ISSN 2085-4552

ULTIMATICS

Jurnal Teknik Informatika

SENTIMENT ANALYSIS OF USER STATISFACTION WITH ACCESS BY KAI APPLICATION USING SUPPORT VECTOR MACHINE AND RANDOM FOREST ALGORITHMS
Muhammad Ilham Nas, Aji Darmawan, Ahmad Awaludin Amri, Kurnia Gusti Ayu (1-8)

SPEECH EMOTION RECOGNITION THROUGH ACOUSTIC DATA ARGUMENTATION AND ATTENTION – DRIVEN CRNN – BIGRU
Muhamad Fitra Kacamarga, Kresna Andika Aprianto (9-16)

EVALUATING THE CAPABILITY OF TRANSFER LEARNING VGG16 FOR DETECTING TOMATO DISEASES IN INDONESIA
Tegar Satriya Wiguna, Febri Liantoni, Yudianto Sujana (17-21)

WEB BASED INFORMATION SYSTEM FOR MONITORING SOIL CONDITIONS (MOISTURE AND TEMPERATURE) INTEGRATED WITH IOT
Ahmad Fazri Anwar, M. Husaini, Khairun Nita Aulia (22-29)

FUZZY EXPERT SYSTEM FOR EARLY HEART DISEASES DIAGNOSIS USING MAMDANI METHOD IN WEB BASED SYSTEM
Ananda Dwi Rizkyta, Muhamad Bahrul Ulum, Fenina Adline Twince Tobing (30-36)

THE IMPLEMENTATION OF AGILE METHODS IN DESIGNING A WEB BASED INFORMATION SYSTEM IN THE GEOMIN LABORATORY OF PT. ANTAM TBK
Dennis Bramastha, Waeisul Bismi (37-44)

MACHINE LEARNING FOR CHILI PAPER PRICE FORECASTING USING EXOGENOUS PUBLIC – ATTENTION SIGNALS AND BAYESIAN HYPERPARAMETER OPTIMIZATION
Wresti Andriani, Gunawan Gunawan, Naella Nabila Putri Wahyuning Naja (45-53)

IMPLEMENTATION OF ARTIFICIAL INTELLIGENCE IN ANEMIA SCREENING FOR ADOLESCENT GIRLS IN PONTIANAK CITY : DEVELOPMENT OF A MECHINE LEARNING BASED EARLY DETECTION SYSTEM
Tilawaty Aprina, Hermanto, Eka Riana, Khoiro Rista Abidin (54-60)

INTEGRATING GIS, IOT AND AI FOR CLIMATE – RESILIENT URBAN FORESTRY : A SMART FRAMEWORK FOR TREE PLANTING STRATEGIES
Julien Nkunduwera Mupenzi, Muhameden El Bousseiry, Adhi Kusnadi, Deden Witarsyah, Aswan Supriyadi Sunge (61-71)

ANDROID – BASED CHILI LEAF DISEASE DETECTION SYSTEM USING DEEP LEARNING FOR HARVEST LOSS MITIGATION)
Wawang Adi Darma, Tia Ernawati, Ridwan, Ariya Fawaz (72-79)

IMPLEMENTATION OF ECDSA AND MLDSA DIGITAL SIGNATURE SCHEMES FOR TRANSACTION AUTHENTICATION ON BLOCKCHAIN
Ridwan Muhammad Raihan, Alam Rahmatulloh (80-90)

AUTOMATED VARED LEXICON LABELING ON SENTIMENT ANALYSIS AGAINST THE E-WALLET
Febri Liantoni, Muhammad Hardiansyah, Elham Syahrian Putra, Erna Piantari (91-97)

DEEP LEARNING BASED ENERGY CONSUMPTION OPTIMIZATION IN SMART HOME SYSTEM AN LSTM APPROACH WITH SHAP INTERPRETABILITY
Mhd. Basri, Krisna Aditya, Andi Zulherry (98-110)

DETECTION OF AI GENERATED VIDEOS USING LOCAL BINARY PATTERN AND SUPPORT VECTOR MACHINE
Ergy David Lundy Tumanggor, Harlen Gilbert Simanullang, Arina Prima Silalahi (111-117)

COMPARATIVE ANALYSIS OF CNN AND MOBILE NET V2 WITH DATA AUGMENTATION AND TRANSFER LEARNING FOR WASTE CLASSIFICATION
Deas Aghellar, Aswan Supriyadi Sunge (118-126)

ANALYSIS OF PUBLIC CONCERNS TOWARDS COVID-19 VACCINES ON X POST PEAK PANDEMIC PHASE
Luthfi Atikah, Novrindah Alvi Hasanah, Binti Kholifah (127-133)

DESISION SUPPORT SYSTEM USING THE WEIGHT PRODUCT (WP) METHOD FOR SELECTING THE BEST EMPLOYEE PERFORMANCE
Budi Permana Putra, Eka Sofiati, Muhammad Harits Pratama, Tomy Nanda Putra (134-140)

Engineering | Vol. 18 | No. 1 | Hal. 1 – 140 | June 2026 | ISSN 2085-4552



UMN

UNIVERSITAS
MULTIMEDIA
NUSANTARA

ISSN 2085-4552

EDITORIAL BOARD

Editor-in-Chief
David Agustriawan, S.Kom., M.Sc., Ph.D.

Managing Editor
Fenina Adline Twince Tobing, S.Kom., M.Kom.
Alexander Waworuntu, S.Kom., M.T.I.
Eunike Endahriana Surbakti, S.Kom., M.T.I
Rena Nainggolan, S.Kom., M.Kom. (UMI)
Nabila Rizky Oktadini, S.SI., M.T. (Unsri)
Rosa Reska Riskiana, S.T., M.T.I.
(Telkom University)
Hastie Audytra, S.Kom., M.T. (Unugiri)

Designer & Layouter
Fenina Adline Twince Tobing, S.Kom., M.Kom.

Members
Rahmat Irsyada, S.Pd., M.Pd. (Unugiri)
Nirma Ceisa Santi (Unugiri)
Shinta Amalia Kusuma Wardhani (Unugiri)
Ariana Tulus Purnomo, Ph.D. (NTUST)
Hani Nurrahmi, S.Kom., M.Kom. (Telkom University)
Aulia Akhrian Syahidi, M.Kom. (Politeknik Negeri Banjarmasin)
Errissya Rasywir, S.Kom, M.T. (Universitas Dinamika Bangsa)
Cian Ramadhona Hassolthine (Universitas Siber Asia)
Wirawan Istiono, S.Kom., M.Kom. (UMN)
Dareen Halim, S.T., M.Sc. (UMN)
Dr. Moeljono Widjaja (UMN)
Wella, S.Kom., M.MSI., COBIT5 (UMN)

EDITORIAL ADDRESS

Universitas Multimedia Nusantara (UMN)
Jl. Scientia Boulevard
Gading Serpong
Tangerang, Banten - 15811
Indonesia
Phone. (021) 5422 0808
Fax. (021) 5422 0800
Email: ultimatics@umn.ac.id



Ultimatics : Jurnal Teknik Informatika is the Journal of the Informatics Study Program at Universitas Multimedia Nusantara which presents scientific research articles in the fields of Computer Science and Informatics, as well as the latest theoretical and practical issues, including Analysis and Design of Algorithm, Software Engineering, System and Network security, Ubiquitous and Mobile Computing, Artificial Intelligence and Machine Learning, Algorithm Theory, World Wide Web, Cryptography, as well as other topics in the field of Informatics. Ultimatics: Jurnal Teknik Informatika is published regularly twice a year (June and December) and is managed by the Informatics Study Program at Universitas Multimedia Nusantara.

Call for Papers



Ultimatics: Jurnal Teknik Informatika is the Journal of the Informatics Study Program at Universitas Multimedia Nusantara which presents scientific research articles in the fields of Analysis and Design of Algorithm, Software Engineering, System and Network security, as well as the latest theoretical and practical issues, including Ubiquitous and Mobile Computing, Artificial Intelligence and Machine Learning, Algorithm Theory, World Wide Web, Cryptography, as well as other topics in the field of Informatics. ULTIMATICS is published twice a year by Faculty of Engineering and Informatics of Universitas Multimedia Nusantara in cooperation with UMN Press.



