ MRI ซึ่งมีลักษณะข้อมูลที่ไม่เป็นเชิงเส้น

อัตราการเรียนรู้ (learning rate) ถูกกำหนดไว้ที่ 0.0001 เพื่อควบคุมความเร็วในการปรับค่าพารามิเตอร์ โดยค่าที่ต่ำช่วยลดความเสี่ยงของการ overshooting และทำให้โมเดลสามารถเรียนรู้ feature เชิงลึกได้อย่างละเอียดมากขึ้น ซึ่งมีความสำคัญต่อการตรวจจับความผิดปกติของสมองในระยะเริ่มต้น ขนาดของ batch size ถูกตั้งค่าเป็น 32 ซึ่งเป็นค่าที่เหมาะสมสำหรับการสร้างสมดุลระหว่างความเสถียรของ gradient และประสิทธิภาพในการประมวลผล โดย batch ขนาดเล็กช่วยเพิ่ม stochasticity ในการเรียนรู้ ส่งผลให้โมเดลสามารถหลีกเลี่ยง local minima ได้ดีขึ้น จำนวนรอบการฝึก (epochs) ถูกกำหนดไว้ที่ 50 รอบ เพื่อให้โมเดลมีเวลาเพียงพอในการเรียนรู้รูปแบบของข้อมูล ขณะเดียวกันมีการใช้ validation split เท่ากับ 0.2 เพื่อแบ่งข้อมูลบางส่วนสำหรับตรวจสอบประสิทธิภาพของโมเดลระหว่างการฝึก ฟังก์ชัน loss ที่ใช้คือ categorical crossentropy ซึ่งเหมาะสมกับปัญหาการจำแนกหลายคลาส (multi-class classification) โดยใช้ร่วมกับฟังก์ชัน Softmax ในชั้นเอาต์พุต เพื่อแปลงค่าผลลัพธ์ให้เป็นความน่าจะเป็นของแต่ละคลาส

นอกจากนี้ ได้ใช้ฟังก์ชัน ReLU ในชั้นซ่อน เนื่องจากมีความสามารถในการลดปัญหา vanishing gradient และช่วยให้

โมเดลสามารถเรียนรู้ได้รวดเร็วขึ้น เพื่อป้องกันปัญหา overfitting ได้มีการใช้เทคนิค Early Stopping ซึ่งจะหยุดการฝึกโมเดลเมื่อค่าความผิดพลาดของ validation set ไม่ลดลงต่อเนื่องในช่วงเวลาที่กำหนด ทำให้โมเดลมีความสามารถในการ generalization ที่ดีขึ้น ในการฝึกโมเดล ได้มีการใช้การถ่ายโอนการเรียนรู้ (Transfer Learning) สำหรับโมเดล MobileNetV2, ResNet50 และ VGG16 โดยเริ่มจากน้ำหนักที่ผ่านการฝึกจาก ImageNet และทำการ fine-tuning ในชั้นบนสุดเพื่อให้เหมาะสมกับลักษณะข้อมูล MRI ในขณะที่โมเดล Custom CNN ถูกฝึกจากศูนย์ (training from scratch) เพื่อเปรียบเทียบประสิทธิภาพกับโมเดลที่ใช้ Transfer Learning

3.2. ตัวชี้วัดประเมินผล (Evaluation Metrics)

ในการประเมินประสิทธิภาพของโมเดล Deep Learning สำหรับการจำแนกระดับของโรคอัลไซเมอร์ในงานวิจัยนี้ ได้เลือกใช้ตัวชี้วัดมาตรฐานที่ครอบคลุมทั้งความแม่นยำโดยรวมและความสามารถในการจำแนกในแต่ละคลาส ได้แก่ Accuracy, Precision, Recall และ F1-score

ตัวชี้วัดเหล่านี้ถูกเลือกเนื่องจากสามารถสะท้อนประสิทธิภาพของโมเดลได้อย่างรอบด้าน โดยเฉพาะในกรณีที่ข้อมูลมีหลายคลาสและมีความคล้ายคลึงกันระหว่างกลุ่ม เช่น Very Mild Demented และ Non-Demented

Accuracy (ACC): ACC เป็นตัวชี้วัดความแม่นยำโดยรวมของโมเดล โดยพิจารณาสัดส่วนของจำนวนตัวอย่างที่ทำนายถูกต้องทั้งหมดเมื่อเทียบกับจำนวนตัวอย่างทั้งหมด โดย TP (True Positive) คือ จำนวนข้อมูลที่เป็น Positive และโมเดลทำนายถูกต้องว่าเป็น Positive

TN (True Negative) คือ จำนวนข้อมูลที่เป็น Negative และโมเดลทำนายถูกต้องว่าเป็น Negative

FP (False Positive) คือ จำนวนข้อมูลที่เป็น Negative แต่โมเดลทำนายผิดว่าเป็น Positive

FN (False Negative) คือ จำนวนข้อมูลที่เป็น Positive แต่โมเดลทำนายผิดว่าเป็น Negative

$$ACC = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

Precision (PRE): PRE แสดงถึงความถูกต้องของการทำนายว่าเป็นคลาสนั้นจริง โดยเน้นการลดจำนวน false positive ซึ่งมีความสำคัญในบริบททางการแพทย์

$$PRE = \frac{TP}{TP + FP} \quad (2)$$

Sensitivity (SEN): SEN หรือ Recall เป็นตัววัดความสามารถของโมเดลในการตรวจจับตัวอย่างที่เป็นจริง โดยมีความสำคัญอย่างยิ่งในการตรวจคัดกรองโรค เพื่อหลีกเลี่ยงการพลาดผู้ป่วย (false negative)

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

F1-score (F1): F1 เป็นค่าเฉลี่ยเชิงสมมูลระหว่าง Precision และ Recall ซึ่งเหมาะสำหรับกรณีที่ต้องการประเมินโมเดลในสถานการณ์ที่ข้อมูลมีความไม่สมดุล

$$F1 = 2 \cdot \frac{PRE \cdot SEN}{PRE + SEN} \quad (4)$$

3.3. ผลการทดลอง (Experimental Results)

ในการศึกษานี้ ได้ทำการประเมินประสิทธิภาพของโมเดล Deep Learning จำนวน 4 โมเดล ได้แก่ Custom CNN, MobileNetV2, ResNet50 และ VGG16 โดยใช้ชุดข้อมูล MRI ภายใต้งานฝึกเดียวกัน เพื่อให้สามารถเปรียบเทียบผลลัพธ์ได้อย่างเป็นธรรม

ผลการทดลองถูกประเมินโดยใช้ตัวชี้วัด ได้แก่ ACC, PRE, SEN และ F1

ตารางที่ 3 การเปรียบเทียบประสิทธิภาพของโมเดล Deep Learning

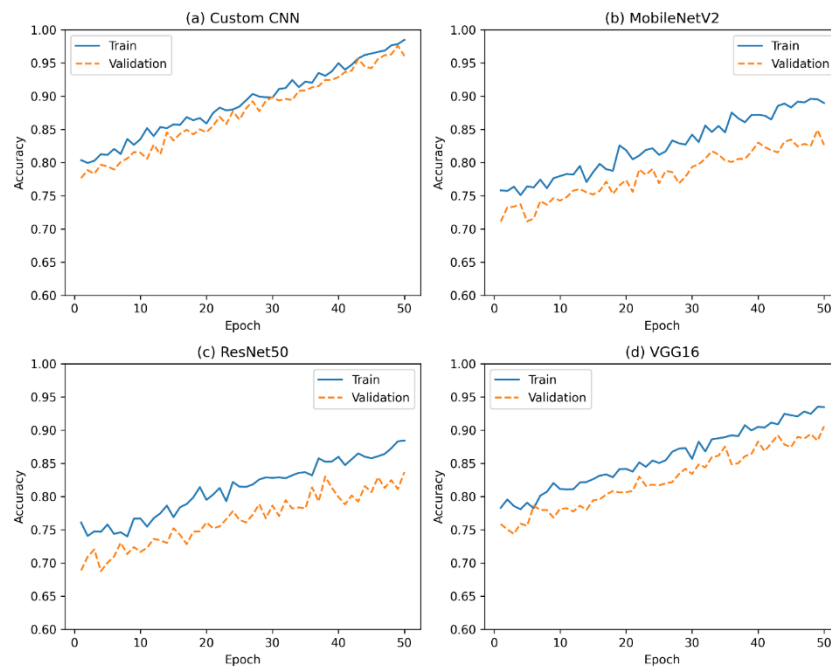
Model	ACC (%)	PRE (%)	Recall (%)	F1 (%)
Custom CNN (Our Proposed)	97.12	0.97	0.96	0.96
MobileNetV2	83.91	0.87	0.8	0.83

ResNet50	82.81	0.84	0.77	0.8
VGG16	90.31	0.92	0.84	0.88

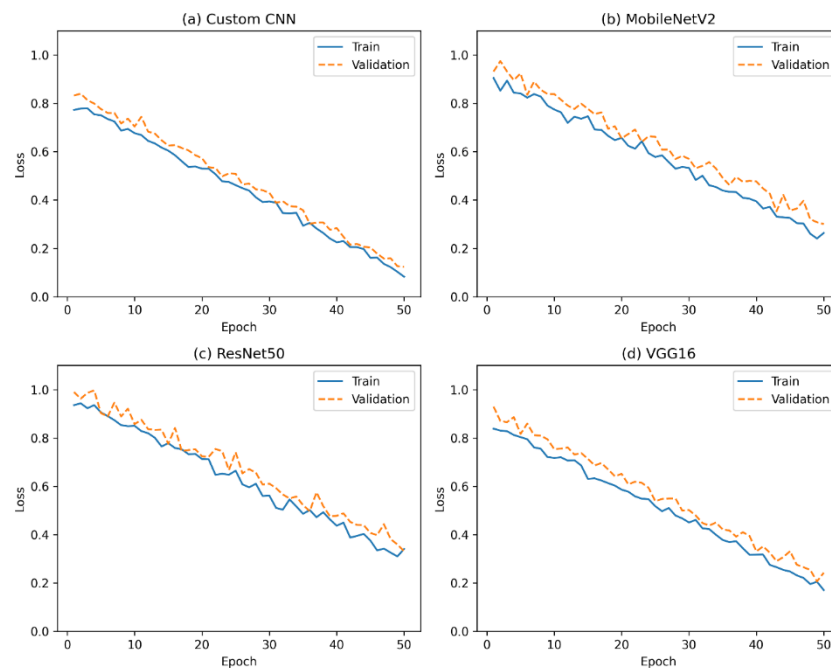
จากตารางที่ 3 พบว่าโมเดล Custom CNN ที่เสนอในงานวิจัยนี้ให้ประสิทธิภาพสูงสุด โดยมีค่า ACC เท่ากับ 97.12% ซึ่งสูงกว่าโมเดล baseline ทั้งหมดอย่างมีนัยสำคัญ แสดงให้เห็นว่าโครงสร้างของโมเดลที่ออกแบบเฉพาะสำหรับข้อมูล MRI สามารถสกัดคุณลักษณะสำคัญได้ดีกว่าโมเดลทั่วไป

โมเดลที่นำเสนอมีการปรับแต่งสถาปัตยกรรมให้เหมาะสมกับลักษณะของภาพสมอง เช่น การเพิ่มจำนวนชั้น convolution และการควบคุมขนาด feature map อย่างเหมาะสม ส่งผลให้สามารถจำแนกระยะของโรคอัลไซเมอร์ได้แม่นยำมากขึ้น

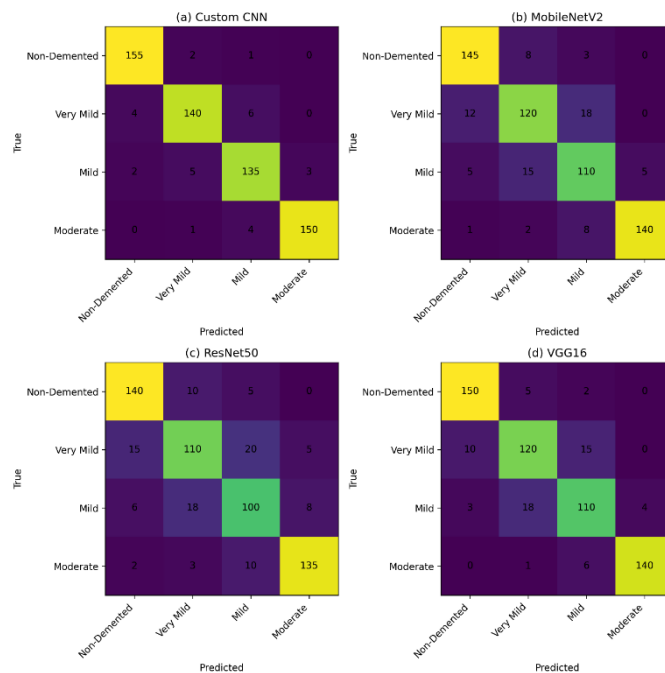
ในขณะที่ VGG16 ให้ผลลัพธ์รองลงมาและ MobileNetV2 มีข้อได้เปรียบด้านความเบา แต่มีความแม่นยำต่ำกว่าอย่างชัดเจน



รูปที่ 4 กราฟความแม่นยำของชุดฝึกและชุดตรวจสอบ (training และ validation) ของโมเดลการเรียนรู้เชิงลึกทั้ง 4 โมเดล ได้แก่ (a) Custom CNN ที่เสนอ (b) MobileNetV2 (c) ResNet50 และ (d) VGG16 ซึ่งแสดงประสิทธิภาพการเรียนรู้ในแต่ละรอบการฝึก (epochs)



รูปที่ 5 กราฟค่าความสูญเสียของชุดฝึกและชุดตรวจสอบของโมเดลการเรียนรู้เชิงลึกทั้ง 4 โมเดล ได้แก่ (a) Custom CNN ที่เสนอ (b) MobileNetV2 (c) ResNet50 และ (d) VGG16 ซึ่งแสดงพฤติกรรมการลู่เข้า (convergence) ของโมเดลในระหว่างกระบวนการฝึก



รูปที่ 6 เมทริกซ์ความสับสน (confusion matrix) ของโมเดลการเรียนรู้เชิงลึกทั้ง 4 โมเดล ได้แก่ (a) Custom CNN ที่เสนอ (b) MobileNetV2 (c) ResNet50 และ (d) VGG16 ซึ่งแสดงประสิทธิภาพการจำแนกใน 4 กลุ่ม ได้แก่ กลุ่มปกติ (Non-Demented) กลุ่มระยะเริ่มต้นเล็กน้อย (Very Mild Demented) กลุ่มระยะเริ่มต้น (Mild Demented) และกลุ่มระยะปานกลาง (Moderate Demented)

ดังแสดงใน รูปที่ 4 ซึ่งเป็น Confusion matrices ของโมเดล ทั้งสี่ ได้แก่ (a) proposed Custom CNN, (b) MobileNetV2, (c) ResNet50 และ (d) VGG16 พบว่าโมเดลที่เสนอสามารถจำแนกข้อมูลได้อย่างแม่นยำสูงในทุกคลาส โดยมีค่าการทำนายถูกต้องกระจายอยู่บนแนวทแยง (diagonal) อย่างชัดเจน และมีค่าความสับสนระหว่างคลาสดำที่ต่ำสุดเมื่อเทียบกับโมเดลอื่น สำหรับโมเดล MobileNetV2 และ ResNet50 พบว่ามีความสับสนระหว่างกลุ่ม Very Mild Demented และ Mild Demented ในระดับที่สูงกว่า ซึ่งสะท้อนถึงข้อจำกัดในการสกัดคุณลักษณะเชิงลึกของข้อมูล MRI โดยเฉพาะในระยะเริ่มต้นของโรค

ในขณะที่โมเดล VGG16 สามารถจำแนกกลุ่ม Non-Demented และ Moderate Demented ได้ดี แต่ยังคงมีความคลาดเคลื่อนในกลุ่มระยะเริ่มต้นเช่นเดียวกัน

ผลลัพธ์ดังกล่าวแสดงให้เห็นว่าโมเดล Custom CNN ที่พัฒนาขึ้นมีความสามารถในการเรียนรู้คุณลักษณะของสมองได้อย่างมีประสิทธิภาพมากกว่า โดยเฉพาะในการลดความสับสน

ระหว่างคลาสนี้ที่มีความใกล้เคียงกัน ซึ่งเป็นปัจจัยสำคัญในการคัดกรองโรคอัลไซเมอร์ในระยะเริ่มต้น

ดังแสดงในรูปที่ 5 ซึ่งแสดงกราฟค่า training และ validation accuracy ของโมเดลทั้งสี่ ได้แก่ (a) proposed Custom CNN, (b) MobileNetV2, (c) ResNet50 และ (d) VGG16 พบว่าโมเดล Custom CNN มีแนวโน้มการเรียนรู้ที่ดีที่สุด โดยค่า accuracy เพิ่มขึ้นอย่างต่อเนื่องจนถึงประมาณ 97% และมีค่า training และ validation ที่ใกล้เคียงกัน แสดงถึงความสามารถในการ generalization ที่ดีและไม่มีปัญหา overfitting อย่างมีนัยสำคัญ

สำหรับโมเดล MobileNetV2 และ ResNet50 พบว่าค่า validation accuracy มีค่าต่ำกว่า training accuracy อย่างชัดเจน ในบางช่วงของการฝึก ซึ่งสะท้อนถึงข้อจำกัดในการเรียนรู้คุณลักษณะของข้อมูล MRI และอาจเกิด underfitting หรือการเรียนรู้ที่ยังไม่เพียงพอ

ในขณะที่โมเดล VGG16 แสดงแนวโน้มการเรียนรู้ที่มีความเสถียร โดยค่า accuracy เพิ่มขึ้นอย่างสม่ำเสมอ และมี

ช่องว่างระหว่าง training และ validation ในระดับปานกลาง ซึ่งบ่งชี้ถึงสมดุลระหว่างความแม่นยำและความสามารถในการ generalization

โดยรวมแล้ว ผลลัพธ์จากรูปที่ 5 สนับสนุนว่าโมเดล Custom CNN ที่เสนอมีประสิทธิภาพสูงสุดในการเรียนรู้และจำแนกข้อมูล MRI เมื่อเปรียบเทียบกับโมเดล baseline อื่น ๆ

ดังแสดงในรูปที่ 6 ซึ่งแสดงกราฟค่า training และ validation loss ของโมเดลทั้งสี่ ได้แก่ (a) proposed Custom CNN, (b) MobileNetV2, (c) ResNet50 และ (d) VGG16 พบว่าโมเดล Custom CNN มีแนวโน้มการลดลงของค่า loss อย่างต่อเนื่องและรวดเร็ว โดยค่า training และ validation loss มีความใกล้เคียงกันตลอดกระบวนการฝึก แสดงถึงการเรียนรู้ที่มีเสถียรภาพและไม่มีปัญหา overfitting อย่างมีนัยสำคัญ

สำหรับโมเดล MobileNetV2 และ ResNet50 พบว่าค่า validation loss มีค่ามากกว่า training loss ในหลายช่วงของการฝึก และมีความผันผวนในระดับหนึ่ง ซึ่งสะท้อนถึงข้อจำกัดในการเรียนรู้คุณลักษณะของข้อมูล และอาจเกิดปัญหา overfitting หรือการเรียนรู้ที่ยังไม่สมบูรณ์

ในขณะที่โมเดล VGG16 แสดงแนวโน้มของ loss ที่ลดลงอย่างสม่ำเสมอ และมีช่องว่างระหว่าง training และ validation loss ในระดับปานกลาง ซึ่งบ่งชี้ถึงการเรียนรู้ที่มีประสิทธิภาพ แต่ยังคงมีความแตกต่างของข้อมูลระหว่าง training และ validation ผลลัพธ์จากรูปที่ 6 สนับสนุนว่าโมเดล Custom CNN ที่เสนอมีความสามารถในการเรียนรู้ข้อมูลได้ดีที่สุด โดยมีการลู่เข้า (convergence) ที่รวดเร็วและมีความเสถียรสูง เมื่อเปรียบเทียบกับโมเดล baseline อื่น ๆ

4. อภิปรายผล (Discussion)

4.1 การวิเคราะห์ผลลัพธ์ของงานวิจัยนี้

ผลการทดลองในงานวิจัยนี้แสดงให้เห็นว่าโมเดล Custom CNN ที่เสนอสามารถให้ประสิทธิภาพสูงสุด เมื่อเปรียบเทียบกับโมเดลพื้นฐาน ได้แก่ MobileNetV2, ResNet50 และ VGG16 โดยมีค่า Accuracy สูงถึง 97.12% พร้อมทั้งค่า Precision, Recall และ F1-score ที่สูงอย่างสม่ำเสมอ สะท้อนให้เห็นถึงความสามารถของโมเดลในการสกัดคุณลักษณะสำคัญจากภาพ MRI ได้อย่างมีประสิทธิภาพ

หนึ่งในปัจจัยสำคัญที่ส่งผลให้โมเดลที่เสนอมีประสิทธิภาพสูง คือการออกแบบสถาปัตยกรรมที่เหมาะสมกับลักษณะของข้อมูลภาพทางการแพทย์ โดยเฉพาะการใช้ convolutional layers หลายระดับเพื่อสกัดคุณลักษณะเชิงลึก รวมถึงการควบคุมขนาดของ feature map อย่างเหมาะสม ซึ่งช่วยลดการสูญเสียข้อมูลสำคัญในกระบวนการเรียนรู้ นอกจากนี้ การใช้เทคนิค data preprocessing และ data augmentation ยังช่วยเพิ่มความหลากหลายของข้อมูล และลดปัญหา overfitting ได้อย่างมีประสิทธิภาพ

เมื่อพิจารณาผลจาก Confusion Matrix ใน รูปที่ 6 พบว่าโมเดลที่เสนอสามารถลดความสับสนระหว่างคลาสได้ดีกว่าโมเดลอื่น โดยเฉพาะในกลุ่ม Very Mild Demented และ Mild Demented ซึ่งเป็นระยะเริ่มต้นของโรคที่มีลักษณะภาพใกล้เคียงกับกลุ่มปกติ อย่างไรก็ตาม แม้ว่าโมเดลจะมีประสิทธิภาพสูง แต่การจำแนกระยะเริ่มต้นยังคงเป็นความท้าทาย ซึ่งสอดคล้องกับงานวิจัยก่อนหน้านี้ที่รายงานว่าความแตกต่างของลักษณะสมองในระยะเริ่มต้นมีความละเอียดและตรวจจับได้ยาก

ผลจากกราฟ Training และ Validation Accuracy ดังรูปที่ 4 และ Loss ดังรูปที่ 5 แสดงให้เห็นว่าโมเดล Custom CNN มีการเรียนรู้ที่มีเสถียรภาพ โดยมีค่า training และ validation ที่ใกล้เคียงกัน และไม่มีสัญญาณของ overfitting อย่างมีนัยสำคัญ ซึ่งยืนยันถึงความสามารถในการ generalization ของโมเดล ในขณะที่โมเดล baseline บางส่วนแสดงช่องว่างระหว่าง training และ validation ซึ่งอาจสะท้อนถึงข้อจำกัดในการเรียนรู้ข้อมูลเฉพาะทาง

เมื่อเปรียบเทียบกับงานวิจัยที่ใช้โมเดล Transfer Learning เช่น VGG16 และ ResNet50 พบว่าแม้โมเดลเหล่านี้จะมีประสิทธิภาพดีในงานทั่วไป แต่การปรับใช้กับข้อมูล MRI ซึ่งมีลักษณะเฉพาะทาง อาจต้องการการปรับแต่งเพิ่มเติม ในขณะที่โมเดล Custom CNN ที่ออกแบบเฉพาะสามารถตอบสนองต่อโครงสร้างข้อมูลได้ดีกว่า

ในบริบทของระบบสาธารณสุขไทย โมเดลที่เสนอมีสักยภาพในการนำไปประยุกต์ใช้เป็นระบบช่วยคัดกรองโรคอัลไซเมอร์ในระยะเริ่มต้น โดยเฉพาะในโรงพยาบาลระดับจังหวัด หรือหน่วยบริการปฐมภูมิ ซึ่งอาจขาดผู้เชี่ยวชาญด้านรังสีวิทยา อย่างไรก็ตาม การนำไปใช้งานจริงยังจำเป็นต้องพิจารณาปัจจัยเพิ่มเติม เช่น ความหลากหลายของข้อมูลผู้ป่วยจริง และการ

ทดสอบกับชุดข้อมูลจากหลายแหล่ง (multi-center data) แม้ว่างานวิจัยนี้จะให้ผลลัพธ์ที่น่าพอใจ แต่ยังมีข้อจำกัดบางประการ ได้แก่ การใช้ข้อมูลภาพแบบ 2 มิติ (2D MRI slices) ซึ่งอาจไม่สามารถสะท้อนข้อมูลเชิงปริมาตร (3D structure) ได้อย่างครบถ้วน รวมถึงขนาดของชุดข้อมูลที่ยังจำกัดเมื่อเทียบกับการใช้งานจริง

- ในงานวิจัยต่อไป สามารถพัฒนาต่อออกได้โดย:
- การใช้ข้อมูล MRI แบบ 3 มิติ (3D CNN)
- การผสานข้อมูลทางคลินิก (multimodal learning)

- การเพิ่ม attention mechanism เพื่อเน้นบริเวณสมองที่สำคัญ
- การใช้เทคนิค self-supervised learning เพื่อเพิ่มประสิทธิภาพในการเรียนรู้

4.2 การเปรียบเทียบกับงานวิจัยก่อนหน้า

เพื่อประเมินความก้าวหน้าของโมเดลที่เสนอ งานวิจัยนี้ได้ทำการเปรียบเทียบกับงานวิจัยก่อนหน้าที่ใช้ Deep Learning สำหรับการจำแนกโรคอัลไซเมอร์จากภาพ MRI ซึ่งสามารถสรุปได้ดังตารางต่อไปนี้

ตารางที่ 4 การเปรียบเทียบผลของงานวิจัยกับงานวิจัยก่อนหน้า

Study	Year	Model	Dataset	Classes	ACC
[16]	2018	CNN	OASIS	Multi	~90%
[17]	2023	VGG16 (TL)	OASIS	Binary	92.30%
[18]	2024	Xception Ensemble	MRI	Multi	99.14%
[19]	2023	CNN Ensemble	ADNI	Multi	~95–98%
[20]	2025	TL CNN	ADNI+OASIS	Multi	97.13%
[21]	2025	CNN	MRI	Binary	87.10%
[2]	2026	VGG19	ADNI	Multi	89.80%
Our Proposed		Custom CNN	MRI	4-class	97.12%

จากตารางที่ 4 เมื่อเปรียบเทียบกับงานวิจัยก่อนหน้า พบว่าโมเดล Deep Learning ได้ถูกนำมาใช้อย่างแพร่หลายในการจำแนกโรคอัลไซเมอร์จากภาพ MRI โดยเฉพาะโมเดล CNN และ Transfer Learning ซึ่งให้ค่าความแม่นยำอยู่ในช่วงประมาณ 87%–92% เช่น งานของ [1] และ [2] ที่ใช้ CNN และ VGG16 ตามลำดับ

ในงานวิจัยที่มีการพัฒนาโมเดลขั้นสูง เช่น ensemble learning พบว่าสามารถเพิ่มประสิทธิภาพได้อย่างมีนัยสำคัญ โดยงานของ [3] รายงานความแม่นยำสูงถึง 99.14% และงานของ [4] ให้ผลลัพธ์อยู่ในช่วง 95–98% อย่างไรก็ตาม โมเดลในกลุ่มนี้มักมีความซับซ้อนสูง และต้องใช้ทรัพยากรในการประมวลผลมาก ซึ่งอาจไม่เหมาะสมต่อการนำไปใช้งานจริงในระบบสาธารณสุข

สำหรับงานวิจัยที่ใช้การรวมข้อมูลหลายชุด (multi-dataset) เช่น [5] พบว่าสามารถเพิ่มความสามารถในการ generalization

ได้ โดยให้ค่าความแม่นยำที่ 97.13% อย่างไรก็ตาม การใช้หลาย dataset อาจมีข้อจำกัดด้านความสอดคล้องของข้อมูลและการปรับมาตรฐานข้อมูล

ในขณะเดียวกัน งานวิจัยที่ใช้โมเดลพื้นฐาน เช่น CNN และ VGG19 [6], [7] ยังมีข้อจำกัดด้านความแม่นยำ โดยเฉพาะในกรณีการจำแนกแบบหลายคลาส (multi-class classification) ซึ่งเป็นปัญหาที่มีความซับซ้อนสูงกว่าการจำแนกแบบ binary

เมื่อพิจารณาผลลัพธ์ของงานวิจัยนี้ พบว่าโมเดล Custom CNN ที่เสนอสามารถให้ค่า Accuracy เท่ากับ 97.12% ซึ่งอยู่ในระดับเดียวกับงานวิจัยขั้นสูง เช่น [3] และ [5] แต่มีข้อได้เปรียบที่สำคัญคือ โครงสร้างของโมเดลมีความเรียบง่ายกว่า และไม่ต้องใช้เทคนิค ensemble หรือ multi-dataset ที่ซับซ้อน

นอกจากนี้ โมเดลที่เสนอสามารถรองรับการจำแนกแบบ 4-class ซึ่งมีความท้าทายสูงกว่าการจำแนกแบบ binary ใน

งานวิจัย [2] และ [6] โดยเฉพาะการแยกกลุ่ม Very Mild Demented และ Mild Demented ซึ่งมีลักษณะภาพที่ใกล้เคียงกัน

โดยรวมแล้ว งานวิจัยนี้แสดงให้เห็นว่า การออกแบบโมเดลเฉพาะสำหรับข้อมูล MRI สามารถให้ประสิทธิภาพที่สูงและมีความเหมาะสมต่อการนำไปใช้งานจริงในบริบทของระบบสาธารณสุขไทย มากกว่าการใช้โมเดลทั่วไปที่พัฒนาจาก dataset ธรรมชาติ

5. สรุปผล

งานวิจัยนี้ได้พัฒนาระบบช่วยคัดกรองโรคอัลไซเมอร์ในระยะเริ่มต้นจากภาพถ่ายสมองด้วยคลื่นแม่เหล็กไฟฟ้า (MRI) โดยใช้เทคนิค Deep Learning ซึ่งมุ่งเน้นการประยุกต์ใช้ในบริบทของระบบสาธารณสุขไทย โดยได้ทำการเปรียบเทียบประสิทธิภาพของโมเดลจำนวน 4 โมเดล ได้แก่ Custom CNN (ที่เสนอ), MobileNetV2, ResNet50 และ VGG16 บนชุดข้อมูล Alzheimer's Disease Multiclass Images Dataset จำนวน 6,400 ภาพ

ผลการทดลองแสดงให้เห็นว่าโมเดล Custom CNN ที่เสนอสามารถให้ประสิทธิภาพสูงสุด โดยมีค่า ACC เท่ากับ 97.12% พร้อมทั้งค่า PRE, SEN และ F1 ที่อยู่ในระดับสูงอย่างสม่ำเสมอเมื่อเปรียบเทียบกับโมเดล baseline อื่น ๆ ซึ่งมีค่าความแม่นยำต่ำกว่าอย่างชัดเจน นอกจากนี้ โมเดลยังแสดงให้เห็นถึงความสามารถในการเรียนรู้ที่มีเสถียรภาพ และสามารถลดความสัมพันธ์ระหว่างคลาสได้อย่างมีประสิทธิภาพ โดยเฉพาะในกลุ่มระยะเริ่มต้นของโรค

เมื่อเปรียบเทียบกับงานวิจัยก่อนหน้านี้ พบว่าโมเดลที่เสนอสามารถให้ประสิทธิภาพในระดับเดียวกับงานวิจัยขั้นสูง แต่มีข้อได้เปรียบในด้านความเรียบง่ายของโครงสร้าง และความเหมาะสมต่อการนำไปใช้งานจริง โดยไม่จำเป็นต้องใช้เทคนิคที่มีความซับซ้อนสูง เช่น ensemble learning หรือการใช้หลายชุดข้อมูลร่วมกัน

ในเชิงการประยุกต์ใช้งาน โมเดลที่พัฒนาขึ้นมีศักยภาพในการนำไปใช้เป็นเครื่องมือช่วยคัดกรองโรคอัลไซเมอร์ในระดับบริการสุขภาพเบื้องต้น เช่น โรงพยาบาลชุมชน หรือหน่วยบริการปฐมภูมิ ซึ่งอาจช่วยสนับสนุนการตัดสินใจของบุคลากรทางการแพทย์ และลดภาระของผู้เชี่ยวชาญเฉพาะทาง

อย่างไรก็ตาม งานวิจัยนี้ยังมีข้อจำกัดบางประการ ได้แก่ การใช้ข้อมูลภาพแบบ 2 มิติ (2D MRI) และการทดสอบบนชุด

ข้อมูลเพียงแหล่งเดียว ซึ่งอาจยังไม่ครอบคลุมความหลากหลายของผู้ป่วยจริง

กิตติกรรมประกาศ

ผู้วิจัยขอขอบพระคุณคณะวิทยาศาสตร์และเทคโนโลยีสุขภาพ มหาวิทยาลัยกาฬสินธุ์ที่เอื้อเฟื้อสถานที่ในการทำวิจัย โดยบทความวิจัยนี้เป็นผลจากการศึกษาในรายวิชาการวิเคราะห์ภาพถ่ายทางการแพทย์ หลักสูตรวิทยาศาสตรบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศทางการแพทย์ คณะวิทยาศาสตร์และเทคโนโลยีสุขภาพ มหาวิทยาลัยกาฬสินธุ์ ประจำปีการศึกษา 2568

เอกสารอ้างอิง

- [1] สำนักงานสถิติแห่งชาติ, "รายงานสถิติประชากรประเศไทย ประจำปี 2551," สำนักงานสถิติแห่งชาติ, กรุงเทพฯ, 2551. [Online]. Available: <https://www.nso.go.th>
- [2] K. Chauhan, N. Kishore, N. Goel, and Monika, "Deep learning based multi-stage classification of Alzheimer's disease using T1-weighted MRI images," Franklin Open, vol. 15, p. 100557, 2026/06/01/ 2026, doi: <https://doi.org/10.1016/j.fraope.2026.100557>.
- [3] H. Zhu, J. Huang, K. Chen, X. Ying, and Y. Qian, "multiPI-TransBTS: A multi-path learning framework for brain tumor image segmentation based on multi-physical information," Computers in Biology and Medicine, vol. 191, p. 110148, 2025/06/01/ 2025, doi: <https://doi.org/10.1016/j.combiomed.2025.110148>.
- [4] X. Zhang et al., "Unsupervised brain MRI tumour segmentation via two-stage image synthesis," Medical Image Analysis, vol. 102, p. 103568, 2025/05/01/ 2025, doi: <https://doi.org/10.1016/j.media.2025.103568>.
- [5] V. Shevchenko et al., "A comparative machine learning study of schizophrenia biomarkers derived from functional connectivity," Scientific Reports, vol. 15, no. 1, p. 2849, 2025/01/22 2025, doi: 10.1038/s41598-024-84152-2.
- [6] V. G. Shankar, D. S. Sisodia, and P. Chandrakar, "Alzheimer's stage progression modeling using graph neural network and MRI biomarkers," Neural Computing

- and Applications, 2025/06/05 2025, doi: 10.1007/s00521-025-11353-9.
- [7] H. Schulz and S. Behnke, "Deep Learning," *KI - Künstliche Intelligenz*, vol. 26, no. 4, pp. 357-363, 2012/11/01 2012, doi: 10.1007/s13218-012-0198-z.
- [8] I. Goodfellow, Y. Bengio, and A. Courville, *Convolutional Neural Networks (CNNs)*. Cambridge, MA: MIT Press, 2016, pp. 326-359.
- [9] r. ranokphanuwat, "ระบบควบคุมโรงเรือนผักไฮโดรโปนิคส์อัตโนมัติโดยใช้เทคโนโลยี IoT และเครื่องมือการเรียนรู้เชิงลึก," *วารสารวิทยาการและเทคโนโลยีสารสนเทศ*, vol. 8, no. 2, pp. 74-82, 12/25 2018, doi: 10.14456/jist.2018.8.
- [10] P. Luengvongsakorn and W. Pijitrojana, "Comparative Study of ML and DL in Optical Transceiver Failure Diagnosis," *วารสารวิทยาการและเทคโนโลยีสารสนเทศ*, vol. 15, no. 2, pp. 24-36, 12/27 2025. [Online]. Available: <https://ph02.tci-thaijo.org/index.php/JIST/article/view/260843>.
- [11] ณ. หงษ์บุญมี and จ. กันยาประสิทธิ์, "การวิเคราะห์ปัญหาสุขภาพจากภาพถ่ายเล็บด้วยเทคนิคการเรียนรู้เชิงลึก," *วารสารวิทยาการและเทคโนโลยีสารสนเทศ*, vol. 11, no. 2, pp. 10-20, 12/28 2021, doi: 10.14456/jist.2021.12.
- [12] M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4510-4520, 2018.
- [13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
- [14] C. W. Chun, N. T. Yousir, S. Mohammed Abdulameer, and A. A. Hezam, "Deep Learning Approach for Predicting Prostate Cancer from MRI Images," *Journal of Soft Computing and Data Mining*, vol. 3, no. 2, pp. 1-9, 2022. [Online]. Available: <https://publisher.uthm.edu.my/ojs/index.php/jscdm/article/view/12220>.
- [15] Kaggle. Alzheimer's Disease Multiclass Images Dataset, Kaggle. [Online]. Available: <https://www.kaggle.com/datasets/tourist55/alzheimers-dataset-4-class-of-images>
- [16] J. Islam and Y. Zhang, "Brain MRI analysis for Alzheimer's disease diagnosis using an ensemble system of deep convolutional neural networks," (in eng), *Brain Inform*, vol. 5, no. 2, p. 2, May 31 2018, doi: 10.1186/s40708-018-0080-3.
- [17] N. Raza, A. Naseer, M. Tamoor, and K. Zafar, "Alzheimer Disease Classification through Transfer Learning Approach," (in eng), *Diagnostics (Basel)*, vol. 13, no. 4, Feb 20 2023, doi: 10.3390/diagnostics13040801.
- [18] C. Mahanty et al., "Effective Alzheimer's disease detection using enhanced Xception blending with snapshot ensemble," (in eng), *Sci Rep*, vol. 14, no. 1, p. 29263, Nov 26 2024, doi: 10.1038/s41598-024-80548-2.
- [19] S. Fathi, A. Ahmadi, A. Dehnad, M. Almasi-Dooghaee, and M. Sadegh, "A Deep Learning-Based Ensemble Method for Early Diagnosis of Alzheimer's Disease using MRI Images," (in eng), *Neuroinformatics*, vol. 22, no. 1, pp. 89-105, Jan 2024, doi: 10.1007/s12021-023-09646-2.
- [20] S. B. Shahid et al., "Novel deep learning for multi-class classification of Alzheimer's in disability using MRI datasets," (in English), *Frontiers in Bioinformatics, Original Research* vol. Volume 5 - 2025, 2025-August-20 2025, doi: 10.3389/fbinf.2025.1567219.
- [21] A. Y. Bayahya, H. Banjar, O. Talabay, and S. H. Alamri, "Comparative analysis of multiple deep learning models with mitigation-driven approaches for enhanced Alzheimer's disease classification," (in eng), *Sci Rep*, vol. 15, no. 1, p. 44098, Nov 22 2025, doi: 10.1038/s41598-025-27914-w.

การเปรียบเทียบประสิทธิภาพของแบบจำลองการเรียนรู้ของเครื่อง

ในการจำแนกแหล่งกำเนิดทางภูมิศาสตร์ของมรกต

Comparison of Machine Learning Model Performance in Classifying the Geographical Origin of Emeralds

ชาวลิต จันคันธา^{1*} และ รัฐพล อำพล²

Chawalit Chankhantha^{1*} and Rattaphon Amphon²

¹กองพัฒนาอุตสาหกรรมสร้างสรรค์ กรมส่งเสริมอุตสาหกรรม

²ภาควิชาธรณีวิทยา คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่

¹*Division of Creative Industry Development, Department of Industrial Promotion*

²*Department of Geological Sciences, Faculty of Science, Chiang Mai University*

E-Mail: charles.valydh@gmail.com

Received: November 14, 2025; Revised: June 26, 2026; Accepted: June 26, 2026; Published: June 30, 2026

ABSTRACT – Currently, the classification of emerald origins is of significant commercial importance, as it directly impacts the value and credibility of the gemstones. However, traditional identification methods often encounter challenges due to the similarity in chemical compositions across different origins, leading to analytical inaccuracies. Therefore, this research aims to develop machine learning models to classify the origin of emeralds and to compare the performance of each model. The research methodology followed the 6 phases of the Cross-Industry Standard Process for Data Mining (CRISP-DM). The dataset utilized in this study consists of trace element data of emeralds from 4 origins, totaling 1088 samples: Colombia, Zambia (Kafubu), Zambia (Musakashi), and Afghanistan. The dataset was partitioned into training and testing sets at an 80:20 ratio. Subsequently, feature selection was performed utilizing One-Way ANOVA and Pearson correlation coefficient techniques, resulting in 10 optimal features. To construct the models, 6 machine learning algorithms were selected: Random Forest, Support Vector Machine, Naïve Bayes, Logistic Regression, Linear Discriminant Analysis, and k-Nearest Neighbors. The results revealed that the Random Forest model yielded the highest performance in classifying emerald origins, achieving an accuracy, precision, recall, and F1 Score of 99.54%, 99.55%, 99.54%, and 99.54%, respectively. These findings demonstrate that machine learning techniques can be applied as highly effective tools for accurately classifying gemstone origins, significantly enhancing confidence in gemstone analysis.

KEY WORDS -- Emerald, Machine Learning, Geographical Origin, Model, Classification

บทคัดย่อ -- ปัจจุบันการจำแนกแหล่งกำเนิดของมรกตมีความสำคัญอย่างยิ่งในการค้าอัญมณี เนื่องจากส่งผลโดยตรงต่อมูลค่าและความน่าเชื่อถือของอัญมณี อย่างไรก็ตามพบว่าการตรวจสอบแบบเดิมมักประสบปัญหาความคลาดเคลื่อนขององค์ประกอบทางเคมีของมรกตจากแหล่งต่าง ๆ ทำให้เกิดความคลาดเคลื่อนในการวิเคราะห์ ดังนั้นงานวิจัยนี้จึงมีวัตถุประสงค์เพื่อพัฒนาแบบจำลองการเรียนรู้ของเครื่องในการจำแนกแหล่งกำเนิดของมรกต และเปรียบเทียบประสิทธิภาพของแบบจำลองแต่ละประเภท การดำเนินงานวิจัยใช้กระบวนการมาตรฐานสำหรับการทำเหมืองข้อมูล (CRISP-DM)

จำนวน 6 ขั้นตอน ชุดข้อมูลที่ใช้ในการศึกษาเป็นข้อมูลธาตุร่องรอยของมรกตจาก 4 แหล่ง จำนวน 1088 ตัวอย่าง ได้แก่ โคลอมเบีย แซมเบีย (คาฟู) แซมเบีย (มุซาคาซี) และอัฟกานิสถาน ชุดข้อมูลถูกแบ่งออกเป็นข้อมูลฝึกสอนและข้อมูลทดสอบในอัตราส่วน 80:20 จากนั้นทำการคัดเลือกคุณลักษณะด้วยเทคนิค One-Way ANOVA และค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน ได้คุณลักษณะที่เหมาะสมจำนวน 10 ตัวแปร สำหรับการสร้างแบบจำลอง การพัฒนาแบบจำลองการเรียนรู้ของเครื่องจำนวน 6 ประเภท ได้แก่ ป่าสุ่ม ซัพพอร์ตเวกเตอร์แมชชีน นาอ์ฟเบย์ การถดถอยโลจิสติกส์ การวิเคราะห์จำแนกประเภทเชิงเส้น และเพื่อนบ้านใกล้ที่สุด-เค ผลการวิจัยพบว่า แบบจำลองป่าสุ่มให้ประสิทธิภาพสูงสุดในการจำแนกแหล่งกำเนิดของมรกต โดยมีค่าความถูกต้อง ความแม่นยำ ความความระลึก และค่า F1 Score เท่ากับร้อยละ 99.54 99.55 99.54 และ 99.54 ตามลำดับ จากผลการวิจัยแสดงให้เห็นว่าเทคนิคการเรียนรู้ของเครื่องสามารถนำมาใช้เป็นเครื่องมือที่มีประสิทธิภาพในการจำแนกแหล่งกำเนิดของอัญมณีได้อย่างแม่นยำ และสามารถช่วยเพิ่มความเชื่อมั่นในการวิเคราะห์อัญมณีได้อย่างมีนัยสำคัญ

คำสำคัญ -- มรกต, การเรียนรู้ของเครื่อง, แหล่งกำเนิดทางภูมิศาสตร์, แบบจำลอง, การจำแนก

1. บทนำ

อุตสาหกรรมอัญมณีและเครื่องประดับมีบทบาทสำคัญในการขับเคลื่อนเศรษฐกิจของประเทศ เนื่องจากเป็นกลุ่มสินค้าส่งออกสำคัญที่อยู่ในอันดับต้น ๆ ติดต่อกันมาหลายปี ปัจจุบันการระบุแหล่งกำเนิดของอัญมณี (Origin Determination of Gemstones) มีความสำคัญต่อการค้าขายอัญมณีเป็นอย่างมาก เนื่องจากแหล่งที่มาส่งผลต่อมูลค่าของอัญมณี โดยเฉพาะกลุ่มอัญมณีที่มีความต้องการสูง เช่น ทับทิมจากเหมืองโมกอก ประเทศเมียนมาร์ มักมีราคาสูงกว่าทับทิมจากแหล่งอื่นหลายเท่า เมื่อเทียบน้ำหนักและคุณภาพใกล้เคียงกัน จึงเป็นความท้าทายของนักวิเคราะห์อัญมณีที่ต้องอาศัยประสบการณ์ความเชี่ยวชาญสูง ในช่วงปี ค.ศ. 1950 เริ่มมีการระบุแหล่งกำเนิดในใบรับรองอัญมณีโดยห้องปฏิบัติการระดับโลก ได้แก่ Gübelin Gem Lab และ Schweizerische Stiftung für Edelstein-Forschung (SSEF) ประเทศสวิตเซอร์แลนด์ ซึ่งในขณะนั้นการค้นพบแหล่งอัญมณียังไม่มากเท่าปัจจุบัน เช่น แหล่งมรกตส่วนใหญ่จะมาจากประเทศโคลอมเบีย และรัสเซีย เป็นต้น [1] การระบุแหล่งกำเนิดของอัญมณียังเป็นกลไกที่ช่วยสร้างความโปร่งใสและตรวจสอบย้อนกลับได้ (Traceability) ตลอดห่วงโซ่อุปทาน สอดคล้องกับข้อตกลงทางการค้าระหว่างประเทศในด้านความรับผิดชอบต่อสังคมและสิ่งแวดล้อม โดยมีองค์กร Responsible Jewellery Council (RJC) ทำหน้าที่กำกับดูแล

เพื่อให้มั่นใจว่าอัญมณีแหล่งที่มาที่ถูกต้องตามกฎหมาย มาจากธุรกิจที่ไม่เกี่ยวข้องกับการละเมิดสิทธิมนุษยชน และไม่ส่งผลกระทบต่อสิ่งแวดล้อม นอกจากนี้ยังสร้างความมั่นใจให้กับผู้บริโภคอีกด้วย [2] ซึ่งทั้งหมดนี้เป็นรากฐานสำคัญในการพัฒนาอุตสาหกรรมอัญมณีและเครื่องประดับให้เติบโตอย่างยั่งยืน

ในการจำแนกแหล่งกำเนิดของมรกตนั้นต้องอาศัยทั้งประสบการณ์และความเชี่ยวชาญของนักอัญมณี โดยทั่วไปจะพิจารณาจากข้อมูลเอกลักษณ์ทางเคมี ข้อมูลทางสเปกโทรสโกปี เช่น UV-Vis-NIR Spectroscopy และลักษณะเฉพาะของมลทินภายใน (Inclusions) เพื่อให้ผลที่แม่นยำมากขึ้น [3] แต่เนื่องจากมรกตเกิดในสภาพแวดล้อมทางธรณีวิทยาคล้ายคลึงกัน บ่อยครั้งที่เจอพบข้อมูลมีความใกล้เคียงกัน ส่งผลต่อความแม่นยำในการตัดสินใจ จากปัญหาดังกล่าว ผู้วิจัยเล็งเห็นถึงความสำคัญในการพัฒนาวิธีการจำแนกแหล่งกำเนิดของมรกตให้มีประสิทธิภาพ โดยนำเทคนิคการเรียนรู้แบบมีผู้สอน (Supervised Learning) ของ ปัญญาประดิษฐ์ (Artificial Intelligence) มาประยุกต์ใช้กับข้อมูลผลวิเคราะห์องค์ประกอบทางเคมีของมรกตจากแหล่งที่มีศักยภาพในเชิงพาณิชย์ ทำการฝึกฝนและทดสอบข้อมูลผ่านแบบจำลองการเรียนรู้ของเครื่องจำนวน 6 ประเภท ได้แก่ ป่าสุ่ม (Random Forests: RF) ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machines: SVM) นาอ์ฟเบย์ (Naïve Bayes: NB) การถดถอยโลจิสติกส์ (Logistic

Regression: LR) การวิเคราะห์การจำแนกเชิงเส้น (Linear Discriminant Analysis: LDA) และเพื่อนบ้านใกล้ที่สุด-เค (k-Nearest Neighbors: k-NN) การเลือกใช้แบบจำลองที่มีประสิทธิภาพที่สุด จะเป็นเครื่องมือสนับสนุนนักอัญมณีให้สามารถจำแนกแหล่งกำเนิดของมรกตได้อย่างถูกต้อง แม่นยำ ลดความผิดพลาดในการสำแดงแหล่งที่มา และยกระดับความน่าเชื่อถือในอุตสาหกรรมอัญมณีและเครื่องประดับ ทั้งนี้ ผลการวิจัยยังสามารถนำไปพัฒนาต่อยอดเทคนิคการเรียนรู้ของเครื่องเพื่อจำแนกแหล่งที่มาของอัญมณีชนิดอื่น ๆ ที่มีมูลค่าสูงในอนาคตได้อีกด้วย