International Journal of New Media Technology (IJNMT) is a scholarly open access, peer-reviewed, and interdisciplinary journal focusing on theories, methods and implementations of new media technology. Topics include, but not limited to digital technology for creative industry, infrastructure technology, computing communication and networking, signal and image processing, intelligent system, control and embedded system, mobile and web-based system, and robotics.



Ultima Computing: Jurnal Sistem Komputer is a Journal of Computer Engineering Study Program, Universitas Multimedia Nusantara which presents scientific research articles in the field of Computer Engineering and Electrical Engineering as well as current theoretical and practical issues, including Edge Computing, Internet-of-Things, Embedded Systems, Robotics, Control System, Network and Communication, System Integration, as well as other topics in the field of Computer Engineering and Electrical Engineering.



Ultima InfoSys: Jurnal Ilmu Sistem Informasi is a Journal of Information Systems Study Program at Universitas Multimedia Nusantara which presents scientific research articles in the field of Information Systems, as well as the latest theoretical and practical issues, including database systems, management information systems, system analysis and development, system project management information, programming, mobile information system, and other topics related to Information Systems.

ISSN 2085-4552

FOREWORD

ULTIMA Greetings!

Ultimatics: Jurnal Teknik Informatika is the Journal of the Informatics Study Program at Universitas Multimedia Nusantara which presents scientific research articles in the fields of Computer Science and Informatics, as well as the latest theoretical and practical issues, including Analysis and Design of Algorithm, Software Engineering, System and Network Security, Ubiquitous and Mobile Computing, Artificial Intelligence and Machine Learning, Algorithm Theory, World Wide Web, Cryptography, as well as other topics in the field of Informatics. Ultimatics: Jurnal Teknik Informatika is published regularly twice a year (June and December) and is published by the Faculty of Engineering and Informatics at Universitas Multimedia Nusantara.

In this June 2026 edition, Ultimatics enters the 1st Edition of Volume 18. In this edition there are seventeen scientific papers from researchers, academics and practitioners in the fields of Computer Science and Informatics. Some of the topics raised in this journal are: Sentiment Analysis of User Satisfaction with Access by KAI Application Using Support Vector Machine and Random Forest Algorithms, Speech Emotion Recognition through Acoustic Data Augmentation and Attention-Driven CRNN-BiGRU, Evaluating the Capability of Transfer Learning VGG16 for Detecting Tomato Diseases in Indonesia, Web-Based Information System for Monitoring Soil Conditions (Moisture and Temperature) Integrated with IoT, Fuzzy Expert System for Early Heart Disease Diagnosis Using Mamdani Method in Web-Based System, Implementation Of Agile Methods in Designing A Web-Based Information System in the Geomin Laboratory of PT. Antam Tbk, Machine Learning for Chili Pepper Price Forecasting Using Exogenous Public-Attention Signals and Bayesian Hyperparameter Optimization, Implementation of Artificial Intelligence in Anemia Screening for Adolescent Girls in Pontianak City, Development of a Machine Learning-Based Early Detection System, Integrating GIS, IOT and AI for Climate-Resilient Urban Forestry: a Smart Framework for Tree Planting Strategies, Android-Based Chili Leaf Disease Detection System Using Deep Learning for Harvest Loss Mitigation, Implementation of ECDSA and MLDSA Digital Signature Schemes for Transaction Authentication on Blockchain, Automated Vader Lexicon Labeling on Sentiment Analysis Against the E-Wallet, Deep Learning-Based Energy Consumption Optimization in Smart Home Systems: an LSTM Approach with SHAP Interpretability, Detection of AI-Generated Videos using Local Binary Pattern and Support Vector Machine, Comparative Analysis of CNN and MobileNetV2 with Data Augmentation and Transfer Learning for Waste Classification, Analysis of Public Concerns Towards COVID-19 Vaccines on X Post-Peak Pandemic Phase, Decision Support System Using the Weight Product (WP) Method for Selecting the Best Employee Performance.

On this occasion we would also like to invite the participation of our dear readers, researchers, academics, and practitioners, in the field of Engineering and Informatics, to submit quality scientific papers to: Ultimatics: Jurnal Teknik Informatics, International Journal of New Media Technology (IJNMT), Ultima Infosys: Journal of Information Systems and Ultima Computing: Journal of Computer Systems. Information regarding writing guidelines and templates, as well as other related information can be obtained through the email address ultimatics@umn.ac.id and the webpage of our Journal here.

Finally, we would like to thank all contributors to this June 2026 Edition of Ultimatics. We hope that scientific articles from research in this journal can be useful and contribute to the development of research and science in Indonesia.

June 2026,

David Agustriawan, S.Kom., M.Sc., Ph.D.
Editor-in-Chief

ISSN 2085-4552

DAFTAR ISI

No	Title	Authors	Pages
1	Sentiment Analysis of User Satisfaction with Access by KAI Application Using Support Vector Machine and Random Forest Algorithms	Muhammad Ilham Nas, Aji Darmawan, Ahmad Awaludin Amri, Kurnia Gusti Ayu	1-8
2	Speech Emotion Recognition through Acoustic Data Augmentation and Attention-Driven CRNN-BiGRU	Muhamad Fitra Kacamarga, Kresna Andika Aprianto	9-16
3	Evaluating the Capability of Transfer Learning VGG16 for Detecting Tomato Diseases in Indonesia	Tegar Satriya Wiguna, Febri Liantoni, Yudianto Sujana	17-21
4	Web-Based Information System for Monitoring Soil Conditions (Moisture and Temperature) Integrated with IoT	Ahmad Farzi Anwar, M. Husaini, Khairun Nita Aulia	22-29
5	Fuzzy Expert System for Early Heart Disease Diagnosis Using Mamdani Method in Web-Based System	Ananda Dwi Rizkyta, Muhamad Bahrul Ulum, Fenina Adline Twince Tobing	30-36
6	Implementation Of Agile Methods in Designing A Web-Based Information System in the Geomin Laboratory of PT. Antam Tbk	Dennis Bhayu Bramastha, Waeisul Bismi	37-44
7	Machine Learning for Chili Pepper Price Forecasting Using Exogenous Public-Attention Signals and Bayesian Hyperparameter Optimization	Wresti Andriani, Gunawan Gunawan, Naella Nabila Putri Wahyuning Naja	45-53
8	Implementation of Artificial Intelligence in Anemia Screening for Adolescent Girls in Pontianak City: Development of a Machine Learning-Based Early Detection System	Tilawaty Aprina, Hermanto, Eka Riana, Khoiro Rista Abidin	54-60
9	Integrating GIS, IOT and AI for Climate-Resilient Urban Forestry: a Smart Framework for Tree Planting Strategies	Julien Nkunduwera Mupenzi, Mohameden El Bousseiry, Adhi Kusnadi, Deden Witarsyah, Aswan Supriyadi Sunge	61-71
10	Android-Based Chili Leaf Disease Detection System Using Deep Learning for Harvest Loss Mitigation	Wawang Adi Darma, Tia Ernawati, Ridwan, Ariya Fawaz	72-79
11	Implementation of ECDSA and MLDSA Digital Signature Schemes for Transaction Authentication on Blockchain	Ridwan Muhammad Raihan, Alam Rahmatulloh	80-90
12	Automated Vader Lexicon Labeling on Sentiment Analysis Against the E-Wallet	Febri Liantoni, Muhammad Hardiansyah, Elham Syahrian Putra, Erna Piantari	91-97
13	Deep Learning-Based Energy Consumption Optimization in Smart Home Systems: an LSTM Approach with SHAP Interpretability	Mhd. Basri, Krisna Aditya, Andi Zulherry	98-110
14	Detection of AI-Generated Videos using Local Binary Pattern and Support Vector Machine	Ergy David Lundy Tumanggor, Harlen Gilbert Simanullang, Arina Prima Silalahi	111-117

15	Comparative Analysis of CNN and MobileNetV2 with Data Augmentation and Transfer Learning for Waste Classification	Deas Aghellar, Aswan Supriyadi Sunge	118-126
16	Analysis of Public Concerns Towards COVID-19 Vaccines on X Post-Peak Pandemic Phase	Luthfi Atikah, Novrindah Alvi Hasanah, Binti Kholifah	127-133
17	Decision Support System Using the Weight Product (WP) Method for Selecting the Best Employee Performance	Budi Permana Putra, Eka Sofiati, Muhammad Harits Pratama, Tomy Nanda Putra	134-140