2. เอกสารและงานวิจัยที่เกี่ยวข้อง

มรกต (Emerald) เป็นอัญมณีสีเขียว อยู่ในสปีชีส์เบริล (Beryl species) มีสูตรเคมี $\text{Be}_3\text{Al}_2\text{Si}_6\text{O}_{18}$ แสดงผลึกในระบบเฮกซะโกนอล (Hexagonal System) มีความแข็ง 7-7.5 ตามมาตราวัดของโมส์ โครงสร้างผลึกประกอบด้วยโครงข่ายของ SiO_4 , BeO_4 และ AlO_6 มีลักษณะคล้ายขบวนเรียงซ้อนกันเป็นชั้น ๆ จนเกิดเป็นท่อ (Chanel Sites) ขนานกันตามแนวแกนผลึก ท่อดังกล่าวนี้มีเส้นผ่านศูนย์กลางขนาด 2.80-5.10 อังสตรอม สามารถกักเก็บอัลคาไลไอออนขนาดเล็กจนถึงโมเลกุลใหญ่ เช่น Na^+ , Li^+ , K^+ , Rb^+ , Cs^+ , H_2O , CO_2 และก๊าซเฉื่อยบางชนิด [4] สีเขียวของมรกตเกิดจากธาตุโครเมียม (Cr) และวานาเดียม (V) เข้าไปแทนที่อะลูมิเนียม (Al) ในโครงสร้าง ส่วนธาตุเหล็ก (Fe) หากมีปริมาณสูงจะทำให้มรกตมีสีเข้ม เช่น มรกตจากประเทศแซมเบียที่มีสีเขียวเข้ม การรวมกันของธาตุเหล่านี้ส่งผลให้มรกตในธรรมชาติมีเจดสีตั้งแต่เขียวอ่อน ถึงเขียวเข้ม [5] โดยธรรมชาติมรกตมักมีรอยแตกร้าวภายในเนื้อพลอยอยู่แล้ว ดังนั้นจึงมักมีการปรับปรุงคุณภาพด้วยการอุดรอยแตกร้าวโดยใช้วัสดุที่มีค่าดัชนีหักเหของแสงใกล้เคียงกับมรกต (1.577-1.583) เช่น น้ำมันซิดาร์วูด เรซิน และพอลิเมอร์ โดยทั่วไปการอุดรอยแตกร้าวด้วยน้ำมันซิดาร์วูดได้รับการยอมรับมากกว่าการอุดด้วยเรซิน มรกตที่มีการค้าขายในปัจจุบันส่วนใหญ่มาจากประเทศโคลอมเบีย บราซิล แซมเบีย รัสเซีย มาดากัสการ์ และเอธิโอเปีย โดยเฉพาะมรกตจากเหมืองมุโซ ประเทศโคลอมเบีย ที่มีชื่อเรียกทางการค้าว่า Muzo Green เป็นมรกตคุณภาพสูง มีเนื้อค่อนข้างใส มีสีเขียวสด เป็นที่ต้องการและมี

ชื่อเสียงมาอย่างยาวนาน โดยในปี 2024 มรกตจากโคลอมเบียที่มีชื่อว่า The Aga Khan Emerald น้ำหนัก 37 กระรัต ถูกประมูลไปในราคาสูงถึง 305 ล้านบาท

ในด้านธรณีวิทยา มรกตเป็นอัญมณีที่เกิดขึ้นได้ค่อนข้างยากในธรรมชาติ เนื่องจากธาตุเบริลเลียม (Be) ที่เป็นองค์ประกอบหลักของมรกต พบได้น้อยมากในเปลือกโลก โดยมีความเข้มข้นเฉลี่ยเพียง 2.1 ส่วนในล้านส่วนโดยน้ำหนัก (ppmw) มรกตส่วนใหญ่เกิดในหินเพกมาไทต์ซึ่งมีธาตุเบริลเลียม อะลูมิเนียม และซิลิกา เข้าไปสัมผัสและทำปฏิกิริยาเคมีกับหินภูเขาไฟชนิดเมฟิก-อัลตราเมฟิก ซึ่งมีธาตุโครเมียม และวานาเดียม ในสภาวะที่อุณหภูมิและความดันเหมาะสมจนเกิดการตกผลึกของมรกตขึ้นมา อย่างไรก็ตามมรกตจากประเทศโคลอมเบียมีกระบวนการเกิดที่แตกต่างจากแหล่งอื่น โดยเกิดจากกระบวนการหมุนเวียนของน้ำแร่ร้อน (Hydrothermal Fluid) ภายในหินดินดานสีดำ (Black Shales) ในช่วงอุณหภูมิ 330–450 องศาเซลเซียส และความดันประมาณ 2–3.5 กิโลบาร์ [6]

การเรียนรู้ของเครื่อง (Machine Learning) เป็นสาขาย่อยของปัญญาประดิษฐ์ (Artificial Intelligence) ที่มุ่งพัฒนาเทคนิคและวิธีการให้คอมพิวเตอร์สามารถเรียนรู้จากการป้อนข้อมูลให้ระบบประมวลผลและค้นหาความสัมพันธ์ระหว่างข้อมูลเข้ากับข้อมูลขาออกเพื่อสร้างขั้นตอนการประมวลผลหรือแบบจำลองสำหรับใช้กับข้อมูลใหม่ในอนาคต การเรียนรู้ของเครื่องแบ่งออกเป็น 3 ประเภทหลัก ได้แก่ การเรียนรู้แบบมีผู้สอน (Supervised Learning) การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) และการเรียนรู้แบบเสริมแรง (Reinforcement Learning) [7]

การประยุกต์ใช้เทคนิคการเรียนรู้ของเครื่องเพื่อจำแนกแหล่งกำเนิดทางภูมิศาสตร์ของอัญมณีได้มีการศึกษาและรายงานว่ามีความแม่นยำในระดับสูง เช่น การศึกษาของ Blodgett and Shen [8] รายงานการประยุกต์ใช้เทคนิคการวิเคราะห์การจำแนกกลุ่มเชิงเส้น (Linear Discriminant Analysis: LDA) ร่วมกับข้อมูลธาตุร่องรอย (Trace element) ที่ได้จากการวิเคราะห์ด้วยเทคนิค LA-ICP-MS เพื่อจำแนกแหล่งกำเนิดของอัญมณี ได้แก่ ทับทิม ไพลิน และพาราอับัว รวมถึงประยุกต์ใช้ในการตรวจแยกเพชรสีที่ผ่านการปรับปรุง

คุณภาพด้วยกระบวนการ High Pressure High Temperature (HPHT) ออกจากเพชรสังเคราะห์ ซึ่งผลการวิเคราะห์แสดงให้เห็นถึงความแม่นยำในการจำแนกได้เป็นอย่างดี

Luo et al. [9] จำแนกแหล่งกำเนิดของหยกเนไฟรต์จาก 8 แหล่งสำคัญในภูมิภาคเอเชียตะวันออกเฉียงใต้ โดยใช้ข้อมูลของธาตุจำนวน 8 ชนิด ได้แก่ Ti, V, Cr, Mn, Fe, Co, Ni และ Cu ร่วมกับเทคนิคการวิเคราะห์จำแนกกลุ่มเชิงเส้นแบบทวิภาควนซ้ำ (Iterative Binary Linear Discriminant Analysis) ผลการศึกษาพบว่าแบบจำลองดังกล่าวให้ความถูกต้องเฉลี่ยสูงถึงร้อยละ 94.40

Homkrajae et al. [10] ศึกษาเมทริกซ์จำนวน 225 ตัวอย่าง ซึ่งประกอบด้วยเมทริกซ์แร่จากประเทศจีน เมทริกซ์แร่ และเมทริกซ์จากสหรัฐอเมริกา โดยใช้เทคนิค LA-ICP-MS เพื่อวิเคราะห์ธาตุร่องรอย และทำการจำแนกด้วยเทคนิค LDA พบว่าแบบจำลองสามารถจำแนกประเภทของเมทริกซ์ได้ถูกต้องที่ร้อยละ 85 ซึ่งแสดงให้เห็นถึงศักยภาพของการใช้ข้อมูลธาตุร่องรอยร่วมกับการวิเคราะห์เชิงสถิติขั้นสูง

Wang and Krzemnicki [11] ประยุกต์ใช้เทคนิค LA-ICP-TOF-MS ในการวิเคราะห์องค์ประกอบธาตุของมรกตจำนวน 168 ตัวอย่าง และใช้เทคนิค t-Distributed Stochastic Neighbor Embedding (t-SNE) ซึ่งเป็นอัลกอริทึมการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) เพื่อจำแนกแหล่งกำเนิดของมรกต ผลการวิจัยชี้ว่า t-SNE ให้ประสิทธิภาพสูงกว่าเทคนิคการวิเคราะห์องค์ประกอบหลัก (Principal Component Analysis: PCA) ในการจำแนกกลุ่มข้อมูล

Zhong et al. [12] ศึกษาการจำแนกแหล่งกำเนิดของเพริโดต (Peridot) จำนวน 262 ตัวอย่างจากสาธารณรัฐประชาชนจีนและเกาหลีเหนือ โดยใช้ข้อมูลทางเคมีจากการวิเคราะห์ด้วยเทคนิค LA-ICP-MS เพื่อพัฒนาแบบจำลองการเรียนรู้ของเครื่อง 6 ประเภท ได้แก่ LDA, Decision Tree, Random Forests, k-Nearest Neighbors (k-NN), Support Vector Machines และ Logistic Regression ในการศึกษาได้แบ่งข้อมูลออกเป็นชุดฝึกสอนและชุดทดสอบในอัตรา 70:30 ผลการศึกษาพบว่าแบบจำลอง SVM ให้ค่าความถูกต้องสูงสุด

นอกจากนี้ Seneewong-Na-Ayutthaya et al. [13] ประยุกต์ใช้เทคนิคการเรียนรู้ของเครื่องเพื่อจำแนกแหล่งกำเนิด

ของคอร์นัม โดยใช้ข้อมูลองค์ประกอบทางเคมีที่ได้จากการวิเคราะห์ด้วยเทคนิค LA-ICP-MS ผลการศึกษาพบว่าแบบจำลองป่าสุ่ม (Random Forests) สามารถจำแนกแหล่งกำเนิดของคอร์นัมได้อย่างมีประสิทธิภาพ โดยให้ค่าความถูกต้อง ร้อยละ 82.50 ในกลุ่มทับทิม และร้อยละ 92.50 ในกลุ่มไพลิน อย่างไรก็ตามการใช้ข้อมูลเคมีจากเทคนิค EDXRF ให้ความแม่นยำน้อยกว่า

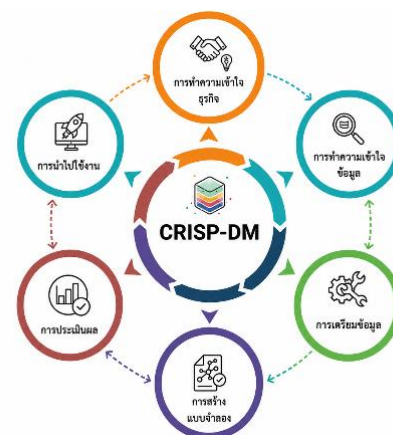
3. วิธีดำเนินการวิจัย

3.1 เครื่องมือการวิจัย

งานวิจัยนี้ใช้เครื่องมือ Google Collaboratory บนเว็บไซต์ <https://colab.research.google.com> ซึ่งเขียนด้วยภาษาไพธอน ซึ่งสามารถเข้าถึง GPU และรองรับการประมวลผลข้อมูล และการเรียนรู้ของเครื่องผ่านการทำงานจากไลบรารีต่าง ๆ ได้

3.2 ขั้นตอนการดำเนินการวิจัย

งานวิจัยนี้นำกระบวนการมาตรฐานสำหรับการทำเหมืองข้อมูล (Cross-Industry Standard Process for Data Mining: CRISP-DM) มาประยุกต์ใช้ ซึ่งประกอบ 6 ขั้นตอน [14] ดังแสดงในรูปที่ 1



รูปที่ 1. การดำเนินงานวิจัยตามกระบวนการของ CRISP-DM 6 ขั้นตอน

3.2.1 การทำความเข้าใจธุรกิจ

การระบุแหล่งกำเนิดของมรกตได้อย่างถูกต้องมีความสำคัญมากในทางการค้าเนื่องจากส่งผลโดยตรงต่อมูลค่าความโปร่งใส และความน่าเชื่อถือ ปัจจุบันอาศัยข้อมูลทางเคมี

และความเชี่ยวชาญของนักวิเคราะห์ข้อมูลเป็นผู้ชี้ขาด แต่มักประสบปัญหาบางแหล่งมีข้อมูลที่คล้ายคลึงกัน ทำให้ไม่สามารถระบุแหล่งได้ชัดเจน งานวิจัยนี้จึงมุ่งพัฒนาแบบจำลองการเรียนรู้ของเครื่องแบบมีผู้สอน จำนวน 6 ประเภท ได้แก่ การวิเคราะห์จำแนกประเภทเชิงเส้น (Linear Discriminant Analysis: LDA) การถดถอยโลจิสติกส์ (Logistic Regression) นาอิวเบย์ (Naïve Bayes) เพื่อนบ้านใกล้ที่สุด-เค (k-Nearest Neighbor, k-NN) ป่าสุ่ม (Random Forest) และซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine, SVM) เพื่อพัฒนาแบบจำลองที่มีประสิทธิภาพสูงสุดในการจำแนกแหล่งกำเนิดของมรดกได้อย่างถูกต้องและแม่นยำ

3.2.2 การทำความเข้าใจข้อมูล

รวบรวมชุดข้อมูลผลการวิเคราะห์ธาตุร่องรอย (Trace elements) ของมรดกที่ได้จากการวิเคราะห์ด้วยเทคนิค Laser Ablation Inductively Coupled Plasma Mass Spectrometry (LA-ICP-MS) [15] ประกอบด้วยตัวแปรตาม (Dependent variable) ซึ่งเป็นแหล่งมรดกจำนวน 4 แหล่ง ได้แก่ Afghanistan (116 ตัวอย่าง) Colombia (306 ตัวอย่าง) Zambia (Kafubu) (393 ตัวอย่าง) และ Zambia (Musakashi) (273 ตัวอย่าง) ส่วนตัวแปรอิสระ (Independent variables) เป็นข้อมูลเชิงปริมาณที่แสดงค่าความเข้มข้นของธาตุร่องรอยแต่ละชนิด โดยวัดค่าออกมาในหน่วยส่วนในล้านส่วนโดยน้ำหนัก (Parts per million by weight: ppmw)

3.2.3 การเตรียมข้อมูล

1) การทำความสะอาดและติดป้ายกำกับข้อมูล

ตรวจสอบข้อมูลที่สูญหาย และแทนค่าสูญหายดังกล่าวด้วยวิธีแทนค่าหลายค่า (Imputation) ปรับข้อมูลให้สมมาตรด้วยวิธี Min-max normalization เพื่อให้สามารถนำข้อมูลเข้าสู่กระบวนการวิเคราะห์ได้ ทำการลบข้อมูลในบางคอลัมน์ที่ไม่เกี่ยวข้อง จากนั้นทำการจัดการข้อมูลที่ไม่สมดุลด้วยวิธี SMOTE (Synthetic Minority Over-Sampling Technique) โดยอาศัยหลักการของเพื่อนบ้านใกล้ที่สุด-เค [16] พร้อมกำหนดป้ายกำกับข้อมูลด้วยเทคนิค Label Encoding

2) การคัดเลือกคุณลักษณะ

คัดเลือกคุณลักษณะที่เหมาะสมด้วยเทคนิคการคัดกรองข้อมูล (Filter Method) โดยประยุกต์ใช้การวิเคราะห์ความ

แปรปรวนทางเดียว (One-Way ANOVA) เพื่อประเมินความแตกต่างของค่าเฉลี่ยของตัวแปรอิสระระหว่างกลุ่มของมรดกแต่ละแหล่ง โดยคัดเลือกเฉพาะตัวแปรที่มีความแตกต่างอย่างมีนัยสำคัญทางสถิติ ($p\text{-value} < .05$) ควบคู่กับการพิจารณาค่าสถิติ F (F-score) เพื่อจัดลำดับความสำคัญของคุณลักษณะ ทั้งนี้ตัวแปรที่มีค่า F-score สูงแสดงถึงความสามารถในการจำแนกค่าเฉลี่ยของตัวแปรอิสระระหว่างกลุ่มได้ดีกว่า จึงมีแนวโน้มเป็นคุณลักษณะที่มีความสำคัญต่อการสร้างแบบจำลองการจำแนกประเภท จากนั้นทำการวิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน (Pearson correlation coefficient) ผ่านแผนภาพความร้อน (Correlation heatmap) เพื่อตรวจสอบปัญหาความสัมพันธ์พหุร่วม (Multicollinearity) การที่ตัวแปรอิสระมีความสัมพันธ์เชิงเส้นระหว่างกันในระดับที่สูงเกินไปซึ่งอาจส่งผลกระทบต่อความแม่นยำของแบบจำลอง

3) การแบ่งชุดข้อมูล

ชุดข้อมูลทั้งหมด ถูกแบ่งออกเป็นชุดข้อมูลฝึกสอน (Training Set) และชุดข้อมูลทดสอบ (Test Set) ในอัตราส่วน 80:20 โดยชุดข้อมูลฝึกสอนจะถูกนำไปใช้สำหรับการสร้างแบบจำลอง และปรับแต่งค่าไฮเปอร์พารามิเตอร์ ในขณะที่ชุดข้อมูลทดสอบจะนำไปใช้ประเมินประสิทธิภาพขั้นสุดท้ายของแบบจำลอง

3.2.4 การสร้างแบบจำลอง

นำชุดข้อมูลฝึกสอนมาสร้างแบบจำลองการเรียนรู้ของเครื่องเพื่อการจำแนกแหล่งกำเนิดของมรดก 6 ประเภท ดังนี้

1) การวิเคราะห์จำแนกเชิงเส้น (Linear Discriminant Analysis: LDA) เป็นวิธีการเรียนรู้แบบมีผู้สอน (Supervised Learning) ที่ใช้สำหรับการจำแนกข้อมูลตั้งแต่ 2 กลุ่มขึ้นไป เพื่อค้นหาการรวมเชิงเส้นของตัวแปรที่สามารถเพิ่มความแตกต่างระหว่างกลุ่มให้มากที่สุด และลดความแปรปรวนภายในกลุ่มให้ต่ำที่สุด ผ่านการวิเคราะห์เมทริกซ์ความแปรปรวนระหว่างกลุ่ม (Between-class Scatter Matrix) และเมทริกซ์ความแปรปรวนภายในกลุ่ม (Within-class Scatter Matrix) จากนั้นจึงแก้ปัญหา Eigenvalue เพื่อหาแกนจำแนกเชิงเส้น (Linear Discriminants) ที่เหมาะสม โดยค่า Eigenvalue ใช้สะท้อนความสามารถในการจำแนกของแต่ละแกนจำแนก ซึ่งค่า

Eigenvalue ที่สูงกว่าจะสามารถอธิบายความแตกต่างระหว่างกลุ่มได้ดีกว่า [17]

2) การถดถอยโลจิสติกส์ (Logistic Regression: LR) เป็นแบบจำลองเชิงเส้นสำหรับการจำแนกประเภท โดยสร้างสมการเชิงเส้นจากตัวแปรอิสระเพื่อประมาณความน่าจะเป็นของการเป็นสมาชิกในแต่ละกลุ่ม สำหรับปัญหาการจำแนกหลายกลุ่ม งานวิจัยนี้ใช้ Multinomial Logistic Regression ซึ่งประยุกต์ใช้ฟังก์ชัน Softmax ในการแปลงค่าผลรวมเชิงเส้นให้เป็นค่าความน่าจะเป็นของแต่ละกลุ่ม โดยผลรวมของความน่าจะเป็นของทุกกลุ่มมีค่าเท่ากับ 1 และกำหนดผลการจำแนกตามกลุ่มที่มีค่าความน่าจะเป็นสูงสุด [18]

3) นาอิวเบย์ (Naïve Bayes) เป็นอัลกอริทึมสำหรับการจำแนกประเภทที่อาศัยทฤษฎีบทของเบย์ (Bayes' theorem) โดยคำนวณความน่าจะเป็นแบบมีเงื่อนไขของการเป็นสมาชิกในแต่ละกลุ่มจากค่าของคุณลักษณะ ภายใต้สมมติฐานว่าคุณลักษณะทุกตัวเป็นอิสระต่อกันแบบมีเงื่อนไข (Conditional Independence) [19] เนื่องจากข้อมูลที่ใช้ในงานวิจัยนี้เป็นข้อมูลเชิงปริมาณ จึงเลือกใช้ Gaussian Naïve Bayes (GNB) ซึ่งสมมติว่าค่าของคุณลักษณะแต่ละตัวภายในแต่ละกลุ่มมีการแจกแจงแบบปกติ (Gaussian Distribution)

4) เพื่อนบ้านใกล้ที่สุด-เค (k-Nearest Neighbors) เป็นอัลกอริทึมแบบไม่อิงพารามิเตอร์ (Non-parametric Algorithm) [20] ที่ใช้หลักการวัดระยะห่างระหว่างข้อมูลเพื่อประเมินความใกล้เคียง เมื่อมีข้อมูลใหม่เข้าสู่แบบจำลอง จะคำนวณระยะห่างระหว่างข้อมูลดังกล่าวกับข้อมูลทั้งหมดในชุดฝึก แล้วเลือกข้อมูลที่อยู่ใกล้ที่สุดจำนวน k ตัวอย่าง (Nearest Neighbors) เพื่อใช้ในการพิจารณาจำแนกประเภท โดยกำหนดผลการจำแนกเป็นกลุ่มที่มีจำนวนสมาชิกมากที่สุดในบรรดาเพื่อนบ้านทั้ง k ตัวอย่าง

5) ป่าสุ่ม (Random Forest) เป็นอัลกอริทึมการเรียนรู้แบบกลุ่ม (Ensemble Learning) ที่สร้างแบบจำลองการจำแนกประเภทจากการรวมผลลัพธ์ของต้นไม้ตัดสินใจ (Decision Trees) หลายต้น แต่ละต้นถูกฝึกด้วยชุดข้อมูลย่อยที่ได้จากการสุ่มตัวอย่างแบบ Bootstrap (Bagging) และสุ่มเลือกบางส่วนของคุณลักษณะในแต่ละโหนดเพื่อใช้ในการแบ่งข้อมูล การแบ่งโหนดจะเลือกเงื่อนไขที่ทำให้ค่าความไม่บริสุทธิ์

(Impurity) ลดลงมากที่สุด ซึ่งสามารถวัดได้ด้วย Gini Impurity หรือ Entropy จนกว่าจะถึงเกณฑ์หยุดที่กำหนด ผลการจำแนกประเภทสุดท้ายได้จากการโหวตเสียงข้างมาก (Majority Voting) ของต้นไม้ตัดสินใจทั้งหมด [19]

6) ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM) เป็นอัลกอริทึมสำหรับการจำแนกประเภทที่มุ่งค้นหาเส้นไฮเปอร์เพลน (Hyperplane) ที่สามารถแบ่งแยกกลุ่มข้อมูลได้ดีที่สุด โดยทำให้ระยะขอบ (Margin) ระหว่างกลุ่มข้อมูลมีค่าสูงสุด (Maximum Margin) ข้อมูลที่อยู่ใกล้กับเส้นไฮเปอร์เพลนมากที่สุดเรียกว่า ซัพพอร์ตเวกเตอร์ (Support Vectors) ซึ่งเป็นข้อมูลสำคัญที่ใช้กำหนดตำแหน่งของเส้นแบ่งกลุ่ม ในกรณีที่ข้อมูลไม่สามารถแบ่งแยกเชิงเส้นได้ จะประยุกต์ใช้ฟังก์ชันเคอร์เนล (Kernel Function) เช่น Radial Basis Function (RBF) Kernel เพื่อคำนวณข้อมูลในปริภูมิคุณลักษณะที่มีมิติสูงขึ้น ทำให้สามารถแบ่งแยกกลุ่มข้อมูลได้อย่างมีประสิทธิภาพ [22]

สำหรับการปรับแต่งพารามิเตอร์ (Hyperparameter Tuning) งานวิจัยนี้ใช้เทคนิค Grid Search ร่วมกับการตรวจสอบความถูกต้องแบบไขว้ชนิด 5-fold cross-validation เพื่อค้นหาชุดพารามิเตอร์ที่ให้ค่าความแม่นยำ (Accuracy) สูงที่สุดสำหรับแต่ละอัลกอริทึมการเรียนรู้ของเครื่อง โดยกำหนดค่า $k=5$ เนื่องจากเป็นค่าที่ให้ความสมดุลระหว่างความน่าเชื่อถือในการประเมินประสิทธิภาพของแบบจำลองและการะในการคำนวณ [23]

4.2.5 การประเมินผล

ประเมินประสิทธิภาพของแบบจำลองในการจำแนกแหล่งกำเนิดของมรดก โดยนำไปทดสอบกับชุดข้อมูลทดสอบ (Testing set) เครื่องมือหลักที่ใช้ในการวัดประสิทธิภาพของการจำแนกประเภทแบบหลายกลุ่ม (Multiclass Classification) คือ เมทริกซ์ความสับสน (Confusion Matrix) ซึ่งจะแสดงผลการเปรียบเทียบระหว่างค่าที่แบบจำลองทำนายได้กับค่าจริงในแต่ละกลุ่มเป้าหมาย การวัดผลจะใช้วิธี One-vs-Rest (OVR) เพื่อพิจารณากลุ่มที่สนใจเทียบกับกลุ่มอื่นๆ ทั้งหมด โดยทำการพิจารณาจาก 4 ค่า ได้แก่ ค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าความถ่วงดุล (F1 Score) [21] สามารถคำนวณได้จากสมการที่ (1) – (5)