ISSN 2085-4552

Sentiment Analysis of User Satisfaction with Access by KAI Application Using Support Vector Machine and Random Forest Algorithms

Muhammad Ilham Nas[1], Aji Darmawan[2], Ahmad Awaludin Amri[3], Kurnia Gusti Ayu[4]
[1,2,3,4] Faculty of Computer Science, Mercu Buana University, Jakarta, Indonesia
ilhamnas29@gmail.com[1], ajidarmawan.id@gmail.com[2],
ahmadawaludinamri18@gmail.com[3], kurnia.gusti@mercubuana.ac.id[4]

Accepted 19 June 2026
Approved 28 June 2026

***Abstract*— PT Kereta Api Indonesia (Persero), a State-Owned Enterprise (SOE) in the railway sector, holds an important responsibility in providing efficient services. To enhance their services, they introduced the application "Access by KAI." Although this application represents a technological innovation, it often falls short of users' expectations due to existing shortcomings. Therefore, this study proposes the use of sentiment analysis to gauge users' perspectives on the "Access by KAI" application. By applying Support Vector Machine (SVM) and Random Forest algorithms to user review datasets, this study classifies user sentiments into three categories: positive, neutral and negative, and compares the performance of these two algorithms. The results of this study offer valuable insights into users' attitudes towards the application, which can assist PT Kereta Api Indonesia in improving service quality. The classification results using the Random Forest and Support Vector Machine (SVM) algorithms, based on the factors used in this study, show varying performance. Random Forest achieved an accuracy of 86% for an 80:20 and 90:10 data ratios, and 85% for a 70:30 ratio, while SVM achieved 83% accuracy across different data ratios.**

***Index Terms*— Access by KAI; Sentiment Analysis; Support Vector Machine; Random Forest.**

I. INTRODUCTION

The development of information and communication technology plays an important role in supporting daily activities, including transportation. PT Kereta Api Indonesia (Persero), the only State-Owned Enterprise (SOE) under the Ministry of Transportation operating in the railway sector, introduced Access by KAI to facilitate ticket booking and related service[1]. On July 7, 2023, the application, formerly known as KAI Access, was updated with features such as ticket booking, rescheduling, cancellation, and e-boarding passes. These features are expected to improve the train travel experience. As one of Indonesia's preferred public transportation modes, trains are widely used. Transactions through KAI Access reached 9.1 million (61.77%), followed by B2B partners with 4 million transactions (27.10%), counters with 1.2 million transactions (8.47%), the KAI website with 374 thousand transactions (2.52%), vending machines with 14 thousand transactions (0.10%), and Contact Center 121 with 6 thousand transactions (0.04%).

Access by KAI is available on both the App Store and Google Play Store. However, the application has not fully met user expectations because several limitations still need to be improved. User reviews are valuable for application development, yet the large number and variety of opinions make systematic evaluation difficult. Alawaji and Aloraini stated that sentiment analysis of digital service application reviews can be used to assess user satisfaction and provide insights for service improvement strategies[2]. Currently, there is no standardized method for classifying Access by KAI user reviews into positive, neutral, or negative sentiment. Therefore, sentiment analysis is proposed as an approach to evaluate user opinions on the Access by KAI application[3]. The neutral class is important because it may represent users who do not express strong dissatisfaction but still show uncertainty, hesitation, or moderate experience with the application. In the context of user satisfaction, neutral reviews may serve as early indicators of potential service issues before they develop into negative perceptions. Therefore, identifying neutral sentiment can assist PT KAI in designing preventive service improvement strategies. Sentiment analysis is a method used to extract opinions, attitudes, and sentiments about a topic, product, or service from unstructured text data[4]. In this study, sentiment classification was conducted using classification-based machine learning methods to systematically categorize user opinions toward the Access by KAI application.

Previous studies have examined sentiment analysis and classification algorithm comparison. Rahayu and

Rahmat compared Naïve Bayes and SVM on Spotify user reviews from the Play Store, with SVM achieving 84% accuracy and Naïve Bayes achieving 86.4% [5]. Another study by Wandani, Fauziah, and Andrianingsih compared K-NN, Random Forest, and Naïve Bayes for sentiment analysis of Twitter users on flash sale events using "flash sale" and "Shopee flash sale" datasets. Their results showed accuracies of 83.53% and 81.48% for Naïve Bayes, 82.94% and 77.78% for K-NN, and 80.59% and 74.07% for Random Forest, respectively [6]. Lustiansyah, Widiyanto, and Wahyono applied Long Short-Term Memory (LSTM) to analyze sentiment toward the PeduliLindungi application based on Play Store reviews, achieving 82.44% accuracy, 78.66% precision, and 87.03% recall [7].

Based on these studies, this research compares SVM and Random Forest in analyzing sentiment toward the Access by KAI application. The dataset was collected from the Google Play Store using a web scraping technique. Unlike previous research that focused on aspect-based sentiment analysis of KAI Access reviews, this study compares SVM and Random Forest across three data-splitting scenarios and evaluates their performance at the class level for positive, neutral, and negative sentiments. This study also emphasizes the neutral class as an early indicator of potential dissatisfaction and provides practical insights for improving the Access by KAI application.

II. METHOD

This study follows a series of stages required to analyze the sentiment of Access by KAI reviews, as illustrated in the flowchart in Figure 1 below.

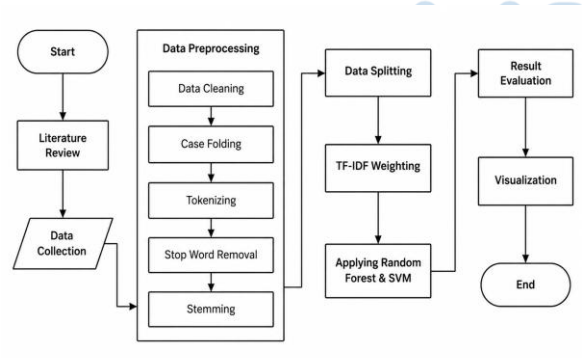


Fig. 1. Research Flowchart

A. Data Scraping

Data scraping, or web scraping, is a method for retrieving data from online sources, particularly from web pages that use markup languages such as HTML or XHTML. This process involves parsing documents to obtain the desired data, regardless of its size [8]. Web scraping automates the retrieval and extraction of data from websites, eliminating the need for manual copying[9].

B. Data Preprocessing

The data preprocessing stage is carried out to convert raw data into a higher-quality form, ensuring it can be effectively processed in subsequent steps[10]. This step addresses problematic or inconsistent data, such as data containing noise or error[11][12]. The preprocessing stage consists of five main steps: data cleaning, case folding, tokenization, stop word removal, and stemming.

1) Data Cleaning

Data cleaning involves preparing text documents by eliminating or correcting irrelevant elements, such as numbers, URLs, usernames, hashtags, and other nonessential components. The goal is to identify and fix inaccurate or incomplete data[13].

2) Case Folding

Case Folding involves converting all uppercase letters in a document to lowercase to ensure uniformity. For example, "PT Astra" is transformed to "pt astra," so that both uppercase and lowercase letters are treated the same [14]

3) Tokenization

Tokenization is a method by segmenting textual data into smaller units, known as tokens, to support individual word-level analysis.

4) Stop Word Removal

At this stage, commonly used words, including "what," "who," "and," and "to," are removed because they have limited relevance to sentiment analysis and may introduce noise into the dataset [15].

5) Stemming

Stemming aims to convert words into their base forms, thereby simplifying the text analysis process[16]. For Indonesian datasets, the Sastrawi library in the Python programming language is employed for this purpose. For instance, "*makanan*" is stemmed to "*makan*".

C. Data Splitting

In machine learning, dataset partitioning is a necessary step to support model development and evaluation. The dataset is separated into two subsets: one for training the model and another for testing its ability to make predictions on data that were not used during training. The split ratio is important because an unsuitable proportion may negatively affect the model's precision and accuracy[17].

D. Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF represents textual data as weighted numerical features, making it suitable for classification models such as SVM and Random Forest that require numerical input[16]. This technique also assigns greater weight to terms that are more informative,

thereby supporting improved model performance in the classification stage.

The formula for calculating TF-IDF is:

$$tf_{td}\, id_{ft} = tf_{td} * \log\left(\frac{N}{df_t}\right) \quad (1)$$

Where:

$tftd\ \ idft$: The weighted score of a word in a document.

$tftd$: The frequency of term *t* in document *d*.

N: Refers to the total documents in the dataset.

dft: Number of documents that contain word *t*.

A high TF-IDF value reflects a term that is frequent in a specific document but uncommon in the overall corpus. In contrast, a low TF-IDF value indicates that the term appears widely across documents, reducing its distinctiveness in the analysis.

E. Support Vector Machine

Support Vector Machine is a machine learning method that applies the Structural Risk Minimization (SRM) principle. This algorithm aims to identify the most optimal hyperplane for separating two classes within the input space[18]. In classification, the hyperplane is chosen to divide the classes by maximizing the margin, defined as the distance between the hyperplane and the nearest data points from each class. SVM is particularly effective in handling data that may not have a linear relationship in high-dimensional spaces[19]. By employing kernel functions, SVM can implement non-linear and high-dimensional input separators, making it a versatile method[20][21]. In sentiment analysis, SVM is used to classify texts as positive or negative. The following is the SVM formula:

$$f(x) = w.x + b \quad (2)$$

Explanation of the formula:

f = The target value

(x) = Vector containing the input data

w = Vector of weight values

b = Bias term

F. Random Forest

Random Forest is an ensemble-based machine learning algorithm that integrates multiple decision trees to perform classification tasks[22]. The final prediction is generated through a majority voting mechanism. This method offers several advantages, including low error rates, strong performance on large datasets, and robustness against overfitting due to its ensemble learning approach[23]. Moreover, Random Forest can be used for both classification and regression tasks and is able to identify the most influential features in the training data, making it useful for feature selection [24].

$$l(y) = argmax_c\ (\sum I_{\text{h}}\text{n}(y)=c) \quad (3)$$

G. Confusion Matrix

A confusion matrix is used to evaluate a model's performance in classifying the target classes. It summarizes the model's ability to classify data correctly and highlights any errors that occur during the process[25].

TABLE I. CONFUSION MATRIX

Actual Class	Predicted Class	
	Positive	**Negative**
Positive	*True Positive* (TP)	*False Negative* (FN)
Negative	*False Positive* (FP)	*True Negative* (TN)
Positive	*True Positive* (TP)	*False Negative* (FN)

A confusion matrix contains four key components used to evaluate the performance of a classification model:

1. True Positive (TP) refers to positive instances that are correctly classified as positive
2. True Negative (TN represents negative instances that are correctly classified as negative
3. False Positive (FP) occurs when negative instances are incorrectly predicted as positive
4. False Negative (FN) occurs when positive instances are incorrectly predicted as negative.

The confusion matrix can also be used to derive several evaluation metrics:

- Accuracy measures the proportion of correct predictions made by the model, both positive and negative, out of the total predictions. The formula for accuracy is[26]:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FP+FN+TN}} \times 100\% \quad (4)$$

- Precision measures how well the model identifies true positives out of all the predicted positive results. The formula for precision is[27] :

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \times 100\% \quad (5)$$

- Recall (or Sensitivity) measures the model's effectiveness in identifying true positive cases from all actual positives. The formula for recall is [28]:

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \times 100\% \quad (6)$$

- F1 Score is the harmonic mean of precision and recall. It is useful when the dataset has an imbalance between false negatives and false

positives. The formula for the F1 score is [29]:

$$\text{F1 Score} = \frac{2 \text{ x recall x precision}}{\text{recall+precision}} \times 100\% \quad (7)$$

SVM and Random Forest were selected because both methods have strong characteristics for text classification. SVM is effective for high-dimensional TF-IDF features because it maximizes the margin between classes, while Random Forest improves classification stability through ensemble learning and reduces overfitting compared with a single decision tree. Therefore, comparing these two methods provides a more comprehensive understanding of their suitability for sentiment classification in application review data.

III. RESULT AND DISCUSSIONS

A. Data Scraping

Review data were collected from the Access by KAI application on the Google Play Store using the google-play-scraper library. The scraping process retrieved 7,341 review records consisting of four attributes: userName, score, date, and content. A sample of the scraped review data is presented in Table 2.

TABLE II. SAMPLE OF DATA SCRAPING RESULTS FROM GOOGLE PLAY STORE REVIEWS

No.	User Name	Rating	Date	Review Content
1	ahmad fajri	1	2024-04-25 13:20:57	Reduksi tidak bisa dipakai pada aplikasi ini, ...
2	Fyanhie 03	1	2024-04-25 13:03:29	Mau daftar aja susah gimana si
3	Gilang Rizki	5	2024-04-25 12:56:27	Bagus. Aplikasi sangat baik, Bintang 5. Semoga...
4	Marmut Merah jambu	1	2024-04-25 12:55:13	Daftar member MAPCLUB ngga bisa ngisi provinsi...
5	wong mendem	1	2024-04-25 12:18:58	Tiket KA jarak jauh tidak bisa dibatalkan, tida...

B. Data Preprocessing

Data preprocessing was conducted to transform raw review data into a cleaner and more structured format for sentiment classification. This stage included sentiment labeling, data cleaning, case folding, tokenization, stop word removal, and stemming. Sentiment labeling was performed based on the rating scores obtained from Google Play Store reviews. Ratings higher than 3 were categorized as positive sentiment (1), ratings equal to 3 were categorized as neutral sentiment (0), and ratings lower than 3 were categorized as negative sentiment (-1). The threshold value of 3 was used because it represents the midpoint of the 1–5 rating scale. Therefore, a rating of 3 indicates a neutral user evaluation, while ratings above and below this value indicate positive and negative evaluations, respectively.

After the labeling process, text preprocessing was applied to reduce noise and improve data quality. Data cleaning was performed to remove URLs, numbers, punctuation marks, symbols, and irrelevant characters. Case folding was used to convert all text into lowercase. Tokenization was applied to split the text into individual terms. Stop word removal was conducted to eliminate common words that do not provide significant semantic contribution to sentiment classification. Finally, stemming was applied to convert Indonesian words into their root forms. Table 3 presents examples of review text before and after preprocessing. The results show that punctuation marks, irrelevant characters, and less meaningful terms were removed, while the remaining words were converted into a cleaner textual form for sentiment classification.

TABLE III. EXAMPLE OF REVIEW TEXT BEFORE AND AFTER PREPROCESSING

No.	Before Preprocessing	After Preprocessing
1	Reduksi tidak bisa dipakai pada aplikasi ini, ...	reduksi tidak bisa pakai aplikasi
2	Mau daftar aja susah gimana si	daftar susah gimana
3	Bagus. Aplikasi sangat baik, Bintang 5. Semoga...	bagus aplikasi sangat baik bintang semoga
4	Daftar member MAPCLUB ngga bisa ngisi provinsi...	daftar member mapclub ngga bisa ngisi provinsi
5	Tiket KA jarak jauh tidak bisa dibatalkan, tida...	tiket ka jarak jauh tidak bisa batal

C. Data Splitting

The dataset was divided into training and testing sets using three data-splitting scenarios: 90:10, 80:20, and 70:30. The training set was used to build the classification models, while the testing set was used to evaluate model performance on unseen data. These scenarios were applied to compare the consistency of the SVM and Random Forest models across different proportions of training and testing data. The number of training and testing data for each scenario is presented in Table 4.

TABLE IV. NUMBER OF TRAINING AND TESTING DATA IN EACH DATA SPLIT

Data Split	Training Data	Testing Data
90:10	6,606	735
80:20	5,872	1,469
70:30	5,138	2,203

D. TF-IDF Weighting

TF-IDF weighting was applied to convert the preprocessed review text into numerical features that could be processed by machine learning algorithms. Term Frequency (TF) represents the frequency of a term in a document, while Inverse Document Frequency (IDF) indicates the importance of a term across the entire document collection. Words that

frequently appear in a specific review but rarely appear in other reviews receive higher weights. The resulting TF-IDF feature matrix was used as input for the SVM and Random Forest classification models.

E. Model Application

The TF-IDF feature matrix was used as input for the SVM and Random Forest models. Both models were trained using the training set and evaluated using the testing set under three data-splitting scenarios: 90:10, 80:20, and 70:30. Model performance was assessed using accuracy, precision, recall, and F1-score for each sentiment class.