1) ค่าความถูกต้อง (Accuracy) เป็นตัววัดที่ใช้ประเมินสัดส่วนของจำนวนตัวอย่างที่แบบจำลองจำแนกได้ถูกต้องต่อจำนวนตัวอย่างทั้งหมด โดยพิจารณาผลการจำแนกของทุกกลุ่ม หากค่า Accuracy เข้าใกล้ร้อยละ 100 แสดงว่าแบบจำลองมีประสิทธิภาพในการจำแนกข้อมูลสูง

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)} \quad (1)$$

2) ค่าความแม่นยำ (Precision)

$$Precision = \frac{TP}{(TP + FP)} \quad (2)$$

3) ค่าความระลึก (Recall)

$$Recall = \frac{TP}{(TP + FN)} \quad (3)$$

4) ค่าประสิทธิภาพโดยรวม (F1 Score)

$$F1 = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} \quad (4)$$

โดยที่

True Positive (TP) คือ จำนวนข้อมูลที่เป็นของกลุ่มที่พิจารณา และแบบจำลองสามารถจำแนกได้ถูกต้องว่าเป็นกลุ่มนั้น

True Negative (TN) คือ จำนวนข้อมูลที่ไม่ใช่กลุ่มที่พิจารณา และแบบจำลองสามารถจำแนกได้ถูกต้องว่าไม่ใช่กลุ่มนั้น

False Positive (FP) คือ จำนวนข้อมูลที่ไม่ใช่กลุ่มที่พิจารณา แต่แบบจำลองจำแนกผิดว่าเป็นกลุ่มนั้น

False Negative (FN) คือ จำนวนข้อมูลที่เป็นของกลุ่มที่พิจารณา แต่แบบจำลองจำแนกผิดว่าเป็นกลุ่มอื่น

3.2.6 การนำไปใช้งาน

แบบจำลองที่ได้รับการพัฒนาสามารถเป็นเครื่องมือสนับสนุนในการจำแนกแหล่งกำเนิดของมรกตได้อย่างถูกต้องและแม่นยำมากขึ้น

4. ผลการวิจัย

4.1 ผลการจัดการกับข้อมูล

จากการตรวจสอบข้อมูลพบว่ามีจำนวนข้อมูลทั้งหมด 1,088 แถว และ 18 คอลัมน์ (รูปที่ 2) โดยคอลัมน์ Origin คือ ตัวแปรตาม (Dependent Variables: X) ประกอบด้วยแหล่งมรกตจำนวน 4 แหล่ง ได้แก่ Afghanistan (116 ตัวอย่าง) Colombia (306 ตัวอย่าง) Zambia (Kafubu) (393 ตัวอย่าง) และ Zambia (Musakashi) (273 ตัวอย่าง) ไม่พบข้อมูลสูญหาย ส่วนตัวแปรต้น (Independent Variables: y) คือ ค่าความเข้มข้นของธาตุร่องรอย จำนวน 16 ชนิด ซึ่งเป็นข้อมูลเชิงปริมาณ ได้แก่ Li, Na, Mg, K, Ca, Sc, V, Cr, Mn, Fe, Co, Ni, Zn, Ga, Rb และ Cs แสดงในหน่วย Parts per million by weight (ppmw) อย่างไรก็ตามในบางตัวอย่างแสดงค่าความเข้มข้นของธาตุเป็น “b.d.” (Below Detection Limit) ซึ่งเกิดจากความเข้มข้นของธาตุต่ำกว่าขีดจำกัดการตรวจวัดของเครื่องมือ (Limit of Detection: LOD) จากการตรวจสอบข้อมูลพบเฉพาะในธาตุ K, Ca, Co, Ni และ Zn และเนื่องจากไม่ทราบค่า LOD ของแต่ละธาตุอย่างชัดเจนจึงไม่สามารถประมาณค่าทดแทนด้วยวิธีเชิงสถิติ เช่น การแทนด้วยค่ากึ่งหนึ่งของ LOD (LOD/2) ได้อย่างเหมาะสม ดังนั้นงานวิจัยนี้จึงเลือกแทนค่าดังกล่าวด้วยวิธีแทนค่าด้วยหลายค่า (Multiple Imputation; MI) เพื่อให้สามารถนำข้อมูลเข้าสู่กระบวนการวิเคราะห์ได้ นอกจากนี้ได้ทำการลบข้อมูลในคอลัมน์ Analysis No. ออกไปเนื่องจากไม่มีผลต่อการวิเคราะห์

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1088 entries, 0 to 1087
Data columns (total 18 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Analysis no.          1088 non-null   object
1   Origin                1088 non-null   object
2   Li                    1088 non-null   float64
3   Na                    1088 non-null   float64
4   Mg                    1088 non-null   float64
5   K                     1082 non-null   float64
6   Ca                    1068 non-null   float64
7   Sc                    1088 non-null   float64
8   V                     1088 non-null   float64
9   Cr                    1088 non-null   float64
10  Mn                    1088 non-null   float64
11  Fe                    1088 non-null   float64
12  Co                    617 non-null    float64
13  Ni                    1033 non-null   float64
14  Zn                    976 non-null    float64
15  Ga                    1088 non-null   float64
16  Rb                    1088 non-null   float64
17  Cs                    1088 non-null   float64
dtypes: float64(16), object(2)
memory usage: 153.1+ KB
```

รูปที่ 2. แสดงรายละเอียดข้อมูลที่ใช้ในงานวิจัยนี้

นอกจากนี้ได้ทำการสังเคราะห์ตัวอย่างเพิ่มในกลุ่มที่มีจำนวนตัวอย่างน้อย เฉพาะในชุดข้อมูลฝึกสอน (Training Set) จำนวนข้อมูลก่อนและหลังการปรับสมดุลด้วยวิธีการสังเคราะห์ข้อมูลเพิ่ม (SMOTE) ทำให้ตัวอย่างในแต่ละกลุ่มมีจำนวนเท่ากันที่ 314 ตัวอย่าง (ตารางที่ 1)

ตารางที่ 1. แสดงผลการปรับสมดุลจำนวนข้อมูลด้วยวิธี SMOTE

กลุ่ม	จำนวนตัวอย่าง ก่อนทำ SMOTE	จำนวนตัวอย่าง หลังทำ SMOTE
Afghanistan	93	314
Colombia	245	314
Zambia (Kafubu)	314	314
Zambia (Musakashi)	218	314

สำหรับผลลัพธ์การติดป้ายกำกับข้อมูลด้วยเทคนิค Label Encoding แสดงในตารางที่ 2

ตารางที่ 2. แสดงผลการติดป้ายกำกับให้กับกลุ่มตัวอย่าง

กลุ่มตัวอย่าง	ป้ายกำกับ
Afghanistan	0
Colombia	1
Zambia (Kafubu)	2
Zambia (Musakashi)	3

4.2 ผลการคัดเลือกคุณลักษณะ

ทำการคัดเลือกคุณลักษณะสำหรับสร้างแบบจำลองด้วยเทคนิค One-Way ANOVA พบว่าค่าสถิติ F (F-Score) ของตัวแปรต้นอยู่ระหว่าง 32.665 ถึง 2270.786 (ตารางที่ 3) ในงานวิจัยนี้ทดลองคัดเลือกคุณลักษณะ 11 อันดับแรกที่มีค่าสถิติ F สูงกว่า 300 ซึ่งแสดงค่าเฉลี่ยของกลุ่มแตกต่างกันสูง มีแนวโน้มเพิ่มประสิทธิภาพในการจำแนกกลุ่มได้ดี ได้แก่ ลิเทียม (Li) โซเดียม (Na) เหล็ก (Fe) แมกนีเซียม (Mg) ซีเซียม (Cs) แมงกานีส (Mn) รูบิเดียม (Rb) นิกเกิล (Ni) วานาเดียม (V) สังกะสี (Zn) และแคลเซียม (Ca) โดยมีค่า F-score อยู่ระหว่าง 305.690 ถึง 2270.786 นอกจากนี้ตัวแปรดังกล่าวยังแสดงค่า p-value น้อยกว่า 0.05 ซึ่งแสดงให้เห็นว่ามีความสามารถในการจำแนกกลุ่มได้อย่างมีนัยสำคัญทางสถิติ

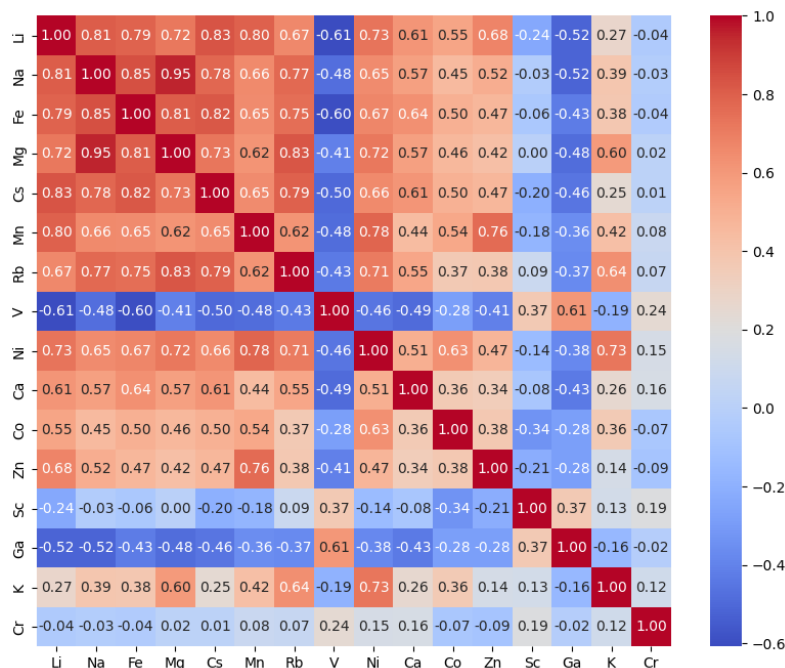
ตารางที่ 3. แสดงค่าสถิติ F และระดับนัยสำคัญของธาตุแต่ละชนิดจากการวิเคราะห์ด้วยเทคนิค One-Way ANOVA

ตัวแปรต้น	คำอธิบาย	F-score	p-value
Li	ลิเทียม	2270.786	< 0.001
Na	โซเดียม	1762.535	< 0.001
Fe	เหล็ก	1260.116	< 0.001
Mg	แมกนีเซียม	797.837	< 0.001
Cs	ซีเซียม	700.895	< 0.001
Mn	แมงกานีส	484.035	< 0.001
Rb	รูบิเดียม	383.838	< 0.001
Ni	นิกเกิล	339.728	< 0.001
V	วานาเดียม	336.771	< 0.001
Zn	สังกะสี	309.685	< 0.001
Ca	แคลเซียม	305.690	< 0.001
Co	โคบอลต์	223.385	< 0.001
Sc	สแกนเดียม	206.622	< 0.001
Ga	แกลเลียม	163.878	< 0.001

K	โพแทสเซียม	56.902	< 0.001
Cr	โครเมียม	32.665	< 0.001

ในการประเมินความสัมพันธ์ระหว่างตัวแปรและลดความซ้ำซ้อนของข้อมูล ผู้วิจัยได้ทำการวิเคราะห์ค่าสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน (Pearson's correlation coefficient: r) ซึ่งค่าที่ได้จะอยู่ในช่วงระหว่าง -1 ถึง 1 จากผลการวิเคราะห์พบว่า โซเดียม (Na) และแมกนีเซียม (Mg) มี

ความสัมพันธ์เชิงบวกในระดับสูงมาก โดยมีค่า r เท่ากับ 0.95 (รูปที่ 3) กรณีนี้อาจก่อให้เกิดปัญหา Multicollinearity ในกระบวนการสร้างแบบจำลองเพื่อหลีกเลี่ยงปัญหาดังกล่าว จึงได้พิจารณาค่าสถิติ F ประกอบการตัดสินใจ พบว่า Na มีค่าสถิติ F สูงกว่า Mg จึงทำการตัดตัวแปร Mg ออกจากชุดข้อมูล ดังนั้นคุณลักษณะที่เหมาะสมสำหรับใช้เป็นข้อมูลนำเข้าในการพัฒนาแบบจำลองในงานวิจัยนี้ ประกอบด้วย 10 ธาตุ ได้แก่ Li, Na, Fe, Co, Cs, Mn, Rb, Ni, V และ Ca



รูปที่ 3. แผนภาพความร้อนแสดงค่าสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน (Pearson Correlation Heatmap) ของคุณลักษณะในชุดข้อมูลการวิจัย

4.3 ผลการวัดประสิทธิภาพแบบจำลอง

แบบจำลอง	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
k-NN	95.87	95.94	95.87	95.87
NB	99.08	99.16	99.08	99.10
SVM	86.70	86.82	86.70	86.73
LR	86.24	86.43	86.24	86.24
LDA	84.86	85.60	84.86	84.98
RF	99.54	99.55	99.54	99.54

การวัดประสิทธิภาพของแบบจำลองทั้ง 6 ประเภท ด้วยเครื่องมือ Confusion matrix แสดงในตารางที่ 4 การวัดผลจะอ้างอิงค่าพารามิเตอร์ที่เหมาะสมที่สุด โดยแสดงผลการประเมินด้วยค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าความถ่วงดุล (F-1 Score) โดยวิธี Weighted Averaging

ตารางที่ 4. ผลการวัดประสิทธิภาพของแบบจำลองทั้ง 6 อัลกอริทึม

จากผลการศึกษาพบว่าแบบจำลองป่าสุ่ม (Random Forest) ให้ประสิทธิภาพโดยรวมสูงสุด สามารถจำแนกแหล่งกำเนิดมลพิษได้อย่างถูกต้อง โดยมีค่าความถูกต้อง ร้อยละ 99.54 ค่าความแม่นยำ ร้อยละ 99.55 ค่าความระลึก ร้อยละ 99.54 และค่าความถ่วงดุล ร้อยละ 99.54 รองลงมา คือแบบจำลองเพื่อนบ้านใกล้ที่สุด-เค (k-Nearest Neighbor) ให้ค่าความค่าความถูกต้อง ร้อยละ 95.87 ค่าความแม่นยำ ร้อยละ 95.94 ค่าความระลึก ร้อยละ 95.87 และค่าความถ่วงดุล ร้อยละ 95.87 ในขณะที่แบบจำลองที่มีประสิทธิภาพต่ำสุดในการศึกษานี้ คือ แบบจำลองการวิเคราะห์จำแนกประเภทเชิงเส้น (Linear Discriminant Analysis) ซึ่งให้ค่าความค่าความถูกต้อง ร้อยละ 84.86 ค่าความแม่นยำ ร้อยละ 85.60 ค่าความระลึก ร้อยละ 84.86 และค่าความถ่วงดุล ร้อยละ 84.98

นอกจากนี้ เพื่อยืนยันประสิทธิภาพของแบบจำลองป่าสุ่ม ได้มีการตรวจสอบด้วยวิธี 5-Fold Cross Validation พบว่าค่าความถูกต้องในแต่ละรอบดังแสดงในตารางที่ 5 โดยมีค่าเฉลี่ยความถูกต้องเท่ากับร้อยละ 99.72 แสดงให้เห็นว่าแบบจำลองมีความสม่ำเสมอในการเรียนรู้และสามารถทำการจำแนกข้อมูลที่ไม่เคยเห็นมาก่อนได้อย่างมีประสิทธิภาพ

ตารางที่ 5. ผลการวัดประสิทธิภาพของแบบจำลองป่าสุ่มด้วยเทคนิค 5-Fold Cross Validation

รอบที่	ค่าความถูกต้อง
1	100%
2	100%
3	100%
4	98.62%
5	100%
ค่าเฉลี่ยความถูกต้อง	99.72
ค่าเบี่ยงเบนมาตรฐาน (S.D.)	0.55

5. บทสรุปและการอภิปราย

งานวิจัยนี้มุ่งเน้นไปที่การพัฒนาแบบจำลองการเรียนรู้ของเครื่องเพื่อจำแนกแหล่งกำเนิดของมลพิษจำนวน 4 แหล่งที่สำคัญ โดยผ่านกระบวนการเลือกคุณลักษณะเฉพาะตัวแปรที่สามารถจำแนกกลุ่มได้ดีในการสร้างแบบจำลองเรียนรู้ และส่งผลให้แบบจำลองการเรียนรู้มีประสิทธิภาพสูงในการจำแนก จากการวิเคราะห์ด้วยเทคนิค One-Way Anova และสัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน พบว่าคุณลักษณะที่สามารถจำแนกกลุ่มได้ดีมีอยู่จำนวน 10 ธาตุ ได้แก่ Li, Na, Fe, Co, Cs, Mn, Rb, Ni, V และ Ca ซึ่งสอดคล้องกับงานวิจัยของ Alonso-Perez et al. [25] ซึ่งทำการคัดเลือกธาตุร่องรอยจำนวน 8 ธาตุ ได้แก่ Li, V, Cr, Fe, Sc, Ga, Rb และ Cs นำมาสร้างแบบจำลองการเรียนรู้ของเครื่องเพื่อจำแนกมลพิษจากแหล่งต่าง ๆ ได้อย่างมีประสิทธิภาพ นอกจากนี้ยังสอดคล้องกับการศึกษาของ มณฑิรา เสนิงศ์ ณ อุทยา และคณะ [21] คัดเลือกธาตุร่องรอย จำนวน 11 ธาตุ ได้แก่ Li, Na, Mg, K, Sc, V, Cr, Fe, Ga, Rb และ Cs ในการสร้างแบบจำลองการเรียนรู้ของเครื่องที่สามารถจำแนกแหล่งกำเนิดของมลพิษได้อย่างมีประสิทธิภาพเช่นเดียวกัน จากงานวิจัยนี้สามารถสรุปได้ว่าคุณลักษณะจำนวน 5 ธาตุ ได้แก่ Li, V, Fe, Rb และ Cs เป็นกลุ่มธาตุร่องรอยที่มีประสิทธิภาพสูงในการจำแนกแหล่งกำเนิดของมลพิษ ซึ่งสอดคล้องกับงานวิจัยนี้โดยพบว่าธาตุดังกล่าวมีค่าสถิติ F อยู่ใน 10 อันดับแรก ทั้งนี้ Alonso-Perez et al. [25] อธิบายเพิ่มเติมว่ามลพิษจากแหล่งโคลอมเบีย มีความเข้มข้นของธาตุรูบิเดียมในระดับต่ำเพียง 0.2–3.3 ppmw เมื่อเทียบกับแหล่งอื่น ๆ เช่น แหล่งมาดากัสการ์ (23–114 ppmw) และแหล่งแซมเบีย (16–138 ppmw) ดังนั้นธาตุรูบิเดียมจึงเป็นตัวแปรที่สำคัญในการจำแนกมลพิษจากโคลอมเบียออกจากแหล่งอื่น เช่นเดียวกับการศึกษาของ Saeseaw et al. [3] ที่สนับสนุนว่ามลพิษจากโคลอมเบียประกอบด้วยธาตุอัลคาไลที่มีปริมาณความเข้มข้นต่ำกว่าแหล่งอื่น โดยเฉพาะธาตุ Li, K, Rb, Cs

และ Fe ในขณะที่มรดกจากแหล่งกาฟูบูในประเทศแซมเบียมีปริมาณธาตุ Cs, Fe และ Li ก่อนข้างสูงกว่าแหล่งอื่น ๆ

การวิจัยนี้แสดงให้เห็นว่าแบบจำลองป่าสุ่ม (Random Forest) ให้ประสิทธิภาพสูงสุดในการจำแนกกำเนิดของมรดก โดยให้ค่าความถูกต้องที่ร้อยละ 99.54 รองลงมาคือแบบจำลองนาอิวเบย์ ซึ่งยังถือว่ามีประสิทธิภาพสูงเช่นกัน โดยมีค่าความถูกต้องร้อยละ 99.08 ซึ่งผลการวิจัยดังกล่าวนี้สอดคล้องกับการศึกษาของ มณฑิรา เสนิงค์ ณ อุรุยา และคณะ [26] ซึ่งได้ศึกษาและพัฒนาแบบจำลองการเรียนรู้ของเครื่องในการจำแนกแหล่งกำเนิดของมรดกจาก 5 แหล่ง ได้แก่ ประเทศโคลอมเบีย แซมเบีย เอธิโอเปีย ปากีสถาน และอัฟกานิสถาน ซึ่งผลการวิจัยพบว่าแบบจำลองป่าสุ่ม และโครงข่ายประสาทเทียม (Artificial Neural Network) โดยให้ค่าความถูกต้องสูงสุดเท่ากันที่ร้อยละ 97 ในขณะที่แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนให้ประสิทธิภาพต่ำสุดที่ร้อยละ 86.70 นอกจากนี้ในงานวิจัยนี้ ยังสอดคล้องกับการศึกษาของ Alonso-Perez et al. [25] ซึ่งทำการวัดประสิทธิภาพของแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนเพื่อจำแนกมรดก 13 แหล่ง ซึ่งพบว่าแบบจำลองให้ค่าความถูกต้องเพียงร้อยละ 79 โดยสาเหตุเกิดจากความคล้ายคลึงกันของข้อมูลระหว่างคลาส (Inter-Class Similarities) ซึ่งส่งผลให้เส้นแบ่งระหว่างกลุ่มมีความซับซ้อนและยากต่อการแยกอย่างชัดเจน

จากผลการวิจัยนี้สามารถนำไปแก้ปัญหาที่เกิดขึ้นบ่อยครั้งในการจำแนกแหล่งกำเนิดของมรดก ซึ่งบ่อยครั้งนักอณูเคมีจะใช้วิธีการพล็อตกราฟเปรียบเทียบระหว่างธาตุสองธาตุ (Bivariate Plot) เช่น การศึกษาของ Krzemnicki et al. [15] ทำการพล็อตกราฟเปรียบเทียบระหว่างธาตุรูบิเดียม (Rb) และซีเซียม (Cs) เพื่อจำแนกมรดกจากประเทศโคลอมเบีย และแซมเบีย (มุซาคาซิ) แต่กลับพบว่าข้อมูลเคมีของทั้งสองแหล่งยังคงทับซ้อนกันอยู่ เช่นเดียวกับการศึกษาของ Saeseaw et al. [3] พบว่าการพล็อตกราฟเทียบกันระหว่าง

โพแทสเซียม (K) และเหล็ก (Fe) ไม่สามารถแยกแหล่งมรดกจากประเทศโคลอมเบีย จีน และแซมเบีย (มุซาคาซิ) ออกจากกันได้ชัดเจน ทั้งนี้จำเป็นต้องใช้ข้อมูลทางสเปกโทรสโกปีและลักษณะเฉพาะของมลทินภายในประกอบการตัดสินใจเพิ่มเติม

6. ข้อเสนอแนะสำหรับการวิจัยในอนาคต

ในการวิจัยครั้งนี้ แม้ว่าแบบจำลองป่าสุ่มจะให้ค่าความถูกต้องสูงถึงร้อยละ 99.54 แต่เมื่อพิจารณาแล้ว อาจเกิดจากจำนวนข้อมูลที่มีจำกัด ซึ่งส่งผลให้แบบจำลองเรียนรู้ข้อมูลที่น้อยเกินไปจนทำให้ง่ายต่อการจำแนก ดังนั้นแนวทางการพัฒนาในอนาคตควรปรับปรุงข้อมูลให้มีความซับซ้อนและความท้าทายมากยิ่งขึ้น ดังนี้

- 1) เพิ่มปริมาณข้อมูลสำหรับชุดข้อมูลฝึกสอน และชุดข้อมูลทดสอบ รวมถึงการจัดการจำนวนตัวอย่างในแต่ละคลาสให้มีความใกล้เคียงกัน เพื่อพัฒนาแบบจำลองการเรียนรู้ของเครื่องให้มีความน่าเชื่อถือมากยิ่งขึ้น
- 2) เพิ่มความหลากหลายของข้อมูลโดยรวบรวมมรดกจากแหล่งอื่น ๆ ซึ่งในปัจจุบันพบว่ามรดกอีกหลายแหล่งมากที่มีจำหน่ายในตลาดอณูเคมี เช่น บราซิล มาดากัสการ์ เอธิโอเปีย รัสเซีย และปากีสถาน เพื่อเพิ่มความถูกต้องและแม่นยำของแบบจำลองในกรณีที่เจอมรดกจากแหล่งอื่น ๆ
- 3) บูรณาการร่วมกับข้อมูลทางอณูเคมีวิทยาเพื่อเพิ่มขีดความสามารถให้กับแบบจำลอง เช่น ลักษณะเด่นของมลทินภายใน ตำแหน่งของสเปกตรัมที่ได้จากเทคนิค UV-Vis-NIR / Raman / Photoluminescence Spectroscopy เป็นต้น
- 4) ประยุกต์ใช้เทคนิคการเรียนรู้ของเครื่องที่หลากหลายมากยิ่งขึ้น เช่น การเรียนรู้เชิงลึก (Deep Learning) หรือการเรียนรู้แบบผสม (Ensemble Learning) เพื่อเปรียบเทียบประสิทธิภาพของแบบจำลองการเรียนรู้ของเครื่องขั้นสูงต่อไป

5) พัฒนาแบบจำลองการเรียนรู้ของเครื่องสำหรับจำแนกแหล่งกำเนิดของอัญมณีชนิดอื่น ๆ ที่มีความสำคัญในเชิงพาณิชย์ เช่น ทับทิม ไพลิน สปิเนล และทัวร์มาลีน เป็นต้น

เอกสารอ้างอิง

[1] S. F. McClure, T. M. Moses, and J. E. Shigley, "The geographic origin dilemma," *Gems & Gemology*, vol. 55, no. 4, pp. 450–453, 2019.

[2] J.-L. Archuleta, "The color of responsibility: Ethical issues and solutions in colored gemstones," *Gems & Gemology*, vol. 52, no. 2, pp. 114–131, 2016.

[3] S. Saeseaw, N. D. Renfro, A. C. Palke, Z. Sun, and S. F. McClure, "Geographic origin determination of emerald," *Gems & Gemology*, vol. 55, no. 4, pp. 478–507, 2019.

[4] D. L. Wood and K. Nassau, "The characterization of beryl and emerald by visible and infrared absorption spectroscopy," *American Mineralogist*, vol. 53, pp. 777–800, 1968.

[5] E. Fritsch and G. R. Rossman, "An update on color in gems. Part I: Introduction and colors caused by dispersed metal ions," *Gems & Gemology*, vol. 23, pp. 126–139, 1987.

[6] D. Schwarz, G. Giuliani, G. Grundmann, and M. Glas, "The origin of emerald...a controversial topic," *extraLapis English*, no. 2, pp. 18–21, 2002.

[7] M. Awad and R. Khanna, *Machine Learning in Action: Examples*. New York, NY, USA: Apress, 2015.

[8] T. Blodgett and A. H. Shen, "Application of discriminant analysis in gemology: Country-of-origin separation in colored stones and distinguishing HPHT-treated diamonds," *Gems & Gemology*, vol. 47, no. 2, p. 145, 2011.

[9] Z. Luo, M. Yang, and A. H. Shen, "Origin determination of dolomite-related white nephrite through IB-LDA," *Gems & Gemology*, vol. 51, no. 3, pp. 268–278, 2015.

[10] A. Homkrajac, Z. Sun, and S. C. Shih, "Gemological and chemical characteristics of natural freshwater pearls from the

Mississippi River system," *Gems & Gemology*, vol. 55, no. 2, pp. 164–185, 2019.

[11] H. A. O. Wang and M. S. Krzemnicki, "Multi-element analysis of minerals using laser ablation inductively coupled plasma time of flight mass spectrometry and geochemical data visualization using t-distributed stochastic neighbor embedding: Case study on emeralds," *Journal of Analytical Atomic Spectrometry*, vol. 36, no. 3, pp. 518–527, 2021.

[12] Y. Zhong, A. H. Shen, Z. Zhang, M. Ye, and Y. Han, "Application of machine learning algorithms in the geographical origin determination of peridot," *Journal of Gems & Gemology*, vol. 25, no. 6, pp. 65–75, 2023 (in Chinese).

[13] M. Seneewong-Na-Ayutthaya, B. Wanthanachaisaeng, W. Suwanmanee, T. Lhuamporn, P. Kamkeaw, S. Phibanchon, T. Sripoonjan, and T. Leelawatanasuk, "An implementation of machine learning in ruby and sapphire origin determination," in *Proc. 37th International Gemmological Conference (IGC 2023)*, Tokyo, Japan, 2023.

[14] G. Mariscal, Ó. Marbán, and C. Fernández, "A survey of data mining and knowledge discovery process models and methodologies," *The Knowledge Engineering Review*, vol. 25, no. 2, pp. 137–166, 2010.

[15] M. S. Krzemnicki, H. A. O. Wang, M. Wälle, P. Lefèvre, W. Zhou, and L. E. Cartier, "Gemmological characterisation of emeralds from Musakashi, Zambia, and implications for their geographic origin determination," *Journal of Gemmology*, vol. 39, no. 4, pp. 338–350, 2024.

[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[17] A. Tharwat, T. Gaber, A. Ibrahim, and A. E. Hassanien, "Linear discriminant analysis: A detailed tutorial," *AI Communications*, vol. 30, no. 2, pp. 169–190, 2017.

- [18] C. Y. J. Peng, K. L. Lee, and G. M. Ingersoll, "An introduction to logistic regression analysis and reporting," *Journal of Educational Research*, vol. 96, no. 1, pp. 3–14, 2002.
- [19] P. Domingos and M. Pazzani, "On the optimality of the simple Bayesian classifier under zero-one loss," *Machine Learning*, vol. 29, no. 2–3, pp. 103–130, 1997.
- [20] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.
- [21] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832–844, 1998.
- [22] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Schölkopf, "Support vector machines," *IEEE Intelligent Systems*, vol. 13, no. 4, pp. 18–28, 1998.
- [23] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. New York, NY, USA: Springer, 2013.
- [24] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.
- [25] F. Alonso-Pérez, L. T. Grisi, F. G. Sardi, A. Martínez, A. L. de Oliveira, L. Lobo-Guerrero, G. Barone, and S. Raneri, "Geographic origin determination of emeralds using machine learning algorithms and handheld LIBS data," *Minerals*, vol. 14, no. 1, Art. no. 89, 2024.
- [26] มณฑิรา เสนีวงศ์ ณ อยุธยา, พะนง ลีลาวัฒนสุข, ภูวดล วรธนะชัยแสง, วรดิษฐ์ชนก สุวรรณมณี, ชีวากร สุพรรณ, พิมพ์ภัทส คำแก้ว, ทศนรา ศรีปิ่นจั่น, และ อนุรักษ บุญมาวงศ์, "กระบวนการพิสูจน์แหล่งกำเนิดอัญมณีโดยใช้เอกลักษณ์ทางเคมีด้วยเทคนิค LA-ICP-MS (เบรล)," ใน *การสัมมนาเผยแพร่งานวิจัยและนวัตกรรมด้านอัญมณีและ*

เครื่องประดับ: วรรณการความร่วมมือหน่วยงานเครือข่ายสู่
การนำไปใช้ประโยชน์, กรุงเทพฯ, ประเทศไทย, 2567.

การประยุกต์ใช้เทคนิคเหมืองข้อมูลเพื่อพัฒนาระบบสารสนเทศ

พยากรณ์การใช้น้ำประปา

Application of Data Mining Techniques for Developing an Information System for Water Demand Forecasting

หฤทัย อาษากิจ^{1*} และ พัชรมัย ชาวสืบ²

Haruethai Asakit^{1*} and Patcharamai Saosueb²

^{1,2}คณะบริหารธุรกิจและศิลปศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา เชียงใหม่

^{1,2}Faculty of Business Administration and Liberal Arts, Rajamangala University of Technology Lanna,

Chiang Mai

aeharuethai@gmail.com

Received: February 3, 2026; Revised: June 26, 2026; Accepted: June 27, 2026; Published: June 30, 2026

ABSTRACT – This study compared the performance of water consumption forecasting models in Thailand using data mining techniques based on the Cross-Industry Standard Process for Data Mining (CRISP-DM). Four models were evaluated: Linear Regression (LR), Artificial Neural Network (ANN), Random Forest (RF), and Gradient Boosted Trees (GBT). Monthly water consumption data published by the Metropolitan Waterworks Authority (MWA) through Thailand's Government Open Data Center, covering 2014–2023, were divided into training, validation, and testing sets (80:10:10). The models were developed using RapidMiner and evaluated using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). The results showed that the Linear Regression model achieved the best performance, with an MAE of 188,158.96 m³ and an RMSE of 310,809.74 m³, outperforming the other models while demonstrating high prediction stability. A web-based prototype decision support system was also developed to visualize forecasting results through an interactive dashboard, supporting water production planning and water distribution management in alignment with Sustainable Development Goal 6 (SDG 6).