1) Support Vector Machine

The SVM model was applied to classify user reviews into positive, neutral, and negative sentiment classes. SVM was selected because it is effective for high-dimensional text data represented by TF-IDF features. The model works by finding an optimal separating hyperplane that maximizes the margin between sentiment classes. In this study, the SVM model achieved consistent accuracy of 83% across all data-splitting scenarios.

2) Random Forest

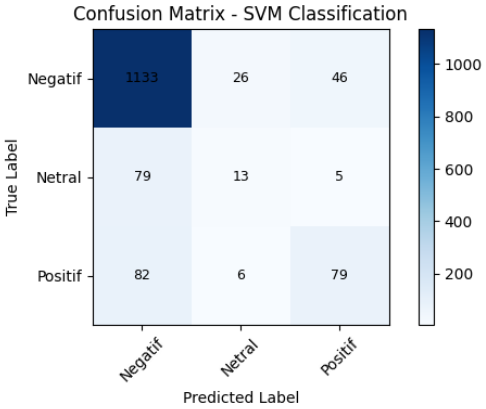
The Random Forest model was applied using the same TF-IDF features and data-splitting scenarios. Random Forest was selected because it applies ensemble learning by combining multiple decision trees, which improves classification stability and reduces the risk of overfitting compared with a single decision tree. In this study, Random Forest achieved 86% accuracy in the 90:10 and 80:20 splits and 85% accuracy in the 70:30 split.

F. Evaluation (Confusion Matrix)

Model evaluation was performed using the confusion matrix and classification report to assess accuracy, precision, recall, and F1-score for each sentiment class. For readability, only the confusion matrix results for the 80:20 data split are presented as the representative scenario, while the complete performance comparison across all data splits is summarized in Table 5.

1) Support Vector Machine

Figure 2 presents the SVM confusion matrix for the 80:20 data split. Based on the classification report, the model achieved an overall accuracy of 83%. The model performed well in detecting negative sentiment, with 88% precision, 94% recall, and 91% F1-score. However, the neutral class showed lower performance, with 29% precision, 13% recall, and 18% F1-score. For the positive class, the model achieved 61% precision, 47% recall, and 53% F1-score.

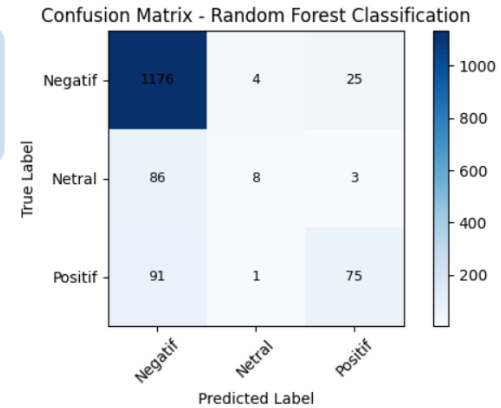


	precision	recall	f1-score	support
-1	0.88	0.94	0.91	1205
0	0.29	0.13	0.18	97
1	0.61	0.47	0.53	167
accuracy			0.83	1469
macro avg	0.59	0.52	0.54	1469
weighted avg	0.81	0.83	0.82	1469

Fig. 2. SVM Confusion Matrix Results on Test Data (20%)

2) Random Forest

Figure 3 presents the Random Forest confusion matrix for the 80:20 data split. Based on the classification report, the model achieved an overall accuracy of 86%. The model also performed strongly in detecting negative sentiment, with 87% precision, 98% recall, and 92% F1-score. However, the neutral class remained difficult to classify, with 62% precision, 8% recall, and 15% F1-score. For the positive class, the model achieved 73% precision, 45% recall, and 56% F1-score.



	precision	recall	f1-score	support
-1	0.87	0.98	0.92	1205
0	0.62	0.08	0.15	97
1	0.73	0.45	0.56	167
accuracy			0.86	1469
macro avg	0.74	0.50	0.54	1469
weighted avg	0.84	0.86	0.83	1469

Fig 3. Random Forest Confusion Matrix Results on Test Data.

IV. DISCUSSION

A. Discussion of Model Performance Results

Table 5 summarizes the performance of SVM and Random Forest across three data-splitting scenarios: 90:10, 80:20, and 70:30. The confusion matrix results are presented in a summarized form to improve readability and reduce redundancy. Overall, Random Forest achieved higher accuracy than SVM, with 86% accuracy for the 90:10 and 80:20 splits and 85% for the 70:30 split, while SVM maintained 83% accuracy across all splits. Both models performed well in detecting negative sentiment. SVM achieved recall values between 94% and 95% for the negative class, while Random Forest achieved higher recall values between 97% and 98%. However, the neutral class showed the weakest performance, with recall values below 15% in all scenarios. This indicates that many neutral reviews were misclassified as either positive or negative.

The low recall of the neutral class may be caused by class imbalance, as negative reviews dominated the dataset, while neutral reviews were relatively limited. Neutral reviews may also contain ambiguous expressions that overlap with other sentiment classes. In addition, TF-IDF focuses on word frequency rather than contextual meaning, which may limit its ability to capture subtle neutral expressions. Overall, Random Forest outperformed SVM in overall accuracy and negative sentiment detection. However, both models still require improvement in detecting neutral sentiment. Future studies may apply SMOTE or class weighting to improve minority-class detection.

TABLE V. SUMMARY OF MODEL PERFORMANCE ACROSS SPLIT DATA

Model	Split	Acc.	Negative P/R/F1	Neutral P/R/F1	Positive P/R/F1
SVM	90:10	83%	87/95/91	42/15/23	54/38/45
SVM	80:20	83%	88/94/91	29/13/18	61/47/53
SVM	70:30	83%	87/94/90	17/8/11	59/41/48
Random Forest	90:10	86%	88/97/92	71/10/17	64/46/54
Random Forest	80:20	86%	87/98/92	62/8/15	73/45/56
Random Forest	70:30	85%	87/97/92	36/6/11	73/42/53

B. Word Visualization Using Word Cloud and Bar Chart

In this study, visualizing the data content is essential for gaining insights. Visualization is a powerful tool for analyzing sentiment and ratings. Bar charts are useful for visualizing sentiment distribution, while word clouds are ideal for understanding text data. Figure 4 shows a visualization of sentiment data classified based on user satisfaction with the Access by KAI application. The data reveals that 6,043 users expressed negative sentiments, 809 users expressed positive sentiments, and 489 expressed neutral sentiments. This indicates that the majority of users are dissatisfied with the service provided by the Access by KAI application.

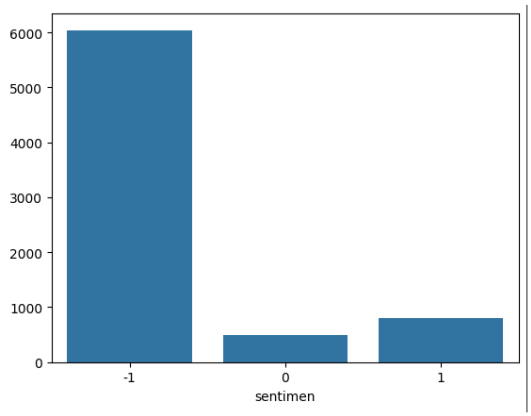


Fig 4. Sentiment Data Visualization

To better understand the content of the review text, word clouds were created for each sentiment class. Three word clouds are displayed:

1) Positive Sentiment Word Cloud

In Figure 5, positive sentiments in the Access by KAI application reviews are visualized. Frequently appearing words such as “application”, “ticket”, “very”, “easy”, “train”, “good”, “help”, “new“, and “more” highlight some of the focal points of positive user feedback.

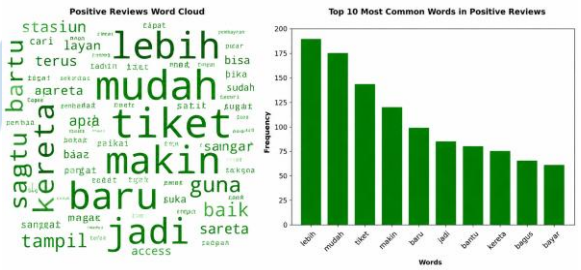


Fig 5. Word Cloud and Bar Chart for Positive Sentiment

2) Neutral Sentiment Word Cloud

Figure 6 presents the word cloud and bar chart for neutral sentiments in Access by KAI reviews. Common words such as "ticket," "application," "old," "more," "new," "good," "pay," "update," and "train" indicate the primary themes within neutral reviews.