KEY WORDS: Water demand forecasting, Data mining, Information system, Decision support system

บทคัดย่อ -- งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของแบบจำลองการพยากรณ์ปริมาณการใช้น้ำประปาในประเทศไทย โดยประยุกต์ใช้เทคนิคเหมืองข้อมูลร่วมกับกระบวนการ CRISP-DM แบบจำลองที่ศึกษา ได้แก่ การถดถอยเชิงเส้น โครงข่ายประสาทเทียม แรนดอมฟอเรสต์ และต้นไม้เสริมกำลังแบบไล่ระดับ ข้อมูลปริมาณการใช้น้ำประปารายเดือนที่เผยแพร่โดยการประปานครหลวงผ่านศูนย์กลางข้อมูลเปิดภาครัฐ ช่วงปี พ.ศ.2557–2566 และแบ่งเป็นชุดฝึกสอน ชุดตรวจสอบ และชุดทดสอบในสัดส่วน 80:10:10 โดยใช้โปรแกรม RapidMiner ในการพัฒนาแบบจำลอง การประเมินผลใช้ค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ยและค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง ผลการวิจัยพบว่าแบบจำลองการถดถอยเชิงเส้นให้ผลดีที่สุด โดยมีค่า MAE เท่ากับ 188,158.96 ลูกบาศก์เมตร และค่า RMSE เท่ากับ 310,809.74 ลูกบาศก์เมตร ซึ่งต่ำกว่าแบบจำลองอื่น และมีความเสถียรสูง นอกจากนี้ยังได้พัฒนาระบบต้นแบบบนแพลตฟอร์มเว็บเพื่อแสดงผลการพยากรณ์ในรูปแบบแดชบอร์ด สนับสนุนการวางแผนการผลิตและการกระจายน้ำ และสอดคล้องกับเป้าหมายการพัฒนาที่ยั่งยืนด้านทรัพยากรน้ำ (SDG 6)

คำสำคัญ: การพยากรณ์การใช้น้ำประปา เหมืองข้อมูล ระบบสารสนเทศ ระบบสนับสนุนการตัดสินใจ

1. บทนำ

น้ำเป็นทรัพยากรธรรมชาติที่มีบทบาทสำคัญต่อการดำรงชีวิตของมนุษย์ พืช และสัตว์ รวมถึงกิจกรรมทางเศรษฐกิจที่สำคัญ เช่น ภาคการเกษตร อุตสาหกรรม และชุมชนเมือง แม้ว่าน้ำจะเป็นทรัพยากรที่สามารถหมุนเวียนได้ตามวัฏจักรธรรมชาติ แต่การเปลี่ยนแปลงสภาพภูมิอากาศ การเพิ่มขึ้นของประชากร และการขยายตัวของเมืองได้ก่อให้เกิดความต้องการใช้น้ำที่เพิ่มสูงขึ้นอย่างต่อเนื่อง โดยเฉพาะในภาคครัวเรือนและอุตสาหกรรม ส่งผลให้หลายพื้นที่ทั่วโลกเผชิญกับความเสี่ยงด้านความมั่นคงทางน้ำ [1] สถานการณ์ดังกล่าวสะท้อนให้เห็นถึงความจำเป็นในการพัฒนาระบบบริหารจัดการน้ำที่มีประสิทธิภาพและยั่งยืนในยุคปัจจุบัน รายงานด้านการประเมินผลกระทบสิ่งแวดล้อมระบุว่า ความผันผวนของอุณหภูมิและปริมาณน้ำฝนมีความสัมพันธ์โดยตรงกับรูปแบบการใช้น้ำของชุมชนเมืองและภาคครัวเรือน [2] ซึ่งสอดคล้องกับรายงานของการประปานครหลวง (2564) ที่พบว่าอุณหภูมิเฉลี่ยรายเดือนมีความสัมพันธ์กับความต้องการใช้น้ำประปาอย่างมีนัยสำคัญ โดยอุณหภูมิที่เพิ่มขึ้น 1 องศาเซลเซียส ส่งผลให้ความต้องการใช้น้ำเพิ่มขึ้นเฉลี่ยประมาณร้อยละ 1.3 ต่อปี นอกจากนี้ รายงานสถานการณ์ทรัพยากรน้ำของประเทศไทยยังชี้ให้เห็นว่าแนวโน้มอุณหภูมิที่สูงขึ้นอย่างต่อเนื่องส่งผลให้ความต้องการใช้น้ำประปาในเขตเมืองเพิ่มสูงขึ้นอย่างมีนัยสำคัญ [3] ทำให้การวางแผนระบบผลิตและกระจายน้ำในอนาคตจำเป็นต้องอาศัยข้อมูลเชิงพยากรณ์ที่มีความแม่นยำและเชื่อถือได้

ในบริบทของประเทศไทย การบริหารจัดการทรัพยากรน้ำยังเผชิญกับความท้าทายทั้งด้านนโยบายและเทคนิค โดยเฉพาะการพัฒนาระบบประปาที่ยั่งยืนและการรองรับการขยายตัวของเมือง [4] ภายใต้แผนนโยบาย Thailand 4.0 และแนวคิดสมรรถนะประเทศไทย ได้เน้นย้ำถึงความสำคัญของการประยุกต์ใช้เทคโนโลยีดิจิทัลและระบบพยากรณ์แบบเรียลไทม์เพื่อสนับสนุนการตัดสินใจเชิงบริหาร [5] ขณะเดียวกัน ภาครัฐยังให้ความสำคัญต่อการบูรณาการข้อมูลและเทคโนโลยีเพื่อยกระดับประสิทธิภาพการบริหารจัดการน้ำในระดับประเทศ [6]

ปัจจุบัน เทคโนโลยีการวิเคราะห์ข้อมูลขนาดใหญ่ เทคนิคเหมืองข้อมูล และการเรียนรู้ของเครื่อง ถูกนำมาใช้ในการพัฒนาระบบพยากรณ์ความต้องการใช้น้ำมากขึ้น เนื่องจากสามารถรองรับข้อมูลจำนวนมากและสร้างแบบจำลองที่มีความแม่นยำสูง รวมทั้งสามารถพิจารณาปัจจัยด้านสภาพอากาศ ปริมาณฝน ความชื้น และจำนวนประชากรควบคู่กันได้ [7] อย่างไรก็ตาม การเลือกแบบจำลองที่เหมาะสมยังคงเป็นประเด็นสำคัญ เนื่องจากแบบจำลองแต่ละประเภทมีข้อจำกัดและประสิทธิภาพที่แตกต่างกัน

จากประเด็นดังกล่าว ผู้วิจัยจึงมีแนวคิดในการประยุกต์ใช้เทคนิคเหมืองข้อมูลเพื่อเปรียบเทียบประสิทธิภาพของแบบจำลองการพยากรณ์การใช้น้ำประปา และคัดเลือกแบบจำลองที่เหมาะสมเพื่อนำไปพัฒนาเป็นระบบพยากรณ์เชิงสารสนเทศสำหรับสนับสนุนการวางแผนและการตัดสินใจด้านการจัดการทรัพยากรน้ำอย่างเป็นระบบและยั่งยืน โดยมุ่งเน้นการบูรณาการการเปรียบเทียบแบบจำลองเหมืองข้อมูลหลายประเภทเข้ากับการพัฒนาระบบสารสนเทศเชิงปฏิบัติการบนแพลตฟอร์มเว็บ ซึ่งยังพบว่ามีการศึกษาจำกัดในบริบทข้อมูลการใช้น้ำประปาระดับประเทศของประเทศไทยในลักษณะดังกล่าว

2. วัตถุประสงค์

2.1 เพื่อศึกษาและเปรียบเทียบประสิทธิภาพของแบบจำลองเหมืองข้อมูลสำหรับการพยากรณ์ปริมาณการใช้น้ำประปา

2.2 เพื่อพัฒนาและทดสอบแบบจำลองพยากรณ์ปริมาณการใช้น้ำประปาโดยใช้เทคนิคเหมืองข้อมูล

2.3 เพื่อพัฒนาระบบสารสนเทศเชิงสนับสนุนการตัดสินใจบนแพลตฟอร์มเว็บสำหรับการจัดการน้ำประปา

3. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

3.1 การทำเหมืองข้อมูลและกระบวนการ CRISP-DM

(Data Mining and CRISP-DM)

การทำเหมืองข้อมูล (Data Mining) เป็นกระบวนการค้นหาและสกัดองค์ความรู้ รูปแบบ และความสัมพันธ์ที่มีคุณค่าจากชุดข้อมูลขนาดใหญ่ โดยอาศัยองค์ความรู้จากหลากหลายศาสตร์ ได้แก่ สถิติ ปัญญาประดิษฐ์ และการเรียนรู้ของเครื่อง (Machine Learning) เพื่อวิเคราะห์ข้อมูลที่มีความซับซ้อนและ

สนับสนุนการตัดสินใจเชิงบริหาร [8] ในยุคดิจิทัลที่องค์กรและหน่วยงานภาครัฐมีการจัดเก็บข้อมูลจำนวนมากจากหลากหลายแหล่ง การทำเหมืองข้อมูลจึงมีบทบาทสำคัญในการแปลงข้อมูลให้เป็นสารสนเทศและองค์ความรู้ที่สามารถนำไปใช้ประโยชน์ได้จริง

การทำเหมืองข้อมูลถือเป็นส่วนหนึ่งของกระบวนการวิเคราะห์ข้อมูล (Data Analytics) ซึ่งเป็นกระบวนการตรวจสอบจัดการ และวิเคราะห์ข้อมูลเพื่อค้นหาสารสนเทศเชิงลึกที่ช่วยสนับสนุนการตัดสินใจและเพิ่มประสิทธิภาพการดำเนินงานขององค์กร [9] โดยการบูรณาการเทคนิคการวิเคราะห์ข้อมูลร่วมกับการทำเหมืองข้อมูลและปัญญาประดิษฐ์ ช่วยเพิ่มความสามารถในการค้นหารูปแบบที่ซับซ้อนของข้อมูล และสนับสนุนแนวคิดการบริหารจัดการองค์กรด้วยข้อมูล (Data-driven Decision Making) [10]

ในการดำเนินงานด้านการทำเหมืองข้อมูลและการวิเคราะห์ข้อมูล กรอบแนวคิด CRISP-DM (Cross-Industry Standard Process for Data Mining) ได้รับการยอมรับอย่างแพร่หลาย เนื่องจากเป็นกระบวนการมาตรฐานที่ช่วยให้การพัฒนาแบบจำลองเป็นระบบและสามารถนำผลลัพธ์ไปประยุกต์ใช้ได้จริง [11] โดย CRISP-DM ประกอบด้วย 6 ขั้นตอน ได้แก่ (1) การทำความเข้าใจปัญหาทางธุรกิจ (Business Understanding) (2) การทำความเข้าใจข้อมูล (Data Understanding) (3) การเตรียมข้อมูล (Data Preparation) (4) การสร้างแบบจำลอง (Modeling) (5) การประเมินผล (Evaluation) และ (6) การนำผลไปใช้งาน (Deployment)

กระบวนการดังกล่าวได้รับการนำไปประยุกต์ใช้อย่างแพร่หลายในหลากหลายสาขา เช่น การแพทย์ การเงิน พลังงาน สิ่งแวดล้อม และการบริหารจัดการทรัพยากรน้ำ โดยเฉพาะงานด้านการพยากรณ์ความต้องการใช้น้ำ ซึ่งต้องอาศัยการวิเคราะห์ข้อมูลเชิงเวลาและการสร้างแบบจำลองเพื่อสนับสนุนการวางแผนและการตัดสินใจเชิงบริหาร [12] ดังนั้น งานวิจัยนี้จึงประยุกต์ใช้กรอบแนวคิด CRISP-DM เป็นแนวทางในการดำเนินงานวิจัย ตั้งแต่การรวบรวมและเตรียมข้อมูล การพัฒนาแบบจำลองพยากรณ์ ไปจนถึงการประเมินประสิทธิภาพของแบบจำลอง เพื่อคัดเลือกแนวทางที่เหมาะสมสำหรับการพัฒนาระบบพยากรณ์ปริมาณการใช้น้ำประปา

3.2 การแบ่งข้อมูล (Data Splitting)

การแบ่งชุดข้อมูลออกเป็น 3 ส่วน ได้แก่ ชุดข้อมูลฝึกสอน (Training set) ร้อยละ 80 สำหรับการเรียนรู้แบบจำลอง ชุดข้อมูลตรวจสอบ (Validation set) ร้อยละ 10 สำหรับการปรับแต่งค่าพารามิเตอร์ของแบบจำลอง และชุดข้อมูลทดสอบ (Testing set) ร้อยละ 10 สำหรับการประเมินประสิทธิภาพขั้นสุดท้ายของแบบจำลอง ถือเป็นแนวปฏิบัติมาตรฐานในงานด้านการเรียนรู้ของเครื่อง เพื่อช่วยลดปัญหาการจดจำรูปแบบเฉพาะของข้อมูลฝึกสอน (overfitting) และเพิ่มความสามารถในการอธิบายทั่วไปของแบบจำลอง (generalization) การใช้ชุดข้อมูลตรวจสอบช่วยให้สามารถคัดเลือกแบบจำลองและค่าพารามิเตอร์ที่เหมาะสมก่อนนำไปประเมินผลด้วยชุดข้อมูลทดสอบจริง ส่งผลให้ผลการประเมินมีความน่าเชื่อถือและสะท้อนประสิทธิภาพของแบบจำลองในสถานการณ์ใช้งานจริงได้ดียิ่งขึ้น [13] [14]

3.3 การวัดประสิทธิภาพ (Evaluation)

การประเมินประสิทธิภาพของแบบจำลองเป็นขั้นตอนสำคัญในกระบวนการวิเคราะห์ข้อมูลและการพยากรณ์ เนื่องจากช่วยให้สามารถตรวจสอบความแม่นยำ ความเสถียร และความเหมาะสมของแบบจำลองก่อนนำไปประยุกต์ใช้งานจริง โดยทั่วไปนิยมใช้ตัวชี้วัดทางสถิติเพื่อประเมินความคลาดเคลื่อนระหว่างค่าที่พยากรณ์และค่าจริง ซึ่งตัวชี้วัดที่ใช้กันอย่างแพร่หลาย ได้แก่ ค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (Root Mean Square Error: RMSE) และค่าเฉลี่ยของค่าความคลาดเคลื่อนสัมบูรณ์ (Mean Absolute Error: MAE) ค่า RMSE เป็นตัวชี้วัดที่ให้น้ำหนักกับค่าความคลาดเคลื่อนขนาดใหญ่เป็นพิเศษ จึงเหมาะสำหรับกรณีที่ต้องการลดผลกระทบจากความผิดพลาดที่มีขนาดสูง ขณะที่ค่า MAE เป็นตัวชี้วัดที่สะท้อนค่าความคลาดเคลื่อนเฉลี่ยในรูปแบบสัมบูรณ์ ทำให้สามารถตีความผลลัพธ์ได้ง่ายและมีความเสถียรต่อค่าผิดปกติ (Outliers) มากกว่า [15] ดังนั้นการใช้ตัวชี้วัดทั้ง RMSE และ MAE ร่วมกันจึงช่วยให้สามารถประเมินประสิทธิภาพของแบบจำลองได้อย่างรอบด้าน และสนับสนุนการเลือกแบบจำลองที่เหมาะสมสำหรับการนำไปใช้งานด้านการพยากรณ์เชิงบริหารและการตัดสินใจเชิงข้อมูล

3.4 งานวิจัยที่เกี่ยวข้อง

A. Antunes et al. [16] ศึกษาการพยากรณ์ความต้องการใช้น้ำประปาในระยะสั้นในเขตเมืองของประเทศโปรตุเกส โดยใช้ข้อมูลอนุกรมเวลารายวันร่วมกับตัวแปรด้านสภาพอากาศ เช่น อุณหภูมิและปริมาณน้ำฝน และเปรียบเทียบแบบจำลอง Machine Learning ได้แก่ Random Forest, Gradient Boosting และ Support Vector Regression (SVR) การประเมินประสิทธิภาพใช้ตัวชี้วัด RMSE, MAE และ R^2 ผลการวิจัยพบว่า Gradient Boosting สามารถลดค่าความคลาดเคลื่อนได้มากที่สุด และให้ค่าความแม่นยำสูงสุดเมื่อเทียบกับโมเดลอื่น เนื่องจากสามารถจัดการความสัมพันธ์เชิงไม่เชิงเส้นของข้อมูลได้อย่างมีประสิทธิภาพ ผู้วิจัยเสนอว่าแบบจำลองดังกล่าวเหมาะสำหรับการนำไปพัฒนาเป็นระบบสนับสนุนการตัดสินใจขององค์กรประปา เพื่อช่วยในการวางแผนกำลังการผลิตและการจัดสรรทรัพยากรน้ำในระยะสั้น

W. Hao et al. [17] ศึกษาการเปรียบเทียบประสิทธิภาพของแบบจำลองการพยากรณ์ความต้องการใช้น้ำประปาในเมืองมิลาน ประเทศอิตาลี สำหรับช่วงระยะสั้น (1 วันล่วงหน้า) และระยะยาว (7 วันล่วงหน้า) โดยใช้ชุดข้อมูลรายวันจากระบบน้ำของเมือง งานวิจัยนี้ทดสอบทั้งแบบจำลองเชิงสถิติและวิธี Machine Learning ได้แก่ ARIMA, Artificial Neural Networks (ANN), Support Vector Regression (SVR), Light Gradient Boosting Machine (LightGBM) และ Long Short-Term Memory (LSTM) พร้อมพิจารณาการใช้กรอบ Wavelet Data-Driven Forecasting Framework (WDDFF) เพื่อเพิ่มประสิทธิภาพของโมเดล ผลการเปรียบเทียบชี้ให้เห็นว่าเทคนิค WDDFF ช่วยพัฒนาความสามารถในการพยากรณ์ของโมเดลทั้งหมด โดยเฉพาะเมื่อใช้ร่วมกับ LSTM ซึ่งให้ค่าความแม่นยำสูง ($R^2 > 0.9$) สำหรับทั้งการพยากรณ์ระยะสั้นและระยะยาว นอกจากนี้ LightGBM ยังแสดงศักยภาพในการลดจำนวนตัวแปรและเลือกคุณลักษณะสำคัญในบริบทของน้ำและทรัพยากรน้ำได้อย่างมีประสิทธิภาพ

N. Fumo et al. [18] ศึกษาการพยากรณ์ความต้องการใช้น้ำโดยใช้ชุดข้อมูลขนาดใหญ่จากระบบสาธารณูปโภคในสหรัฐอเมริกา และเปรียบเทียบแบบจำลอง Random Forest, XGBoost (ซึ่งอยู่ในกลุ่ม Gradient Boosted Trees) และ Linear Regression โดยใช้ตัวชี้วัด RMSE และ MAE เป็นเกณฑ์

ประเมินผล ผลการวิจัยพบว่า XGBoost ให้ประสิทธิภาพสูงสุดในการลดค่าความคลาดเคลื่อน เนื่องจากสามารถจัดการความสัมพันธ์เชิงซับซ้อนและข้อมูลที่มีความไม่เชิงเส้นได้ดี ขณะที่ Linear Regression ให้ผลลัพธ์ที่มีเสถียรภาพและสามารถตีความเชิงบริหารได้ง่าย ผู้วิจัยเสนอว่าการผสมผสานโมเดลขั้นสูงกับโมเดลเชิงสถิติสามารถช่วยเพิ่มความยืดหยุ่นในการพัฒนาเครื่องมือพยากรณ์สำหรับองค์กรด้านสาธารณูปโภค

Z. Pu et al. [19] นำเสนอโมเดล deep learning แบบผสมระหว่าง wavelet multi-resolution analysis, convolutional Artificial Neural Networks และ long short-term memory เพื่อพยากรณ์ปริมาณการใช้น้ำประปาในเขตเมืองระยะสั้น ผลการศึกษาแสดงให้เห็นว่าระบบ hybrid นี้สามารถให้ความแม่นยำในการพยากรณ์ได้ดีกว่าโมเดล ANN, LSTM, Conv1D และ GRUN แบบดั้งเดิม ซึ่งสอดคล้องกับแนวคิดการพัฒนาแบบพยากรณ์เชิงดิจิทัลที่มีความสามารถตอบสนองต่อรูปแบบข้อมูลที่ซับซ้อนได้อย่างมีประสิทธิภาพ

I. I. Aminu [7] ได้ทบทวนเชิงระบบของงานวิจัยด้านการพยากรณ์และการจำลองความต้องการใช้น้ำในช่วง 10 ปีที่ผ่านมา โดยมุ่งเน้นการประยุกต์ใช้แบบจำลองเชิงคำนวณอัจฉริยะ (Computational Intelligence Models) เช่น Artificial Neural Networks (ANN), Support Vector Machines (SVM), Random Forest, Gradient Boosting, Deep Learning และ Hybrid Models ผลการทบทวนพบว่าแบบจำลองปัญญาประดิษฐ์และการเรียนรู้ของเครื่องสามารถจัดการข้อมูลที่มีความไม่เชิงเส้นและความผันผวนได้ดีกว่าแบบจำลองเชิงสถิติแบบดั้งเดิม อย่างไรก็ตาม ยังไม่มีแบบจำลองใดที่เหมาะสมกับทุกบริบท เนื่องจากประสิทธิภาพขึ้นอยู่กับลักษณะข้อมูล ความละเอียดเชิงเวลา และปัจจัยด้านสภาพแวดล้อม โดยเสนอแนวทางการพัฒนาในอนาคต ได้แก่ การใช้ Hybrid Models การบูรณาการข้อมูลจาก Internet of Things (IoT) การประยุกต์ใช้ Big Data Analytics และการพัฒนาระบบพยากรณ์อัจฉริยะเพื่อสนับสนุนการตัดสินใจด้านการบริหารจัดการทรัพยากรน้ำอย่างยั่งยืน

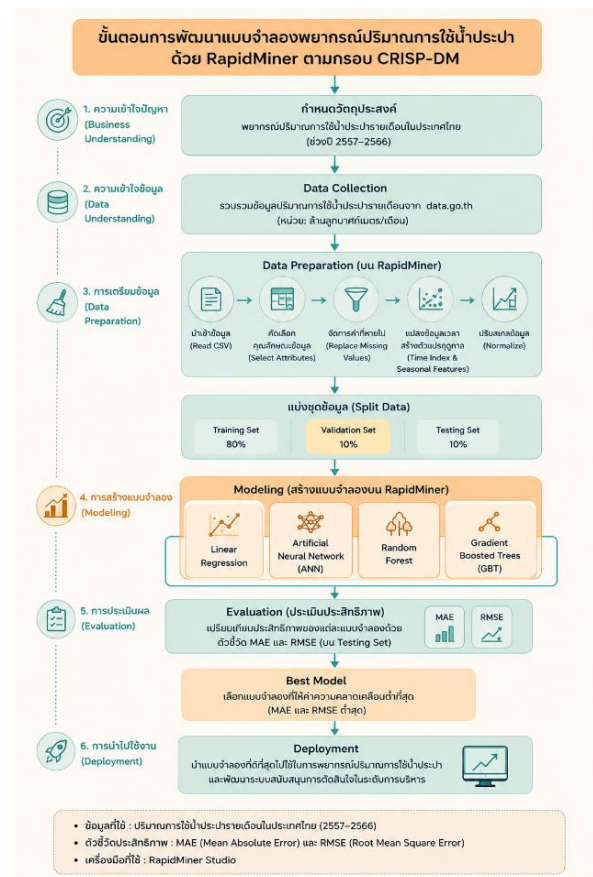
จากการทบทวนวรรณกรรมในช่วงปี ค.ศ. 2022–2025 พบว่า งานวิจัยด้านการพยากรณ์ปริมาณการใช้น้ำประปามีแนวโน้มประยุกต์ใช้เทคนิค Deep Learning และ Hybrid

Models มากขึ้น โดยเฉพาะ Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU) และการผสมเทคนิค Deep Learning ร่วมกับการประมวลผลสัญญาณหรือการคัดเลือกคุณลักษณะ ซึ่งสามารถเพิ่มความสามารถในการเรียนรู้รูปแบบข้อมูลอนุกรมเวลาและความสัมพันธ์ที่ซับซ้อนได้ อย่างไรก็ตาม แบบจำลองดังกล่าวมักมีความซับซ้อนในการพัฒนา ต้องการทรัพยากรด้านการคำนวณสูง และอาศัยข้อมูลจากหลายแหล่งร่วมกันเพื่อให้ได้ประสิทธิภาพสูงสุด

นอกจากนี้ งานวิจัยส่วนใหญ่ใช้ข้อมูลจากระบบประปาเฉพาะพื้นที่หรือข้อมูลที่มีตัวแปรสนับสนุนจำนวนมาก ขณะที่การเปรียบเทียบประสิทธิภาพของเทคนิคเหมือนข้อมูลภายใต้ข้อมูลเปิดภาครัฐที่สามารถเข้าถึงได้จริงยังมีจำนวนจำกัด ซึ่งสะท้อนให้เห็นถึงช่องว่างของงานวิจัย (Research Gap)

ดังนั้น งานวิจัยนี้จึงมุ่งเปรียบเทียบประสิทธิภาพของเทคนิคเหมือนข้อมูลที่ได้รับความนิยม ได้แก่ Artificial Neural Network (ANN), Linear Regression, Random Forest และ Gradient Boosted Trees บนแพลตฟอร์ม RapidMiner เพื่อประเมินความเหมาะสมในการพัฒนาระบบพยากรณ์ปริมาณการใช้น้ำประปาภายใต้บริบทของข้อมูลภาครัฐ พร้อมเสนอให้การประยุกต์ใช้แบบจำลอง Deep Learning เช่น LSTM, GRU และ Hybrid Models เป็นแนวทางสำหรับการศึกษาค้นคว้าในอนาคต

3.5 กรอบแนวคิดงานวิจัย



รูปที่ 1 กรอบแนวคิดและกระบวนการพัฒนาแบบจำลองพยากรณ์ปริมาณการใช้น้ำประปา ตามกรอบ CRISP-DM

จากรูปที่ 1 แสดงกรอบแนวคิดและกระบวนการดำเนินงานของงานวิจัยตามแนวทาง CRISP-DM โดยเริ่มจากการรวบรวมข้อมูลปริมาณการใช้น้ำประปารายเดือนจากศูนย์กลางข้อมูลเปิดภาครัฐ (data.go.th) จากนั้นดำเนินการเตรียมข้อมูล ประกอบด้วย การคัดเลือกรูปแบบ การจัดการข้อมูลที่ไม่สมบูรณ์ การสร้างตัวแปรเชิงเวลา และการปรับสเกลข้อมูลก่อนแบ่งชุดข้อมูลออกเป็น Training Set, Validation Set และ Testing Set เพื่อใช้ในการพัฒนาและประเมินแบบจำลอง

ในขั้นตอนการสร้างแบบจำลอง ผู้วิจัยได้เปรียบเทียบประสิทธิภาพของแบบจำลอง 4 ประเภท ได้แก่ Linear Regression, Artificial Neural Network (ANN), Random Forest และ Gradient Boosted Trees โดยใช้ค่า Mean Absolute Error (MAE) และ Root Mean Squared Error (RMSE) เป็นตัวชี้วัดประสิทธิภาพ ก่อนคัดเลือกแบบจำลองที่ให้ผลการพยากรณ์ดีที่สุดเพื่อนำไปประยุกต์ใช้ในการพัฒนาระบบพยากรณ์ปริมาณการใช้น้ำประปา

4. วิจัยดำเนินการวิจัย

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อศึกษาและเปรียบเทียบประสิทธิภาพของแบบจำลองพยากรณ์ปริมาณการใช้น้ำประปาโดยประยุกต์ใช้เทคนิคเหมืองข้อมูล เพื่อพัฒนาแบบจำลองที่มีความเหมาะสมสำหรับการนำไปใช้งานเชิงบริหาร งานวิจัยได้ดำเนินการตามกรอบกระบวนการ CRISP-DM (Cross-Industry Standard Process for Data Mining) ซึ่งเป็นมาตรฐานสากลด้านการวิเคราะห์ข้อมูล ประกอบด้วย 6 ขั้นตอนหลัก ได้แก่ การทำความเข้าใจปัญหาทางธุรกิจ การทำความเข้าใจข้อมูล การเตรียมข้อมูล การสร้างแบบจำลอง การประเมินผล และการนำไปใช้งาน ดังนี้

4.1 การทำความเข้าใจปัญหา (Business Understanding)

กระบวนการทำความเข้าใจปัญหาทางธุรกิจ (Business Understanding) เป็นขั้นตอนแรกของกรอบการวิเคราะห์ข้อมูลตามแนวคิด CRISP-DM ซึ่งมุ่งเน้นการกำหนดขอบเขตของปัญหา วัตถุประสงค์ขององค์กร และความต้องการเชิงบริหาร เพื่อแปลงโจทย์ทางธุรกิจให้อยู่ในรูปแบบของปัญหาการวิเคราะห์ข้อมูลที่สามารถดำเนินการได้อย่างเป็นระบบ ในการวิจัยครั้งนี้ ผู้วิจัยได้กำหนดปัญหาหลักเป็นการพยากรณ์ปริมาณการใช้น้ำประปาในเขตเมือง เพื่อสนับสนุนการวางแผนการผลิตและการกระจายน้ำของหน่วยงานที่เกี่ยวข้อง ข้อมูลที่นำมาใช้ในการศึกษาได้ถูกรวบรวมจากแหล่งข้อมูลภาครัฐ ได้แก่ ศูนย์กลางข้อมูลเปิดภาครัฐ (Open Government Data) และหน่วยงานด้านการประปา ซึ่งเป็นแหล่งข้อมูลที่มีความน่าเชื่อถือและสามารถนำมาใช้ในการวิเคราะห์เชิงนโยบายและเชิงบริหารได้ ผู้วิจัยได้วิเคราะห์ความต้องการของผู้มีส่วนได้ส่วนเสีย เช่น หน่วยงานด้านการบริหารจัดการน้ำ ผู้กำหนดนโยบาย และผู้ใช้น้ำ เพื่อกำหนดเป้าหมายของระบบพยากรณ์ให้สามารถตอบสนองต่อการตัดสินใจด้านการจัดสรรทรัพยากร การวางแผนกำลังการผลิต และการบริหารความเสี่ยงจากภาวะขาดแคลนน้ำในอนาคต ผลจากขั้นตอนการทำความเข้าใจปัญหาทางธุรกิจถูกนำไปใช้เป็นกรอบกำหนดตัวแปรนำเข้า แนวทางการเลือกแบบจำลองพยากรณ์ และเกณฑ์การประเมินประสิทธิภาพของแบบจำลอง เพื่อให้กระบวนการวิเคราะห์ข้อมูลมีความสอดคล้องกับวัตถุประสงค์ของงานวิจัย

และสามารถนำผลลัพธ์ไปใช้ประโยชน์เชิงปฏิบัติได้อย่างเป็นรูปธรรม

4.2 การทำความเข้าใจข้อมูล (Data Understanding)

ผู้วิจัยได้รวบรวมข้อมูลปริมาณการใช้น้ำประปาจากศูนย์กลางข้อมูลเปิดภาครัฐ (Open Government Data of Thailand: data.go.th) ซึ่งเป็นแพลตฟอร์มกลางสำหรับการเผยแพร่ข้อมูลของหน่วยงานภาครัฐในประเทศไทย โดยข้อมูลชุดนี้เผยแพร่โดยการประสานครหลวง (Metropolitan Waterworks Authority: MWA) ครอบคลุมช่วงระยะเวลาตั้งแต่ปี พ.ศ. 2557–2566 รวมระยะเวลา 10 ปี และมีจำนวนข้อมูลทั้งสิ้น 257,046 ระเบียบ

ข้อมูลที่ใช้ในการวิจัยเป็นข้อมูลปริมาณการใช้น้ำประปารายเดือน (Monthly Water Consumption Data) ซึ่งอยู่ในรูปแบบข้อมูลสรุปรายเดือน (Monthly Aggregated Data) ที่หน่วยงานเจ้าของข้อมูลได้ดำเนินการรวบรวมและสรุปผลเรียบร้อยแล้ว ก่อนเผยแพร่ผ่านระบบ data.go.th โดยแต่ละระเบียนแสดงปริมาณการใช้น้ำประปารวมของพื้นที่และประเภทผู้ใช้น้ำในเดือนนั้น ๆ มีหน่วยวัดเป็นลูกบาศก์เมตร (m³) ดังนั้น ผู้วิจัยจึงไม่จำเป็นต้องดำเนินการรวมข้อมูล (Data Aggregation) เพิ่มเติมก่อนการวิเคราะห์

ข้อมูลต้นฉบับประกอบด้วยตัวแปรจำนวน 11 ตัวแปร ได้แก่ (1) ปี (2) เดือน (3) รหัสจังหวัด (4) ชื่อจังหวัด (5) รหัสอำเภอ (6) ชื่ออำเภอ (7) รหัสตำบล (8) ชื่อตำบล (9) รหัสประเภทผู้ใช้น้ำ (10) ชื่อประเภทผู้ใช้น้ำ และ (11) ปริมาณการใช้น้ำประปา ซึ่งเป็นตัวแปรเป้าหมาย (Target Variable) สำหรับการพยากรณ์

ข้อมูลดังกล่าวจัดเก็บในระดับพื้นที่เชิงภูมิศาสตร์ ประกอบด้วยข้อมูลระดับจังหวัด อำเภอ และตำบล รวมถึงประเภทผู้ใช้น้ำในแต่ละพื้นที่ ซึ่งช่วยสะท้อนลักษณะการใช้น้ำประปาในมิติเชิงพื้นที่และเชิงเวลาได้อย่างครอบคลุม เหมาะสำหรับการนำมาพัฒนาแบบจำลองพยากรณ์ปริมาณการใช้น้ำประปา

ในการพัฒนาแบบจำลอง ผู้วิจัยได้คัดเลือกตัวแปรจำนวน 6 ตัวแปร ได้แก่ ปี เดือน รหัสจังหวัด รหัสอำเภอ รหัสประเภทผู้ใช้น้ำ และปริมาณการใช้น้ำประปา เนื่องจากเป็นตัวแปรที่สะท้อนลักษณะเชิงเวลาและเชิงพื้นที่ที่มีผลต่อความต้องการใช้

น้ำประปาโดยตรง ขณะที่ตัวแปรประเภทชื่อจังหวัด ชื่ออำเภอ ชื่อตำบล และชื่อประเภทผู้ใช้น้ำเป็นข้อมูลเชิงข้อความที่ซ้ำซ้อนกับตัวแปรรหัส จึงไม่นำมาใช้ในการสร้างแบบจำลอง เพื่อลดความซ้ำซ้อนของข้อมูล

4.3 การเตรียมข้อมูล (Data Preparation)

การเตรียมข้อมูลเป็นขั้นตอนสำคัญในกระบวนการวิเคราะห์ข้อมูลและการพัฒนาแบบจำลองพยากรณ์ โดยมีวัตถุประสงค์เพื่อปรับปรุงคุณภาพของข้อมูลให้อยู่ในรูปแบบที่ถูกต้อง ครบถ้วน และเหมาะสมต่อการนำไปประมวลผลเชิงคำนวณ ในงานวิจัยนี้ ผู้วิจัยได้ดำเนินการกระบวนการเตรียมข้อมูลตามขั้นตอนดังต่อไปนี้

1) การคัดเลือกข้อมูล (Data Selection) ผู้วิจัยได้ทำการคัดเลือกเฉพาะข้อมูลที่มีความเกี่ยวข้องโดยตรงกับวัตถุประสงค์ของการวิจัย เพื่อลดความซ้ำซ้อนของข้อมูลและเพิ่มประสิทธิภาพในการวิเคราะห์ โดยเลือกใช้ข้อมูลปริมาณการใช้น้ำประปารายเดือนในช่วงปี พ.ศ. 2557–2566 ซึ่งรวบรวมจากศูนย์กลางข้อมูลเปิดภาครัฐ (Open Government Data of Thailand: data.go.th) และเผยแพร่โดยการประปานครหลวง (Metropolitan Waterworks Authority: MWA)

ข้อมูลต้นทางถูกจัดเก็บและเผยแพร่ในรูปแบบข้อมูลสรุปรายเดือน (Monthly Aggregated Data) อยู่แล้ว โดยแต่ละระเบียนแสดงปริมาณการใช้น้ำประปารวมของพื้นที่และประเภทผู้ใช้น้ำในแต่ละเดือน ดังนั้น ผู้วิจัยจึงนำข้อมูลมาใช้ในการวิเคราะห์โดยตรง โดยไม่ต้องดำเนินการรวมข้อมูลตามช่วงเวลาเพิ่มเติม (Temporal Aggregation) ทั้งนี้เพื่อรักษาค่าปริมาณการใช้น้ำประปาที่สะท้อนการใช้งานจริงในแต่ละเดือน และให้สอดคล้องกับวัตถุประสงค์ของการพยากรณ์ปริมาณการใช้น้ำประปารายเดือน

สำหรับการสร้างแบบจำลอง ผู้วิจัยได้คัดเลือกตัวแปรที่สำคัญจำนวน 6 ตัวแปร ได้แก่ (1) ปี พ.ศ. (2) เดือน (3) รหัสจังหวัด (4) รหัสอำเภอ (5) รหัสประเภทผู้ใช้น้ำ และ (6) ปริมาณการใช้น้ำประปา ซึ่งเป็นตัวแปรเป้าหมาย (Target Variable) สำหรับการพยากรณ์

2) การกลั่นกรองข้อมูล (Data Cleaning) ในขั้นตอนการกลั่นกรองข้อมูล ผู้วิจัยได้ดำเนินการตรวจสอบความถูกต้อง ความครบถ้วน และความสอดคล้องของข้อมูล เพื่อเพิ่มคุณภาพ

ของชุดข้อมูลก่อนนำไปสร้างแบบจำลอง โดยมีขั้นตอนสำคัญดังนี้

- การปรับรูปแบบภาษา (Data Standardization) เนื่องจากข้อมูลต้นทางจัดเก็บในรูปแบบภาษาไทย ผู้วิจัยได้ดำเนินการแปลงข้อมูลบางตัวแปร เช่น เดือน และประเภทผู้ใช้น้ำ ให้อยู่ในรูปแบบภาษาอังกฤษ เพื่อให้สามารถประมวลผลด้วยเครื่องมือวิเคราะห์ข้อมูลและซอฟต์แวร์ทางสถิติได้อย่างมีประสิทธิภาพ

- การจัดการค่าว่าง (Missing Values) ทำการตรวจสอบค่าที่ขาดหาย และแทนค่าที่สูญหายด้วยค่ากลางหรือค่าที่เหมาะสมตามลักษณะของข้อมูล เพื่อหลีกเลี่ยงผลกระทบต่อความแม่นยำของแบบจำลอง

- การตรวจสอบและลบข้อมูลซ้ำ (Duplicate Records) ดำเนินการตรวจสอบระเบียนข้อมูลที่ซ้ำซ้อนทั้งในด้านตัวแปรและค่า และลบข้อมูลที่ซ้ำออกจากชุดข้อมูล

- การตรวจสอบค่าผิดปกติ (Outliers Detection) วิเคราะห์ค่าที่มีการเบี่ยงเบนจากรูปแบบข้อมูลทั่วไปอย่างมีนัยสำคัญ และดำเนินการจัดการข้อมูลที่เป็นค่าผิดปกติเพื่อลดผลกระทบต่อประสิทธิภาพของแบบจำลอง

3) การแปลงข้อมูล (Data Transformation) ผู้วิจัยได้ดำเนินการแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสมสำหรับการวิเคราะห์เชิงเวลาและการพัฒนาแบบจำลองพยากรณ์ โดยมีขั้นตอนสำคัญ ได้แก่

- แปลงข้อมูลเดือนให้อยู่ในรูปแบบเชิงตัวเลข (1–12) เพื่อใช้ในการจัดลำดับข้อมูลอนุกรมเวลา

- รวมข้อมูลปีและเดือนเป็นตัวแปรเชิงเวลาในรูปแบบ Year-Month และสร้างตัวแปรดัชนีเวลา (Time Index) จากลำดับของเดือนในชุดข้อมูล โดยกำหนดให้เดือนมกราคม พ.ศ. 2557 มีค่าเท่ากับ 1 และเพิ่มขึ้นทีละ 1 ตามลำดับเดือนจนถึงเดือนธันวาคม พ.ศ. 2566 ทั้งนี้ ระเบียนข้อมูลที่อยู่ในเดือนเดียวกันจะมีค่า Time Index เดียวกัน เพื่อใช้แทนแนวโน้มการเปลี่ยนแปลงของปริมาณการใช้น้ำประปาตามช่วงเวลา

- สร้างตัวแปรฤดูกาล (Seasonal Feature) จากตัวแปรเดือน (Month) โดยแบ่งข้อมูลออกเป็น 3 ฤดูกาล ได้แก่ ฤดูร้อน (มีนาคม–พฤษภาคม) ฤดูฝน (มิถุนายน–ตุลาคม) และฤดูหนาว (พฤศจิกายน–กุมภาพันธ์) จากนั้นแปลงข้อมูลให้อยู่ในรูปแบบ

One-Hot Encoding เพื่อให้แบบจำลองสามารถเรียนรู้ผลกระทบของฤดูกาลที่มีต่อปริมาณการใช้น้ำประปาได้อย่างเหมาะสม

4) การเลือกตัวแปร (Feature Selection) ผู้วิจัยได้ทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรต้น (Independent Variables) และตัวแปรตาม คือ ปริมาณการใช้น้ำประปา โดยใช้เครื่องมือทางสถิติและการวิเคราะห์ความสัมพันธ์เชิงสหสัมพันธ์ เพื่อคัดเลือกรูปแบบตัวแปรที่มีอิทธิพลต่อการพยากรณ์มากที่สุด ทั้งนี้เพื่อช่วยลดมิติของข้อมูลและเพิ่มประสิทธิภาพของแบบจำลอง

5) การแบ่งชุดข้อมูล (Data Splitting) เพื่อให้การประเมินประสิทธิภาพของแบบจำลองมีความถูกต้องและน่าเชื่อถือ งานวิจัยนี้ได้ดำเนินการแบ่งชุดข้อมูลออกเป็น 3 ส่วน ได้แก่ ชุดข้อมูลฝึกสอน (Training set) ร้อยละ 80 สำหรับการเรียนรู้แบบจำลอง ชุดข้อมูลตรวจสอบ (Validation set) ร้อยละ 10 สำหรับการปรับแต่งค่าพารามิเตอร์ของแบบจำลอง และชุดข้อมูลทดสอบ (Testing set) ร้อยละ 10 สำหรับการประเมินประสิทธิภาพขั้นสุดท้ายของแบบจำลอง ซึ่งเป็นแนวปฏิบัติมาตรฐานในงานด้านการเรียนรู้ของเครื่อง (Machine Learning) เพื่อหลีกเลี่ยงปัญหาการจดจำข้อมูลเฉพาะ (Overfitting) และเพิ่มความสามารถในการอธิบายทั่วไป (Generalization) ของแบบจำลอง ในกระบวนการแบ่งข้อมูล ผู้วิจัยได้ใช้วิธีการแบ่งข้อมูลตามลำดับเวลา (Time-based Splitting) โดยกำหนดให้ข้อมูลในช่วงเวลาก่อนหน้านี้เป็นชุดฝึกสอนและชุดตรวจสอบ ขณะที่ข้อมูลในช่วงเวลาล่าสุดถูกใช้เป็นชุดทดสอบ เพื่อป้องกันปัญหาการรั่วไหลของข้อมูล (Data Leakage) และสะท้อนลักษณะการใช้งานจริงของระบบพยากรณ์เชิงเวลา ชุดข้อมูล Validation ถูกนำมาใช้สำหรับการปรับแต่งพารามิเตอร์ของแบบจำลอง (Hyperparameter Tuning) และการคัดเลือกโครงสร้างแบบจำลองที่เหมาะสม ก่อนนำแบบจำลองที่ผ่านการปรับแต่งแล้วไปประเมินผลขั้นสุดท้ายด้วยชุดข้อมูล Testing ซึ่งช่วยให้ผลการประเมินมีความน่าเชื่อถือและสามารถสะท้อนประสิทธิภาพของแบบจำลองได้อย่างเหมาะสมตามบริบทการใช้งานจริง

4.4 การสร้างแบบจำลอง (Modeling)

ในขั้นตอนการสร้างแบบจำลอง ผู้วิจัยได้นำข้อมูลผ่านกระบวนการเตรียมข้อมูลและการแบ่งชุดข้อมูลมาใช้ในการ

พัฒนาแบบจำลองพยากรณ์ปริมาณการใช้น้ำประปา โดยเลือกใช้แบบจำลองจำนวน 4 ประเภท ได้แก่ Artificial Neural Network (ANN), Linear Regression (LR), Random Forest (RF) และ Gradient Boosted Trees (GBT) ซึ่งเป็นแบบจำลองที่ได้รับความนิยมในการวิเคราะห์ข้อมูลเชิงพยากรณ์และสามารถรองรับข้อมูลที่มีความซับซ้อนได้

เพื่อเพิ่มประสิทธิภาพของแบบจำลอง ผู้วิจัยได้ดำเนินการปรับแต่งค่าพารามิเตอร์ (Hyperparameter Tuning) โดยใช้ชุดข้อมูล Validation เป็นเกณฑ์ในการประเมินผล และทดลองปรับค่าพารามิเตอร์หลายชุดสำหรับแต่ละแบบจำลอง เช่น จำนวนชั้นซ่อนและจำนวนโหนดของ ANN จำนวนต้นไม้ของ Random Forest และจำนวนต้นไม้ร่วมกับอัตราการเรียนรู้ของ Gradient Boosted Trees จากนั้นเลือกชุดค่าพารามิเตอร์ที่ให้ค่าความคลาดเคลื่อนต่ำที่สุดบนชุดข้อมูล Validation เพื่อนำไปใช้ในการสร้างแบบจำลองขั้นสุดท้ายก่อนประเมินผลด้วยชุดข้อมูล Testing

ตารางที่ 1. ค่าพารามิเตอร์ที่ได้รับการคัดเลือกสำหรับแต่ละแบบจำลอง

แบบจำลอง	ค่าพารามิเตอร์ที่ได้รับการคัดเลือก
ANN	Hidden Layers = 2, Neurons = (64, 32), Learning Rate = 0.001, Batch Size = 32, Epochs = 100
Random Forest	Number of Trees = 100, Max Depth = 20
Gradient Boosted Trees	Number of Trees = 100, Learning Rate = 0.05, Max Depth = 5
Linear Regression	ใช้ค่าพารามิเตอร์เริ่มต้นของอัลกอริทึม

ค่าพารามิเตอร์ที่แสดงในตารางที่ 1 เป็นค่าที่ให้ผลการพยากรณ์ดีที่สุดบนชุดข้อมูล Validation และถูกนำมาใช้ในการสร้างแบบจำลองขั้นสุดท้ายก่อนประเมินประสิทธิภาพด้วยชุดข้อมูล Testing

กระบวนการ Grid Search ทำการทดสอบชุดค่าพารามิเตอร์ที่กำหนดไว้ล่วงหน้าหลายรูปแบบ และประเมินประสิทธิภาพของแต่ละชุดด้วยค่า Root Mean Squared Error (RMSE) บนชุด Validation จากนั้นเลือกชุดค่าพารามิเตอร์ที่ให้ผลการพยากรณ์ดีที่สุดเพื่อนำไปใช้ในการสร้างแบบจำลองขั้นสุดท้าย

สำหรับแบบจำลอง ANN ผู้วิจัยใช้โครงสร้างแบบ Feedforward Neural Network ประกอบด้วยชั้นซ่อนจำนวน 2 ชั้น โดยกำหนดจำนวนโหนดเท่ากับ 64 และ 32 โหนด ตามลำดับ ซึ่งเลือกจากผลการทดลองเบื้องต้นโดยพิจารณาค่า Validation Error ต่ำสุด ใช้ฟังก์ชันกระตุ้นแบบ ReLU และ อัลกอริทึม Adam สำหรับการปรับค่าน้ำหนัก กำหนดอัตราการเรียนรู้เท่ากับ 0.001 ขนาดชุดข้อมูลฝึกต่อรอบเท่ากับ 32 และ จำนวนรอบการฝึกเท่ากับ 100 รอบ โดยใช้ฟังก์ชันสูญเสียแบบ Mean Squared Error เพื่อเพิ่มความแม่นยำในการพยากรณ์

นอกจากตัวแปรเชิงพื้นที่และตัวแปรเชิงเวลาแล้ว ผู้วิจัยได้ สร้างตัวแปรฤดูกาล (Seasonal Feature) จากข้อมูลเดือนและ แปลงให้อยู่ในรูปแบบ One-Hot Encoding เพื่อสะท้อน ผลกระทบของฤดูกาลต่อปริมาณการใช้น้ำประปา โดยตัวแปร ดังกล่าวถูกนำไปใช้ร่วมกับตัวแปร ปีรหัสจังหวัด รหัสอำเภอ และรหัสประเภทผู้ใช้น้ำ เป็นข้อมูลนำเข้าสำหรับแบบจำลอง Artificial Neural Network, Random Forest และ Gradient Boosted Trees

สำหรับแบบจำลอง Linear Regression ผู้วิจัยใช้ตัวแปร ดัชนีเวลา (Time Index) เป็นตัวแปรอิสระในการอธิบาย แนวโน้มการเปลี่ยนแปลงของปริมาณการใช้น้ำประปาตาม เวลา โดยสมมติให้ตัวแปรเป้าหมายสามารถอธิบายได้ด้วย ความสัมพันธ์เชิงเส้นระหว่าง Time Index และปริมาณการใช้น้ำประปา ทั้งนี้ ก่อนการสร้างแบบจำลองได้ดำเนินการปรับ สเกลข้อมูลและแปลงตัวแปรเวลาให้อยู่ในรูปแบบดัชนีเวลา (Time Index) เพื่อเพิ่มเสถียรภาพของการประมาณค่า สัมประสิทธิ์และลดปัญหาความเอนเอียงเชิงตัวเลข (Numerical Instability) แบบจำลองถูกประเมินด้วยข้อมูล Validation Set เพื่อคัดเลือกพารามิเตอร์ที่เหมาะสมก่อนนำไปทดสอบ ประสิทธิภาพกับชุดข้อมูล Testing Set

สำหรับแบบจำลอง Random Forest ผู้วิจัยใช้เทคนิค Ensemble Learning โดยสร้างชุดต้นไม้ตัดสินใจหลายต้นและ รวมผลการพยากรณ์ด้วยวิธีการเฉลี่ยค่า (Averaging) เพื่อเพิ่ม เสถียรภาพของแบบจำลองและลดความแปรปรวนของผลลัพธ์ โดยกำหนดจำนวนต้นไม้ (Number of Trees) เท่ากับ 100 ต้น และกำหนดความลึกสูงสุดของต้นไม้ (Maximum Depth) เท่ากับ 20 ระดับ ทั้งนี้ ใช้วิธีสุ่มตัวอย่างข้อมูลแบบ Bootstrap

Sampling และการสุ่มเลือกตัวแปรย่อยในแต่ละโหนด เพื่อเพิ่ม ความสามารถในการเรียนรู้รูปแบบข้อมูลที่มีความไม่เชิงเส้น และลดปัญหา overfitting ของแบบจำลอง

สำหรับแบบจำลอง Gradient Boosted Trees ผู้วิจัยใช้ แนวคิด Boosting ซึ่งเป็นการสร้างต้นไม้ตัดสินใจแบบลำดับขั้น โดยแต่ละต้นจะเรียนรู้จากค่าความคลาดเคลื่อนของต้นก่อน หน้า เพื่อเพิ่มความแม่นยำของแบบจำลองโดยรวม โดยกำหนด จำนวนต้นไม้เท่ากับ 100 ต้น ค่าอัตราการเรียนรู้ (Learning Rate) เท่ากับ 0.05 และความลึกสูงสุดของต้นไม้เท่ากับ 5 ระดับ ทั้งนี้ การกำหนดพารามิเตอร์ดังกล่าวช่วยควบคุมความซับซ้อน ของโมเดล ลดความเสี่ยงของการเกิด overfitting และเพิ่ม เสถียรภาพของผลการพยากรณ์ในข้อมูลอนุกรมเวลา

โดยแบบจำลองทั้ง 4 เทคนิคถูกพัฒนาและปรับแต่ง พารามิเตอร์โดยอาศัยชุดข้อมูล Validation เพื่อคัดเลือกค่าที่ ให้ผลการเรียนรู้มีเสถียรภาพและลดความคลาดเคลื่อนในการ พยากรณ์ จากนั้นจึงนำแบบจำลองที่ได้ไปประเมิน ประสิทธิภาพขั้นสุดท้ายด้วยชุดข้อมูล Testing Set เพื่อสะท้อน สมรรถนะของโมเดลภายใต้สภาพการใช้งานจริง

4.5 การประเมินผล (Evaluation)

การประเมินประสิทธิภาพของแบบจำลองพยากรณ์ ผู้วิจัยได้ใช้ ตัวชี้วัดทางสถิติที่ได้รับการยอมรับอย่างแพร่หลายในงานด้าน การพยากรณ์และการเรียนรู้ของเครื่อง ได้แก่ ค่าเฉลี่ยของความ คลาดเคลื่อนสัมบูรณ์ (Mean Absolute Error: MAE) และ ค่ารากที่สองของความคลาดเคลื่อนกำลังสอง (Root Mean Square Error: RMSE) เพื่อเปรียบเทียบความแม่นยำของ แบบจำลองแต่ละประเภท โดยตัวชี้วัดทั้งสองสะท้อนระดับ ความคลาดเคลื่อนระหว่างค่าที่พยากรณ์และค่าจริง และสามารถประเมินประสิทธิภาพของแบบจำลองเชิงตัวเลขได้ อย่างมีประสิทธิภาพ

1) ค่ารากที่สองของความคลาดเคลื่อนกำลังสอง (Root Mean Squared Error: RMSE) ดังสมการที่ 1 ซึ่งใช้ วัดขนาดของความคลาดเคลื่อนโดยรวมของแบบจำลอง โดย ค่าที่ต่ำกว่าบ่งชี้ถึงความแม่นยำที่สูงกว่า

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (1)$$

โดย n คือ จำนวนข้อมูลทั้งหมด

y_i คือ ค่าจริงของข้อมูลลำดับที่ i

$\hat{y}_i$ คือ ค่าที่โมเดลพยากรณ์ได้ของลำดับที่ i

$(y_i - \hat{y}_i)^2$ คือ ค่าคลาดเคลื่อนกำลังสองของแต่ละจุดข้อมูล

2) ค่าเฉลี่ยของผลต่างสัมบูรณ์ (Mean Absolute Error: MAE) ดังสมการที่ 2 ซึ่งใช้วัดความคลาดเคลื่อนโดยเฉลี่ยในหน่วยเดียวกับข้อมูลจริง ให้ภาพรวมที่ชัดเจนของค่าที่ทำนายผิดพลาด

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (2)$$

โดย n คือ จำนวนข้อมูลทั้งหมด

y_i คือ ค่าจริงของข้อมูลลำดับที่ i

$\hat{y}_i$ คือ ค่าที่โมเดลพยากรณ์ได้ของลำดับที่ i

$|y_i - \hat{y}_i|$ คือ ค่าสัมบูรณ์ของค่าคลาดเคลื่อน

ในงานวิจัยนี้ แบบจำลองที่ให้ค่า MAE และ RMSE ต่ำที่สุดจะถูกพิจารณาว่าเป็นแบบจำลองที่มีประสิทธิภาพสูงสุดและเหมาะสมสำหรับการนำไปประยุกต์ใช้ในการพัฒนาระบบพยากรณ์การใช้น้ำประปาเชิงดิจิทัล

4.6 การนำไปใช้งาน (Deployment)

การนำแบบจำลองที่ผ่านกระบวนการพัฒนาและประเมินประสิทธิภาพไปใช้งานจริงเป็นขั้นตอนสำคัญของกรอบ CRISP-DM โดยมีเป้าหมายเพื่อแปลงผลลัพธ์เชิงเทคนิคให้อยู่ในรูปแบบที่สามารถนำไปใช้สนับสนุนการตัดสินใจเชิงบริหารและการกำหนดนโยบายด้านการจัดการทรัพยากรน้ำได้อย่างมีประสิทธิภาพ ในงานวิจัยนี้ กระบวนการนำแบบจำลองไปใช้งานประกอบด้วย 3 ขั้นตอนหลัก ดังนี้

1) ขั้นตอนการคัดเลือกแบบจำลองที่เหมาะสม ภายหลังจากการพัฒนาแบบจำลองทั้ง 4 ประเภท ได้แก่ Artificial Neural Network (ANN), Linear Regression (LR), Random Forest (RF) และ Gradient Boosted Trees (GB) ผู้วิจัยได้ประเมินประสิทธิภาพของแต่ละแบบจำลองด้วยตัวชี้วัดทางสถิติ ได้แก่ Mean Absolute Error (MAE) และ Root Mean Square Error (RMSE) โดยแบบจำลองที่ให้ค่าความคลาดเคลื่อนต่ำที่สุดและมีความเสถียรในการพยากรณ์ถูกเลือกเป็นแบบจำลองหลักสำหรับการนำไปใช้งานจริง เพื่อให้สามารถพยากรณ์ปริมาณการใช้น้ำประปาได้อย่างแม่นยำและเชื่อถือได้

2) ขั้นตอนการนำแบบจำลองที่ผ่านการคัดเลือกไปใช้ในการพยากรณ์ข้อมูลในอนาคต โดยใช้ข้อมูลชุดใหม่ (New Incoming Data) เป็นข้อมูลนำเข้า เพื่อสร้างผลลัพธ์การพยากรณ์ที่สะท้อนสถานการณ์การใช้น้ำในช่วงเวลาถัดไป ผลลัพธ์ดังกล่าวสามารถนำไปใช้สนับสนุนการวางแผนเชิงบริหาร เช่น การเตรียมความพร้อมด้านกำลังการผลิตน้ำในช่วงฤดูแล้ง การบริหารจัดการน้ำในช่วงความต้องการสูง และการจัดสรรทรัพยากรน้ำในแต่ละพื้นที่ให้สอดคล้องกับความต้องการใช้งานจริง

3) ขั้นตอนการนำเสนอและสื่อสารผลลัพธ์การพยากรณ์ผ่านระบบสารสนเทศบนแพลตฟอร์มเว็บ โดยผู้วิจัยได้พัฒนาเว็บไซต์ด้วยเทคโนโลยี HTML, PHP และ JavaScript เพื่อแสดงผลลัพธ์ในรูปแบบกราฟเส้น (Line Chart) และตารางสรุปข้อมูล ซึ่งช่วยให้ผู้ใช้งานสามารถเข้าใจแนวโน้มการใช้น้ำได้อย่างชัดเจนและสามารถเข้าถึงข้อมูลได้แบบออนไลน์อย่างสะดวกและรวดเร็ว ระบบดังกล่าวสามารถนำไปใช้เป็นเครื่องมือสนับสนุนการตัดสินใจของหน่วยงานภาครัฐ องค์กรท้องถิ่น และผู้มีส่วนได้ส่วนเสียอื่น ๆ รวมถึงประชาชนทั่วไปที่ต้องการข้อมูลเพื่อวางแผนการใช้น้ำในระดับครัวเรือนและชุมชน

5. ผลการวิจัย

การวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของแบบจำลองการพยากรณ์ปริมาณการใช้น้ำประปาในประเทศไทย โดยประยุกต์ใช้กระบวนการวิเคราะห์ข้อมูลตามกรอบ CRISP-DM และเทคนิคเหมืองข้อมูลจำนวน 4 แบบ ได้แก่ โครงข่ายประสาทเทียม (Artificial Neural Network: ANN) การถดถอยเชิงเส้น (Linear Regression) แรนดอมฟอเรสต์ (Random Forest) และต้นไม้เสริมกำลังแบบไล่ระดับ (Gradient Boosted Trees) ข้อมูลที่ใช้เป็นข้อมูลการใช้น้ำประปารายเดือนในช่วงปี พ.ศ. 2557–2566 ซึ่งรวบรวมจากศูนย์กลางข้อมูลเปิดภาครัฐ และผ่านกระบวนการเตรียมข้อมูลก่อนนำไปสร้างแบบจำลอง

ในการแบ่งชุดข้อมูล ผู้วิจัยได้ดำเนินการแบ่งข้อมูลออกเป็น 3 ส่วน ได้แก่ ชุดข้อมูลสำหรับการฝึกสอนแบบจำลอง (Training Set) ร้อยละ 80 ชุดข้อมูลตรวจสอบ (Validation Set) และปรับแต่งพารามิเตอร์ของแบบจำลอง ร้อยละ 10 และชุดข้อมูลสำหรับการทดสอบประสิทธิภาพขั้นสุดท้าย (Testing

Set) ร้อยละ 10 โดยใช้วิธีการแบ่งข้อมูลตามลำดับเวลา (Time-based Splitting) เพื่อป้องกันปัญหาการรั่วไหลของข้อมูล (Data Leakage) และสะท้อนลักษณะการใช้งานจริงของระบบพยากรณ์เชิงเวลา จากนั้นใช้โปรแกรม RapidMiner เป็นเครื่องมือในการพัฒนาแบบจำลอง การประเมินประสิทธิภาพของแบบจำลองดำเนินการโดยใช้ตัวชี้วัดทางสถิติ ได้แก่ ค่าเฉลี่ยของค่าความคลาดเคลื่อนสัมบูรณ์ (Mean Absolute Error: MAE) และค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (Root Mean Square Error: RMSE) เพื่อเปรียบเทียบระดับความแม่นยำของแต่ละเทคนิค ผลการเปรียบเทียบประสิทธิภาพของแบบจำลองแสดงดังตารางที่ 2

ตารางที่ 2. ผลการเปรียบเทียบประสิทธิภาพของแบบจำลองพยากรณ์การใช้น้ำประปา

Model	MAE (ลบ.ม.)	RMSE (ลบ.ม.)
Artificial Neural Network	264,804.09	459,365.09
Linear Regression	188,158.96	310,809.74
Random Forest	543,430.74	806,258.97
Gradient Boosted Trees	1,631,260.09	2,609,506.10

จากตารางที่ 2 พบว่า แบบจำลอง Linear Regression ให้ค่าความคลาดเคลื่อนต่ำที่สุดทั้งในด้านค่า Mean Absolute Error (MAE) และ Root Mean Square Error (RMSE) เมื่อเปรียบเทียบกับแบบจำลองอื่นทั้งหมด แสดงให้เห็นว่าแบบจำลองดังกล่าวมีประสิทธิภาพสูงสุดในการพยากรณ์ปริมาณการใช้น้ำประปาภายใต้ชุดข้อมูลและบริบทของการศึกษานี้

เมื่อพิจารณาเปรียบเทียบกับปริมาณการใช้น้ำประปาสูงสุดในชุดข้อมูล ซึ่งมีค่าเท่ากับ 1,240,771 ลูกบาศก์เมตร พบว่า ค่า MAE ของแบบจำลอง Linear Regression คิดเป็นร้อยละ 15.16 และค่า RMSE คิดเป็นร้อยละ 25.04 ของค่าสูงสุด ผลการวิเคราะห์ดังกล่าวแสดงให้เห็นว่าแบบจำลอง Linear Regression มีประสิทธิภาพสูงสุดเมื่อเปรียบเทียบกับผลดังกล่าว สะท้อนว่าข้อมูลที่ใช้ในการศึกษามีแนวโน้ม (Trend) ชัดเจน และยังไม่มีรูปแบบฤดูกาลหรือความไม่เชิงเส้นที่ซับซ้อนมากนัก จึงทำให้แบบจำลอง Linear Regression สามารถอธิบายความสัมพันธ์ของข้อมูลได้ดีกว่าแบบจำลองอื่นภายใต้ชุดข้อมูลนี้

อย่างไรก็ตาม ผลการศึกษานี้ควรพิจารณาภายใต้ขอบเขตของชุดข้อมูลที่ใช้ในการวิจัย เนื่องจากตัวแปรที่นำมาวิเคราะห์ประกอบด้วยข้อมูลเชิงเวลาและข้อมูลเชิงพื้นที่ที่ได้จากฐานข้อมูลเปิดภาครัฐเป็นหลัก โดยยังไม่ได้รวมปัจจัยภายนอกที่อาจส่งผลกระทบต่อปริมาณการใช้น้ำประปา เช่น อุณหภูมิ ปริมาณน้ำฝน ความชื้น จำนวนประชากร การขยายตัวของเมือง และกิจกรรมทางเศรษฐกิจ ดังนั้น แม้แบบจำลอง Linear Regression จะให้ผลการพยากรณ์ที่ดีที่สุดภายใต้ชุดข้อมูลที่ใช้ศึกษา แต่ประสิทธิภาพของแบบจำลองอาจเปลี่ยนแปลงได้เมื่อมีการบูรณาการตัวแปรภายนอกเพิ่มเติมเข้าสู่กระบวนการวิเคราะห์

ผลการทดลองพบว่าแบบจำลอง Gradient Boosted Trees (GBT) ให้ค่าความคลาดเคลื่อนสูงกว่า Linear Regression และ Random Forest ทั้งในด้าน RMSE และ MAE ซึ่งอาจสะท้อนถึงแนวโน้มการเกิด Overfitting ของแบบจำลอง กล่าวคือแบบจำลองสามารถเรียนรู้รูปแบบเฉพาะของข้อมูลฝึกสอนได้ดี แต่ไม่สามารถอธิบายข้อมูลที่ไม่เคยพบมาก่อนได้อย่างมีประสิทธิภาพ

สาเหตุหนึ่งอาจเกิดจากลักษณะการทำงานของ GBT ที่สร้างต้นไม้ตัดสินใจหลายต้นแบบลำดับต่อเนื่อง (Sequential Learning) เพื่อแก้ไขข้อผิดพลาดของต้นไม้ก่อนหน้า หากกำหนดจำนวนต้นไม้ (Number of Trees) มากเกินไป หรือกำหนดความลึกของต้นไม้ (Tree Depth) สูงเกินความจำเป็น แบบจำลองอาจเรียนรู้รายละเอียดและความผันผวนเฉพาะของข้อมูลฝึกสอนมากเกินไป ส่งผลให้ความสามารถในการอธิบายทั่วไป (Generalization) ลดลงเมื่อทดสอบกับข้อมูลใหม่