Fig 6. Word Cloud and Bar Chart for Neutral Sentiment

3) Negative Sentiment Word Cloud

Figure 7 visualizes negative sentiments in Access by KAI reviews. Frequently used words like "application," "ticket," "old," "new," "update," "train," and "good" reflect the predominant issues raised by users in their negative feedback. The word cloud and bar chart provide a visual summary of the main words associated with negative sentiments in the app reviews.

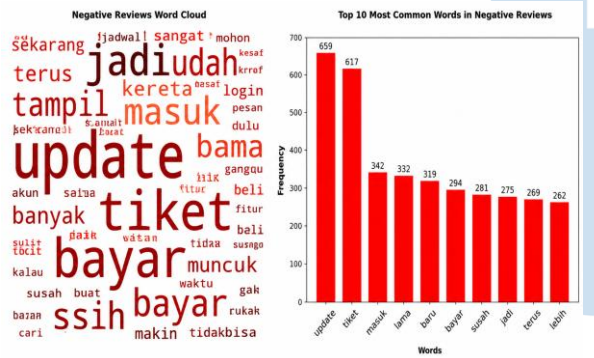


Fig 7. Word Cloud and Bar Chart for Negative Sentiment

V. CONCLUSIONS

Classification using SVM and Random Forest showed that Random Forest achieved higher accuracy, with 86% in the 90:10 and 80:20 splits and 85% in the 70:30 split, while SVM consistently achieved 83% across all splits. Both models performed well in detecting negative sentiment but showed weak performance in identifying neutral sentiment, indicating that many neutral reviews were misclassified as positive or negative. This issue may be addressed in future studies by applying SMOTE or class weighting to improve minority-class detection. The sentiment distribution showed dominant user dissatisfaction, with 6,043 negative reviews compared to 809 positive and 489 neutral reviews. Frequently occurring words indicate that user concerns were mainly related to application usability, ticketing services, payment processes, application updates, and the transition from the old version to the new version. Therefore, PT KAI should prioritize improving payment reliability, simplifying ticket booking, cancellation, and rescheduling procedures, enhancing application stability after updates, and monitoring neutral reviews as early indicators of potential dissatisfaction.

REFERENCES

[1] N. B. Sidauruk and N. Riza, "SENTIMEN ANALISIS DATA PENGGUNA TERHADAP KAI ACCESS Systematic Literature Review," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 2, pp. 1297–1303, 2023.

[2] R. Alawaji and A. Aloraini, "Sentiment Analysis of Digital Banking Reviews Using Machine Learning and Large Language Models," *Electron. 2025,* vol. 14, 2025, doi: https://doi.org/ 10.3390/electronics14112125 Copyright:

[3] G. Radiena *et al.*, "APLIKASI KAI ACCESS MENGGUNAKAN METODE SUPPORT VECTOR MACHINE," *J. Pendidik. Teknol. Inf. e-ISSN*, no. April, pp. 1–10, 2023.

[4] A. Dwiki *et al.*, "Analisis Sentimen Pada Ulasan Pengguna Aplikasi Bibit Dan Bareksa Dengan Algoritma KNN," *J. Tek. Inform. dan Sist. Inf.*, vol. 8, no. 2, pp. 636–646, 2021.

[5] A. S. Rahayu and A. Fauzi, "Komparasi Algoritma Naïve Bayes Dan Support Vector Machine (SVM) Pada Analisis Sentimen Spotify," *J. Sist. Komput. dan Inform.*, vol. 4, pp. 349–354, 2022, doi: 10.30865/json.v4i2.5398.

[6] A. Wandani, "Sentimen Analisis Pengguna Twitter pada Event Flash Sale Menggunakan Algoritma K-NN , Random Forest , dan Naive Bayes," *J. Sains Komput. Inform.*, vol. 5, no. September, pp. 651–665, 2021.

[7] G. A. Lustiansyah *et al.*, "Analisis klasifikasi sentimen pengguna aplikasi pedulilindungi berdasarkan ulasan dengan menggunakan metode long short term memory," *Jakarta-Indonesia, 20 Agustus 2022*, pp. 327–336, 2022.

[8] R. Sistem, M. Hutomi, and M. Qamal, "Sistem Pengecekan Toko Online Asli atau Dropship pada Shopee," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 1, no. 10, pp. 1117–1123, 2021.

[9] S. Satriajati and S. Panuntun, Satria Bagus; Pramana, "IMPLEMENTASI WEB SCRAPING DALAM PENGUMPULAN BERITA KRIMINAL PADA MASA PANDEMI COVID-19 (Implementation of Web Scraping in Criminal News Collection during Covid-19 Pandemic)," *Semin. Nas. Off. Stat. 2020 Stat. New Norm. A Chall. Big Data Off. Stat. IMPLEMENTASI*, pp. 300–308, 2020.

[10] S. Dwiasnati and Y. Devianto, "Classification of forest fire areas using machine learning algorithm," *World J. Adv. Eng. Tehnol. Sci.*, vol. 03, no. May, pp. 8–15, 2021.

[11] M. A. Riwanto, "ANALISIS DATA KUNJUNGAN WISATAWAN MANCANEGARA KE INDONESIA MENGGUNAKAN MICROSOFT POWER BI M . Akbar Riwanto Sistem Informasi , Fakultas Teknik dan Ilmu Komputer , Universitas Islam Indragiri Email : akbarrwnt@gmail.com Abstrak Pendahuluan Metode," *J. Sist. Inf.*, vol. 2, no. 3, pp. 112–119, 2024.

[12] H. T. Duong, T. Anh, and N. Thi, "A review : preprocessing techniques and data augmentation for sentiment analysis," *Comput. Soc. Networks*, pp. 1–16, 2021, doi: 10.1186/s40649-020-00080-x.

[13] J. A. Septian, T. M. Fahrudin, and A. Nugroho, "Analisis Sentimen Pengguna Twitter Terhadap Polemik Persepakbolaan Indonesia Menggunakan Pembobotan TF - IDF dan K - Nearest Neighbor," *J. Intell. Syst. Comput.*, pp. 43–49, 2020.

[14] L. Merinda, A. Abdurrahim, and L. Syafa, "Analisis Sentimen Tweet Vaksin COVID-19 Menggunakan Recurrent," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 10, pp. 802–808, 2021.

[15] J. David, K. V. I. Saputra, A. M. Panjaitan, F. Victor, and P. Samosir, "Sentiment Analysis of University X Students : Comparing Naive Bayes and BERT Approaches," *Ultim. J.*

Tek. Inform., vol. 17, no. 2, 2026.

[16] D. Musfiroh, U. Khaira, P. Eko, P. Utomo, and T. Suratno, "Sentiment Analysis of Online Lectures in Indonesia from Twitter Dataset Using InSet Lexicon Analisis Sentimen terhadap Perkuliahan Daring di Indonesia dari Twitter Dataset Menggunakan InSet Lexicon," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 1, no. April, pp. 24–33, 2021.

[17] V. R. Joseph, "Optimal ratio for data splitting," *Wiley*, no. April, pp. 531–538, 2022, doi: 10.1002/sam.11583.

[18] B. Mahesh, "Machine Learning Algorithms - A Review," *Int. J. Sci. Res.*, vol. 9, no. 1, pp. 381–386, 2020, doi: 10.21275/ART20203995.

[19] P. C. M. F, "Confidence intervals for the random forest generalization error," *Pattern Recognit. Lett.*, vol. 158, pp. 171–175, 2022, doi: 10.1016/j.patrec.2022.04.031.

[20] B. W. Rauf, "Sentimen Analisis Pertambangan Di Konawe Utara Dengan Metode Naïve Bayes," *Pros. Semin. Nas. Pemanfaat. SAINS DAN Teknol. Inf.*, vol. 1, no. 1, pp. 97–102, 2023.