แนวทางในการลดปัญหา Overfitting ของ GBT ได้แก่ การลดค่า Learning Rate การจำกัดความลึกของต้นไม้ การลดจำนวนต้นไม้ และการใช้เทคนิค Early Stopping เพื่อหยุดการเรียนรู้เมื่อประสิทธิภาพบนชุดข้อมูล Validation ไม่ปรับตัวดีขึ้น ซึ่งอาจช่วยเพิ่มประสิทธิภาพของแบบจำลองในการศึกษาครั้งต่อไป

สมการของแบบจำลอง Linear Regression สามารถเขียนได้ดังสมการที่ 3

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \varepsilon \quad (3)$$

เมื่อ

Y = ปริมาณการใช้น้ำประปาที่ต้องการพยากรณ์

β_0 = ค่าคงที่ (Intercept)

$\beta_1, \beta_2, \dots, \beta_n$ = ค่าสัมประสิทธิ์ของตัวแปรอิสระ

$X_1, X_2, \dots, X_n$ = ตัวแปรอิสระ

$\mathcal{E}$ = ค่าความคลาดเคลื่อน (Error Term)

จากผลการวิเคราะห์ด้วยโปรแกรม RapidMiner ดังแสดงในรูปสมการที่ 3 สามารถนำผลลัพธ์มาสร้างสมการพยากรณ์ปริมาณการใช้น้ำประปาโดยใช้เทคนิคการถดถอยเชิงเส้น (Linear Regression) ซึ่งใช้สำหรับอธิบายความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตาม โดยตัวแปรอิสระในงานวิจัยนี้คือดัชนีเวลา (Time Index) ที่แทนลำดับช่วงข้อมูลในช่วงปี พ.ศ. 2557–2566 ซึ่งผ่านกระบวนการปรับสเกลก่อนนำเข้าสู่แบบจำลองรูปแบบสมการถดถอยเชิงเส้นสามารถแสดงได้ดังสมการที่ 4

$$Y_t = \beta_0 + \beta_1 T_t \quad (4)$$

โดยที่

Y_t คือ ปริมาณการใช้น้ำประปาที่ถูกพยากรณ์ในช่วงเวลา t

T_t คือ ดัชนีเวลา (Time Index) ที่ผ่านการปรับสเกลแล้ว

β_0 คือ ค่าคงที่ของแบบจำลอง (Intercept)

β_1 คือ ค่าสัมประสิทธิ์ถดถอยของตัวแปรเวลา

ทั้งนี้ เพื่อหลีกเลี่ยงปัญหาที่เกิดจากการใช้ค่าปีจริงซึ่งมีขนาดตัวเลขสูงและอาจส่งผลต่อเสถียรภาพของการประมาณค่าสัมประสิทธิ์ ผู้วิจัยได้ทำการแปลงข้อมูลปีให้อยู่ในรูปแบบดัชนีเวลา (Time Index) และปรับสเกลข้อมูลก่อนนำเข้าสู่กระบวนการสร้างแบบจำลอง ซึ่งช่วยลดปัญหาความเอนเอียงเชิงตัวเลข (Numerical Instability) และเพิ่มความเสถียรของแบบจำลองในการเรียนรู้แนวโน้มเชิงเวลา

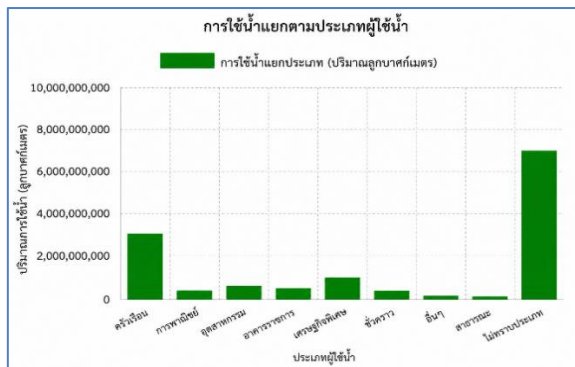
สมการพยากรณ์ดังกล่าวถูกนำไปใช้ในการคาดการณ์ปริมาณการใช้น้ำประปาในอนาคต โดยการแทนค่าดัชนีเวลาของช่วงเวลาที่ต้องการพยากรณ์ลงในสมการ จากนั้นทำการคำนวณเพื่อให้ได้ค่าพยากรณ์ Y_t ซึ่งใช้เป็นข้อมูลสนับสนุนการแสดงผลบนระบบสารสนเทศเชิงพยากรณ์บนแพลตฟอร์มเว็บ

ภายหลังจากการคัดเลือกแบบจำลองที่ให้ประสิทธิภาพดีที่สุด ผู้วิจัยได้นำแบบจำลองดังกล่าวไปพัฒนาเป็นระบบสนับสนุนการตัดสินใจบนเว็บ (Web-based Decision Support System: DSS) เพื่อสนับสนุนการติดตาม วิเคราะห์ และพยากรณ์ปริมาณการใช้น้ำประปา โดยระบบได้รับการออกแบบให้แสดงผลข้อมูลในรูปแบบ Dashboard เพื่อช่วยให้ผู้ใช้งานสามารถติดตามแนวโน้มการใช้น้ำประปาและใช้ข้อมูลประกอบการวางแผนด้านการบริหารจัดการน้ำได้อย่างสะดวก

ระบบต้นแบบได้รับการพัฒนาในลักษณะ Client–Server Architecture โดยสามารถเชื่อมโยงฐานข้อมูลกับแบบจำลองพยากรณ์และแสดงผลผ่านเว็บไซต์ ตัวอย่างดังรูปที่ 2(a) ซึ่งแสดงกราฟแนวโน้มปริมาณการใช้น้ำประปารายปี ช่วยให้ผู้ใช้งานสามารถติดตามการเปลี่ยนแปลงของปริมาณการใช้น้ำในแต่ละปีและวิเคราะห์แนวโน้มของข้อมูลได้อย่างชัดเจน ขณะที่รูปที่ 2(b) แสดงปริมาณการใช้น้ำจำแนกตามประเภทผู้ใช้น้ำ ซึ่งช่วยให้สามารถเปรียบเทียบลักษณะการใช้น้ำของผู้ใช้น้ำแต่ละประเภท และใช้ประกอบการวางแผนการผลิตและการจัดสรรน้ำให้เหมาะสมกับความต้องการของแต่ละกลุ่มผู้ใช้น้ำ นอกจากนี้ การนำเสนอข้อมูลในรูปแบบแดชบอร์ดยังช่วยให้ผู้ใช้งานสามารถวิเคราะห์แนวโน้มและเปรียบเทียบข้อมูลได้อย่างรวดเร็ว ลดความซับซ้อนในการตีความข้อมูล และสนับสนุนการตัดสินใจเชิงบริหารได้อย่างมีประสิทธิภาพ



รูปที่ 2 (a) Dashboard แสดงแนวโน้มปริมาณการใช้น้ำประปารายปี



รูปที่ 2 (b) Dashboard แสดงปริมาณการใช้น้ำประจำวันแยกตามประเภทผู้ใช้น้ำ

ผลการพัฒนาระบบแสดงให้เห็นว่า เว็บไซต์สามารถทำหน้าที่เป็นศูนย์กลางข้อมูลด้านการใช้น้ำประปา (Water Information Dashboard) ที่สนับสนุนการจัดเก็บ การประมวลผล การพยากรณ์ และการแสดงผลข้อมูลผ่านเว็บได้อย่างครบถ้วน อีกทั้งยังรองรับการขยายระบบในอนาคต เช่น การเชื่อมต่อกับฐานข้อมูลภายนอกผ่าน API หรือข้อมูลจากอุปกรณ์ Internet of Things (IoT) เพื่อเพิ่มศักยภาพของระบบในการรองรับข้อมูลแบบเรียลไทม์ และสนับสนุนการพัฒนาระบบบริหารจัดการน้ำประปาเชิงอัจฉริยะ

6. สรุปและอภิปรายผล

ผลการวิจัยแสดงให้เห็นว่าแบบจำลอง Linear Regression ให้ประสิทธิภาพในการพยากรณ์ปริมาณการใช้น้ำประปาสูงที่สุดเมื่อเปรียบเทียบกับแบบจำลองอื่น ได้แก่ Artificial Neural Network, Random Forest และ Gradient Boosted Trees ภายใต้อัตราการผิดพลาดค่าเฉลี่ย (MAE) และค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (RMSE) ที่ต่ำที่สุด ผลลัพธ์ดังกล่าวสะท้อนถึงลักษณะของข้อมูลที่มีแนวโน้มเชิงเวลาและรูปแบบเชิงเส้นค่อนข้างสม่ำเสมอ ส่งผลให้แบบจำลองเชิงสถิติสามารถอธิบายความสัมพันธ์ระหว่างตัวแปรได้อย่างมีประสิทธิภาพและเหมาะสมต่อการนำไปประยุกต์ใช้งานจริง

จากผลการเปรียบเทียบประสิทธิภาพของแบบจำลองพบว่าแบบจำลอง Linear Regression ให้ค่าความคลาดเคลื่อนต่ำที่สุดเมื่อเปรียบเทียบกับ Artificial Neural Network (ANN),

Random Forest และ Gradient Boosted Trees ซึ่งสามารถอธิบายได้จากลักษณะของข้อมูลที่ใช้ในการศึกษา

เมื่อพิจารณาข้อมูลปริมาณการใช้น้ำประจำวันรายเดือนในช่วงปี พ.ศ. 2557–2566 พบว่าข้อมูลมีแนวโน้ม (Trend) ที่เปลี่ยนแปลงอย่างค่อยเป็นค่อยไปตามช่วงเวลา และมีรูปแบบฤดูกาล (Seasonality) ที่สัมพันธ์กับเดือนของปี อันเนื่องมาจากการเปลี่ยนแปลงของสภาพภูมิอากาศและพฤติกรรมการใช้น้ำในแต่ละฤดูกาล นอกจากนี้ ความสัมพันธ์ระหว่างตัวแปรเชิงเวลา ได้แก่ ปี เดือน และดัชนีเวลา (Time Index) กับปริมาณการใช้น้ำประปามีลักษณะใกล้เคียงเชิงเส้น (Linear Pattern) ในหลายช่วงของข้อมูล ส่งผลให้แบบจำลอง Linear Regression สามารถอธิบายแนวโน้มของข้อมูลได้อย่างมีประสิทธิภาพ

ในทางกลับกัน แบบจำลองที่มีความซับซ้อนมากกว่า เช่น Artificial Neural Network, Random Forest และ Gradient Boosted Trees แม้จะสามารถเรียนรู้ความสัมพันธ์แบบไม่เชิงเส้นได้ แต่ภายใต้ชุดข้อมูลที่ใช้ในการศึกษาซึ่งมีรูปแบบแนวโน้มและฤดูกาลค่อนข้างชัดเจน ประสิทธิภาพของแบบจำลองที่ซับซ้อนกว่าไม่ได้เหนือกว่า Linear Regression อย่างมีนัยสำคัญ อีกทั้ง Gradient Boosted Trees ยังมีแนวโน้มเกิด Overfitting จากการเรียนรู้รายละเอียดของข้อมูลฝึกสอนมากเกินไป ทำให้ประสิทธิภาพในการพยากรณ์ข้อมูลใหม่ลดลงเมื่อเปรียบเทียบกับแบบจำลองเชิงเส้น

ผลการศึกษานี้สอดคล้องกับงานวิจัยของ N. Fumo et al. [18] ซึ่งรายงานว่า Linear Regression สามารถให้ผลการพยากรณ์ที่มีเสถียรภาพในกรณีที่ข้อมูลความต้องการใช้น้ำมีแนวโน้มเชิงเส้น ขณะเดียวกัน งานของ A. Antunes et al. [16] และ I. I. Aminu [7] ชี้ให้เห็นว่าแบบจำลองในกลุ่ม Gradient Boosting และ Artificial Neural Network มีประสิทธิภาพสูงในบริบทของข้อมูลที่มีความไม่เชิงเส้นและมีความซับซ้อนสูง ผลการเปรียบเทียบดังกล่าวสนับสนุนข้อค้นพบของงานวิจัยนี้ว่าความเหมาะสมของแบบจำลองพยากรณ์ขึ้นอยู่กับลักษณะของข้อมูล โครงสร้างเชิงเวลา และบริบทการนำไปใช้งาน มากกว่าการเลือกใช้เทคนิคที่มีความซับซ้อนเพียงอย่างเดียว

ในเชิงการบริหารจัดการทรัพยากรน้ำ ผลลัพธ์จากแบบจำลอง Linear Regression ที่มีค่าความคลาดเคลื่อนอยู่ในระดับที่ยอมรับได้สะท้อนถึงศักยภาพในการนำไปประยุกต์ใช้

เป็นเครื่องมือสนับสนุนการตัดสินใจ (Decision Support Tool) สำหรับหน่วยงานที่เกี่ยวข้อง เช่น การวางแผนกำลังการผลิตน้ำ การบริหารการกระจายน้ำในช่วงฤดูแล้ง และการจัดลำดับความสำคัญของพื้นที่เสี่ยงต่อการขาดแคลนน้ำ นอกจากนี้ ความเรียบง่ายของแบบจำลองยังช่วยให้ผู้ใช้งานสามารถตีความผลลัพธ์ได้ง่ายและสามารถนำไปใช้ประกอบการตัดสินใจเชิงบริหารได้อย่างมีประสิทธิภาพ

เมื่อพิจารณาพร้อมกับผลการพัฒนาเว็บไซต์ ระบบสารสนเทศที่พัฒนาขึ้นสามารถเชื่อมโยงผลลัพธ์จากแบบจำลองพยากรณ์เข้ากับการแสดงผลในรูปแบบกราฟและแดชบอร์ดออนไลน์ ช่วยเพิ่มประสิทธิภาพในการสื่อสารข้อมูลและการเข้าถึงสารสนเทศเชิงพยากรณ์ของผู้มีส่วนได้ส่วนเสีย ระบบดังกล่าวจึงสามารถมองได้ว่าเป็นต้นแบบของเทคโนโลยีการจัดการที่บูรณาการข้อมูล การวิเคราะห์ และการนำเสนอผลลัพธ์ไว้ในแพลตฟอร์มเดียว และสามารถนำไปต่อยอดสู่ระบบบริหารจัดการน้ำอัจฉริยะในอนาคต

อย่างไรก็ตาม งานวิจัยนี้ยังมีข้อจำกัด ได้แก่ การใช้ข้อมูลเชิงเวลาระดับรายเดือน ซึ่งอาจไม่สามารถสะท้อนความผันผวนของปริมาณการใช้น้ำในระยะสั้นได้อย่างละเอียด นอกจากนี้ แบบจำลองที่พัฒนาขึ้นยังอาศัยข้อมูลเชิงเวลาและข้อมูลเชิงพื้นที่จากฐานข้อมูลเปิดภาครัฐเป็นหลัก โดยยังไม่ได้บูรณาการปัจจัยภายนอกที่อาจส่งผลกระทบต่อความต้องการใช้น้ำ เช่น อุณหภูมิ ปริมาณน้ำฝน ความชื้น จำนวนประชากร การขยายตัวของเมือง และปัจจัยด้านเศรษฐกิจและสังคม

แม้ว่างานวิจัยนี้จะมุ่งเน้นการเปรียบเทียบประสิทธิภาพของเทคนิคเหมืองข้อมูลบนแพลตฟอร์ม RapidMiner ได้แก่ Linear Regression, Artificial Neural Network, Random Forest และ Gradient Boosted Trees แต่แบบจำลองอนุกรมเวลา เช่น ARIMA, SARIMA, Holt-Winters และ Prophet ก็เป็นแนวทางที่ได้รับความนิยมในการพยากรณ์ข้อมูลอนุกรมเวลา เนื่องจากสามารถอธิบายองค์ประกอบของแนวโน้ม (Trend) และฤดูกาล (Seasonality) ได้อย่างมีประสิทธิภาพ อย่างไรก็ตาม เพื่อให้การเปรียบเทียบเป็นไปอย่างเป็นธรรมภายใต้ชุดข้อมูลและสภาพแวดล้อมเดียวกัน งานวิจัยนี้จึงมุ่งเปรียบเทียบเฉพาะเทคนิคเหมืองข้อมูลที่สามารถพัฒนาได้บนแพลตฟอร์ม RapidMiner

สำหรับการวิจัยในอนาคต ควรพิจารณาบูรณาการข้อมูลด้านสภาพภูมิอากาศ เช่น อุณหภูมิ ปริมาณน้ำฝน และความชื้น รวมถึงข้อมูลด้านประชากร การขยายตัวของเมือง และปัจจัยทางเศรษฐกิจและสังคม เพื่อเพิ่มความสามารถในการอธิบายความสัมพันธ์ของข้อมูลและปรับปรุงความแม่นยำของการพยากรณ์ นอกจากนี้ ควรเปรียบเทียบประสิทธิภาพของเทคนิคเหมืองข้อมูลกับแบบจำลองอนุกรมเวลา เช่น ARIMA, SARIMA, Holt-Winters และ Prophet รวมถึงศึกษาการประยุกต์ใช้แบบจำลองเชิงลึก (Deep Learning) เช่น Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU) และ Hybrid Models เพื่อประเมินศักยภาพในการเพิ่มประสิทธิภาพของระบบพยากรณ์ปริมาณการใช้น้ำประปาในอนาคต

7. ข้อเสนอแนะ

งานวิจัยการประยุกต์ใช้เทคนิคเหมืองข้อมูลเพื่อพัฒนาระบบพยากรณ์การใช้น้ำประปานี้ ได้นำเสนอข้อเสนอแนะสำหรับการปรับปรุงและพัฒนางานในอนาคต รวมถึงแนวทางเชิงนโยบาย โดยแบ่งออกเป็น 2 ประเด็น ดังนี้

7.1 ข้อเสนอแนะการวิจัยในอนาคต

1) ควรเพิ่มเติมแปรด้านสภาพภูมิอากาศ เช่น อุณหภูมิ ปริมาณน้ำฝน ความชื้น และความเร็วลม รวมถึงปัจจัยทางเศรษฐกิจและสังคม เช่น จำนวนประชากร การขยายตัวของเมือง และกิจกรรมทางอุตสาหกรรม เพื่อเพิ่มความสามารถของแบบจำลองในการอธิบายความผันผวนของความต้องการใช้น้ำประปา

2) ควรศึกษาการประยุกต์ใช้แบบจำลองขั้นสูง เช่น Long Short-Term Memory (LSTM), Hybrid Model และ Ensemble Learning เพื่อเปรียบเทียบประสิทธิภาพกับแบบจำลองเชิงสถิติและ Machine Learning แบบดั้งเดิม โดยเฉพาะในกรณีที่ข้อมูลรายวันหรือรายชั่วโมง ซึ่งมีความละเอียดเชิงเวลาสูง

3) ควรพัฒนาระบบให้รองรับข้อมูลแบบเรียลไทม์จากอุปกรณ์ Internet of Things (IoT) เช่น มาตรวัดน้ำอัจฉริยะ (Smart Meter) และเซนเซอร์สิ่งแวดล้อม เพื่อเพิ่มความแม่นยำของการพยากรณ์และสนับสนุนการบริหารจัดการน้ำแบบอัจฉริยะ

4) การศึกษาการประยุกต์ใช้เทคนิคการเรียนรู้เชิงลึก (Deep Learning) ร่วมกับข้อมูลเชิงพื้นที่ (Spatial Data) เพื่อพัฒนาแบบจำลองพยากรณ์ระดับพื้นที่ เช่น ระดับเขตหรือชุมชน ซึ่งจะช่วยเพิ่มความละเอียดของผลการพยากรณ์

สำหรับการวิจัยในอนาคต ควรพิจารณาบูรณาการข้อมูลปัจจัยภายนอกที่อาจส่งผลกระทบต่อปริมาณการใช้น้ำประปา เช่น ข้อมูลสภาพอากาศ (อุณหภูมิ ปริมาณน้ำฝน ความชื้น และความเร็วลม) ข้อมูลด้านประชากร การขยายตัวของเมือง และกิจกรรมทางเศรษฐกิจและสังคม เพื่อเพิ่มความสามารถในการอธิบายความสัมพันธ์ของข้อมูลและปรับปรุงประสิทธิภาพของแบบจำลองพยากรณ์ให้มีความแม่นยำมากยิ่งขึ้น นอกจากนี้ ควรศึกษาการประยุกต์ใช้แบบจำลอง Deep Learning สำหรับข้อมูลอนุกรมเวลา เช่น Long Short-Term Memory (LSTM) และ Gated Recurrent Unit (GRU) รวมถึง Hybrid Models เพื่อเปรียบเทียบประสิทธิภาพกับเทคนิคเหมืองข้อมูลที่ใช้ในการศึกษาครั้งนี้ เนื่องจากแบบจำลองดังกล่าวสามารถเรียนรู้ความสัมพันธ์เชิงเวลาและรูปแบบข้อมูลที่ซับซ้อนได้ดีกว่าแบบจำลอง ANN แบบดั้งเดิม อันจะช่วยเพิ่มศักยภาพของระบบพยากรณ์ปริมาณการใช้น้ำประปาในอนาคต

7.2 ข้อเสนอแนะเชิงนโยบาย

1) หน่วยงานที่เกี่ยวข้องกับการบริหารจัดการน้ำประปา เช่น การประปานครหลวง การประปาส่วนภูมิภาค และองค์กรปกครองส่วนท้องถิ่น ควรนำระบบพยากรณ์เชิงดิจิทัลไปใช้เป็นเครื่องมือสนับสนุนการตัดสินใจเชิงบริหาร เพื่อเพิ่มประสิทธิภาพในการวางแผนกำลังการผลิตและการกระจายน้ำ

2) ภาครัฐควรส่งเสริมการบูรณาการข้อมูลจากหลายหน่วยงานเข้าสู่แพลตฟอร์มกลางด้านการจัดการน้ำ เพื่อพัฒนาเป็นระบบสารสนเทศเชิงบูรณาการสำหรับการบริหารจัดการน้ำในระดับประเทศ ซึ่งสอดคล้องกับแนวทาง Smart City และเป้าหมาย SDG 6 และ SDG 11

3) ควรสนับสนุนการเปิดเผยข้อมูลภาครัฐ (Open Data) อย่างเป็นระบบและมีมาตรฐาน เพื่อเอื้อต่อการนำข้อมูลไปใช้ประโยชน์เชิงวิชาการและเชิงนโยบาย รวมถึงการพัฒนาเครื่องมือพยากรณ์เชิงดิจิทัล

4) ควรส่งเสริมความร่วมมือระหว่างภาครัฐ ภาคเอกชน และสถาบันการศึกษาในการพัฒนาเทคโนโลยีและนวัตกรรมด้านการจัดการน้ำ เพื่อสนับสนุนการบริหารจัดการทรัพยากรน้ำอย่างยั่งยืนในระยะยาว

เอกสารอ้างอิง

- [1] The Water Project, “Water scarcity and global water crisis,” 2025. [Online]. Available: <https://thewaterproject.org/water-scarcity-water-crisis>
- [2] World Bank, “Climate change impact assessment on urban water demand,” 2025. [Online]. Available: <https://www.worldbank.org/en/topic/water>
- [3] UNICEF Thailand, “Water, sanitation and climate resilience in Thailand,” 2024. [Online]. Available: <https://www.unicef.org/thailand/wash>
- [4] Pollution Control Department, “Thailand environmental situation report,” 2025. [Online]. Available: <https://www.pcd.go.th>
- [5] Mordor Intelligence, “Smart irrigation market – growth, trends and forecast,” 2025. [Online]. Available: <https://www.mordorintelligence.com/industry-reports/smart-irrigation-market>
- [6] สำนักงานปลัดกระทรวงมหาดไทย, “สุนทรพจน์วันน้ำโลก ประจำปี พ.ศ. 2567,” 2567. [ออนไลน์]. แหล่งที่มา: <https://www.moi.go.th>
- [7] I. I. Aminu, “State-of-art implementation of computing intelligent models for water demand modelling: A decade review and future direction,” Journal of Water and Land Development, 2025. [Online]. Available: <https://www.jwld.pl/files/2025-03-JWLD-22.pdf>
- [8] J. Han, J. Pei, and M. Kamber, Data Mining: Concepts and Techniques, 4th ed. San Francisco, CA, USA: Morgan Kaufmann, 2022.

- [9] B. Marr, *Data Strategy: How to Profit from a World of Big Data, Analytics and Artificial Intelligence*. London, UK: Kogan Page, 2021.
- [10] Y. Zhang, X. Wang, J. Chen, and H. Li, "Application of data mining techniques in water resources management: A review," *Environmental Modelling & Software*, vol. 143, pp. 105104, 2021.
- [11] F. Martínez-Plumed, L. Contreras-Ochando, C. Ferri, J. Hernández-Orallo, M. Kull, N. Lachiche, M. J. Ramírez-Quintana, and P. Flach, "CRISP-DM twenty years later: From data mining processes to data science trajectories," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 8, pp. 3048–3061, 2021.
- [12] T. H. Davenport and S. Miller, "Working with AI: Real stories of human-machine collaboration," *MIT Sloan Management Review*, 2020.
- [13] S. Raschka, Y. Liu, V. Mirjalili, and D. Dzhulgakov, *Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2*. Birmingham, UK: Packt Publishing, 2020.
- [14] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 3rd ed. Sebastopol, CA, USA: O'Reilly Media, 2022.
- [15] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate Research*, vol. 30, pp. 79–82, 2021.
- [16] A. Antunes, J. P. Santos, and R. Martins, "Short-term water demand forecasting using machine learning techniques," *Water Resources Management*, vol. 35, no. 4, pp. 1327–1343, 2021.
- [17] W. Hao, A. Cominola, and A. Castelletti, "Comparing predictive machine learning models for short- and long-term urban water demand forecasting in Milan, Italy," *IFAC-PapersOnLine*, vol. 55, no. 33, pp. 92–98, 2022.
- [18] N. Fumo and M. R. Biswas, "Water demand forecasting using machine learning algorithms," *Applied Water Science*, vol. 13, no. 1, pp. 12, 2023.
- [19] Z. Pu, J. Yan, L. Chen, Z. Li, W. Tian, T. Tao, and K. Xin, "A hybrid Wavelet-CNN-LSTM deep learning model for short-term urban water demand forecasting," *Frontiers of Environmental Science & Engineering*, vol. 17, no. 2, pp. 22, 2023.

INFORMATION FOR AUTHORS

The Journal of Information Science and Technology (JIST) aims to be the forum through which researchers, faculties, graduate students and experts of the computer and information technology and other technological related fields share and discuss their high quality research work as well as innovation. Original research articles, practical applications and innovations in the broad area of computer and information technology are suitable for publication in JIST. Periodicity (Publication) is 2 issues per year (JAN - JUN and JULY - DECEMBER).

Accepted papers will be Double-blind peer reviewed by minimum 3 reviewers with "Accept" result.

1. Soft Computing
2. Human-Computer Interaction
3. Information Assurance and Security
4. Information Systems
5. Networking
6. Programming
7. Platform Technologies
8. System Integration and Architecture
9. Social and Professional Issues
10. Web Systems and Technologies
11. Multidisciplinary
12. e-Business and m-Business
13. Business and Information System
14. Social and Business Aspects of Convergence IT and Ubiquitous Computing
15. Internet of Things (IoT)
16. Cloud Computing
17. Fintech and blockchain

Manuscript Preparation Guide

The template for writing the journal is available for downloading (<https://ph02.tci-thaijo.org/index.php/JIST/>). Length of the paper should be in between 8-12 pages of A4 size. Text should be typewritten or printed with double-spaced in 11-point of A4 white paper with margins of 1.5" for top, 1.25" for left, and 1" for bottom and right sides. All pages must be numbered sequentially. Here are some guidelines.

- Abstract which length 100-200 words.
- Introduction which talks about the problem and related research.
- Materials and methods use for experimentation.
- Results of the experiment.
- Discussion and Conclusion, References

Submissions (Publication Charge: Free) (Indexed by TCI tier 2)

Please submit paper in MS Word via <https://ph02.tci-thaijo.org/index.php/JIST/> (Online Journal Submission)

Contact Address

The Association of Council of IT Deans (CITT)
140 Moo 1, Cheumsampan Road, Nongchok
Bangkok, Thailand 10530
E-mail: jist@citt.or.th