[21] M. T. Anjasmoros, I. Istiadi, and F. Marisa, "Analisis sentimen aplikasi go-jek menggunakan metode svm dan nbc (studi kasus: komentar pada play store)," *Conf. Innov. Appl. Sci. Technol. (CIASTECH 2020)*, no. Ciastech, pp. 489–498, 2020.

[22] H. A. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," *Babylonian J. Mach. Learn.*, vol. 2024, pp. 69–79, 2024.

[23] M. P. Little, P. S. Rosenberg, and A. Arsham, "Alternative stopping rules to limit tree expansion for random forest models," *Sci. Rep.*, no. 0123456789, pp. 1–6, 2022, doi: 10.1038/s41598-022-19281-7.

[24] G. A. B. Suryanegara, Adiwijaya, and M. D. P. D. Purbolaksono, "JURNAL RESTI Peningkatan Hasil Klasifikasi pada Algoritma Random Forest untuk," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 1, no. 10, pp. 114–122, 2021.

[25] K. G. Ayu, D. W. Sari, I. Farida, D. Harsono, and R. W. Kosaman, "Classification of Employee Competency Assessment Using Naï ve Bayes and K-Nearest Neighbor (KNN) Algorithms," *J. Adv. Inf. Technol.*, vol. 15, no. 7, 2024, doi: 10.12720/jait.15.7.879-885.

[26] U. Purba, Mariana ; Ermatita; Niprison , Handrie; Ayumi, Vina; Salamah, "Effect of Random Splitting and Cross Validation for Indonesian Opinion Mining using Machine Learning Approach," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 9, pp. 145–151, 2022.

[27] S. Sathyanarayanan and B. R. Tantri, "Confusion Matrix-Based Performance Evaluation Metrics," *African J. Biomed. Res.*, vol. 27, no. 4, 2024.

[28] G. Zeng, "Invariance Properties and Evaluation Metrics Derived from the Confusion Matrix in Multiclass Classification," *Math. J.*, no. 1997, 2025.

[29] K. M. Sujon, R. Hassan, K. Choi, and A. Samad, "empirical evidence from advanced statistics , ML , and XAI for evaluating business predictive models," *J. Big Data*, 2025

UMN

ISSN 2085-4552

Speech Emotion Recognition through Acoustic Data Augmentation and Attention-Driven CRNN-BiGRU

Muhamad Fitra Kacamarga[1], Kresna Andika Aprianto[2]
[1,2] Computer Science Department, BINUS Online Learning, Jakarta, Indonesia 11480
fitra.kacamarga@binus.ac.id[1], kresna.aprianto@binus.ac.id[2]

Accepted 21 June 2026
Approved 28 June 2026

***Abstract*— Speech emotion recognition (SER) systems have transformed human-computer interactions by enabling machines to identify emotional cues in speech. This study presents a *systematic* approach that combines complementary data-augmentation techniques with an efficient neural architecture to address data scarcity and model-robustness limitations. The proposed methodology applies a multi-dimensional augmentation framework that retains the original recordings while adding mildly perturbed copies targeting acoustic-environment variability (background-noise injection), speaking-rate variability (time stretching at rates of 0.9 and 1.1), and vocal-pitch variability (pitch shifting by one to two semitones). This augmented dataset feeds into a novel Convolutional Recurrent Neural Network (CRNN) integrated with a Bidirectional Gated Recurrent Unit (BiGRU) and a Bahdanau attention mechanism, designed to capture both local spectral and temporal emotional features effectively. The lightweight model (166,729 parameters, ≈0.64 MB) processes log-Mel spectrograms with Voice Activity Detection. On the RAVDESS database, the model achieves a weighted accuracy (WA) of 85.42% and an unweighted accuracy (UA) of 85.35%; data augmentation provides a consistent +1.74-point WA improvement over the non-augmented baseline, and the model outperforms comparable prior methods (e.g., +6.1 points WA over a CNN with Multi-Head attention) while using far fewer parameters than typical deep SER models. These results validate the effectiveness of integrating systematic data augmentation with an efficient neural architecture for SER applications.**

Index Terms*— *Speech Emotion Recognition; CRNN; BiGRU; Attention Mechanism; Data Augmentation; Deep Learning; Audio Processing; Log-Mel Spectrogram.

I. INTRODUCTION

Human-Computer Interaction (HCI) encompasses the study, planning, and design of the interaction between users and computers. While traditional HCI focuses on efficiency and usability, modern systems increasingly aim to emulate the nuanced characteristics of human-to-human communication, particularly in recognizing and responding to emotional states. Among the various modalities of HCI, three key domains have emerged as critical for natural interactions: speech processing [1]–[7], textual analysis [8], [9], and facial recognition [10], [11]. The effective integration of these modalities enables systems to comprehend and respond to diverse human experiences [12].

Speech Emotion Recognition (SER), an automated process for identifying emotional states from voice signals, has emerged as a promising advancement in this field. Recent developments in deep learning and signal processing have created new opportunities for SER applications, including healthcare and wellness for psychological assessment [14], consumer technology for smart-home automation [13], and professional settings for human-robot interaction [15]. This broad spectrum demonstrates the transformative potential of SER in enhancing human-computer interactions.

Despite these advances, current SER systems face significant limitations. Unlike Automatic Speech Recognition (ASR), which benefits from vast standardized datasets, SER research is impeded by data scarcity due to the expense of recording emotions in controlled environments and the prevalence of noisy real-life data. The subjective nature of emotions, influenced by language and culture, further increases annotation complexity [21]. Recognizing emotions in spontaneously expressed speech also poses a greater challenge than scripted performances [22], [23], particularly under environmental noise and cross-cultural variation.

A typical SER system comprises three components: a comprehensive corpus of speech emotions, effective feature-extraction techniques, and efficient classification algorithms. Commonly employed features include amplitude, Mel-frequency cepstral coefficients (MFCC), log-Mel spectrograms, and the zero-crossing rate [24]. While traditional methods such as Support Vector Machines and Hidden Markov Models have been used for acoustic-feature classification, the advent of deep learning has driven substantial progress. For example, Lee *et al.* [25] achieved 63.89% accuracy using Bidirectional Long Short-Term Memory (BiLSTM) networks, while Lim *et al.* [26] reached 86.06% on the EMO-DB database with convolutional neural networks.

To address these persistent challenges, this study introduces two key innovations:

A systematic data augmentation framework that enhances model robustness through strategic, multi-dimensional manipulation of acoustic properties while retaining the original samples.

A lightweight CRNN-BiGRU architecture (166,729 parameters) that integrates local feature extraction with temporal pattern recognition and an attention mechanism.

These innovations address significant deficiencies in contemporary SER research, namely the need for reliable performance across varied acoustic environments and an efficient model design suitable for practical deployment.

II. Literature Review

Speech Emotion Recognition (SER) is the task of automatically identifying human emotions from spoken utterances. Over the past five years, deep-learning approaches have significantly improved SER performance [1]. Models based on Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) can learn directly from raw audio or spectrograms, eliminating manual feature engineering [2].

2.1 CNN-Based Models

CNNs excel at extracting local spectral-temporal patterns through learned filters. Bhangale and Kothandaraman (2023) developed a lightweight 1-D CNN that achieved 93–94% accuracy on the Emo-DB and RAVDESS datasets [4]. The primary advantages of CNN-based SER include automatic feature learning and a degree of noise robustness through pooling operations [5]. However, CNNs with a limited receptive field may not capture the long-range temporal dependencies that are critical for emotion recognition, as emotions can evolve over an entire utterance.

2.2 RNN-Based Models

RNNs directly model the temporal sequence of acoustic features, making them well suited to capturing the dynamic patterns of emotion in speech. Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs) are the most prominent RNN variants in recent SER research. Trinh *et al.* (2022) compared a CNN, a CNN-RNN hybrid, and a pure GRU model on the IEMOCAP corpus; the GRU network achieved the highest accuracy (97.47%), outperforming both the CNN and the CNN+RNN models under identical conditions [7], demonstrating that the GRU gating mechanism effectively captures emotional dynamics over time.

2.3 Hybrid CNN-RNN (CRNN) Models

Hybrid architectures that combine convolutional and recurrent layers have become a cornerstone of modern SER systems. CNN layers first extract high-level features from spectrogram segments, and RNN layers then process the sequence of these features over time, leveraging both local feature extraction and global temporal modeling [8], [9]. Yadav *et al.* (2021) proposed a convolutional recurrent network using CNN followed by a bidirectional LSTM; with noise-reduction preprocessing and carefully chosen augmentation, their CRNN attained 85.8% accuracy on RAVDESS and 91.6% on Emo-DB [10], [11]. Another study reported approximately 90% accuracy on RAVDESS [12].

2.4 Bidirectional RNNs and BiGRU

Bidirectional architectures (such as BiLSTM or BiGRU) process sequences in both forward and backward directions, providing access to past and future context simultaneously. Xu *et al.* (2024) presented an SER model employing a Bi-GRU with a multi-head attention mechanism that significantly outperformed comparable unidirectional models across multiple metrics [14]–[16], emphasizing the benefit of capturing the complete temporal context of an utterance.

2.5 Attention Mechanisms

Attention mechanisms allow networks to focus on the most salient parts of a speech signal. In an RNN-based SER model, an attention layer learns to weigh the importance of each time frame before the final emotion prediction — valuable in speech, where not all moments are equally informative [26]. In addition to RNNs, pure attention-based architectures have emerged: a recent parallel CNN-transformer architecture achieved 89.3% accuracy, outperforming a comparable CNN–BiLSTM with attention under identical conditions [23].

2.6 Data Augmentation Techniques

Data augmentation addresses the limited size and imbalance of emotion datasets and improves model robustness. Common methods include:

Noise Injection: Adding background or white Gaussian noise to simulate different acoustic environments [24].

Pitch and Speed Variations: Altering pitch or playback speed to mimic different speakers or speaking styles [25].

Room and Channel Effects: Applying filters to simulate different recording conditions.

Segment Augmentation: Splitting utterances or random cropping to increase the sample count.

Empirical evidence shows that augmentation can substantially boost SER performance. Yadav *et al.* (2021) improved their CNN-BiLSTM model from the low 80s to 85.8% on RAVDESS [26], and an augmented CNN model achieved 92.6% on RAVDESS [27]

III. METHOD

This study adopts an experimental methodology to assess a Convolutional Recurrent Neural Network (CRNN) integrated with a Bidirectional Gated Recurrent Unit (BiGRU) for the classification of a speech-emotion dataset. The stages of the study — data preparation, data augmentation, feature extraction, model training, and evaluation — are depicted in Fig. 1. The model's performance is evaluated on the test set using weighted and unweighted accuracy, and is further analyzed through a confusion matrix.

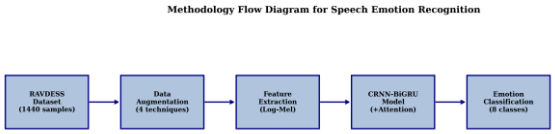


Fig. 1. *Flow diagram of the proposed speech-emotion-recognition methodology, showing the pipeline from dataset input through augmentation, feature extraction, model training, and emotion classification.*

3.1 Emotion Dataset

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) was used in this study. Developed at Ryerson University in Canada [30], the full database is a high-quality collection designed for emotion-recognition research. This work uses the **speech audio** subset, which comprises **1,440 recordings** from 24 professional actors (12 male and 12 female), all native North American English speakers.

TABLE I. **RAVDESS (SPEECH SUBSET) STATISTICS**

Attribute	Value
Total audio files	1,440
Number of emotion classes	8
Number of actors	24 (12 male, 12 female)
Files per actor	60
Sampling rate	48 kHz
Recording environment	Professional soundproof booth

TABLE II. **EMOTION CLASS DISTRIBUTION**

Emotion	Code	Sample Count	Percentage
Neutral	1	96	6.67%
Calm	2	192	13.33%
Happy	3	192	13.33%
Sad	4	192	13.33%
Angry	5	192	13.33%
Fearful	6	192	13.33%
Disgust	7	192	13.33%
Surprised	8	192	13.33%

Dataset balance analysis. The RAVDESS speech subset exhibits a near-balanced distribution across emotion classes. The neutral class contains 96 samples (normal intensity only), whereas the other seven classes contain 192 samples each (96 normal + 96 strong intensity). This slight imbalance (a neutral-to-other ratio of 1:2) is inherent to the dataset design, since neutral expressions lack the intensity variation present in the other emotions. The dataset nonetheless maintains balanced representation across speakers (60 files per actor) and gender (equal male/female distribution), which is why no class-weighting strategy was required.

The dataset's key characteristics for SER research are: (1) **emotional diversity** — eight expressions (neutral, calm, happy, sad, angry, fearful, disgust, surprised); (2) **controlled conditions** — all recordings captured in a soundproof booth at 48 kHz; (3) **standardized lexical content** — each actor utters the same two statements, *"Kids are talking by the door"* and *"Dogs are sitting by the door,"* preventing lexical bias; and (4) **two intensity levels** (normal and strong) for all emotions except neutral.

3.2 Systematic Data Augmentation Framework

Prior to feature extraction, a systematic multi-dimensional augmentation framework was applied to the training data. Here, *"systematic"* denotes a framework that targets several distinct acoustic dimensions of speech variability rather than a single transformation. Crucially, the framework **retains the original (clean) recordings and adds mildly perturbed copies** (two augmented copies per training utterance); the model therefore learns from both the clean distribution and its variations, which avoids the train/test distribution mismatch that overly aggressive augmentation can cause. Each augmented copy applies the following transforms probabilistically, illustrated in Fig. 2:

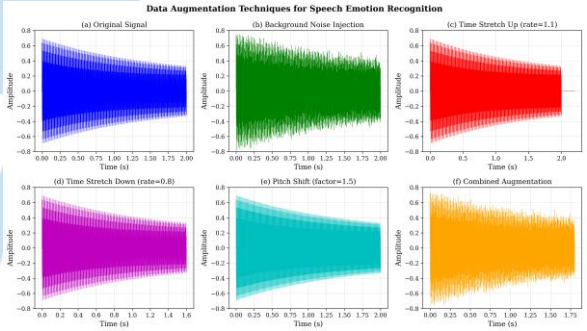


Fig. 2. *Illustration of the data-augmentation techniques applied to a speech sample: (a) original signal, (b) background-noise injection, (c)–(d) time stretching, (e) pitch shifting, and (f) combined augmentation.*

Acoustic-environment variability (background-noise injection): low-level Gaussian noise (amplitude factor 0.02 of the signal peak) added with 30% probability per copy, to simulate diverse recording environments.

Speaking-rate variability (time stretching): the time axis is stretched with 50% probability at a rate of either 0.9 or 1.1, to account for slower or faster speaking patterns.

Vocal-pitch variability (pitch shifting): the pitch is shifted with 50% probability by ±1–2 semitones, to simulate speaker differences.

This multi-dimensional strategy helps the model learn emotional features that are robust to recording

conditions and speaker variation; retaining the clean samples ensures the training distribution remains aligned with the (clean) test distribution.

3.3 Feature Extraction

3.3.1 Feature Selection: Log-Mel Spectrogram vs. MFCC

This study employs log-Mel spectrograms rather than Mel-Frequency Cepstral Coefficients (MFCCs) as the primary acoustic feature, for the following technical reasons:

Preservation of spectral structure. Log-Mel spectrograms retain the complete time–frequency representation of speech, including harmonic relationships and spectral dynamics that are important for emotion recognition. MFCCs apply a Discrete Cosine Transform (DCT) that decorrelates and compresses this information, potentially discarding fine-grained spectral patterns that encode emotional state.

Compatibility with deep learning. CNNs are designed to learn hierarchical features from structured input. Log-Mel spectrograms preserve the spatial relationship between frequency bands, allowing convolutional filters to capture local spectral–temporal patterns effectively.

Retention of prosodic information. The Mel filterbank preserves perceptually relevant frequency resolution along with fundamental frequency (F0), formant transitions, and energy contours — critical emotional indicators. The cepstral analysis in MFCCs tends to remove pitch-related information that helps distinguish, for example, anger (high F0) from sadness (low F0).

State-of-the-art practice. Recent high-performing SER systems consistently employ log-Mel or raw spectrograms as input features [4], [32].

TABLE III. **FEATURE-EXTRACTION PARAMETERS**

Parameter	Value
Number of Mel bands	64
Maximum frequency	6,000 Hz
Window length	2,048 samples
Hop length	1,365 samples (2048 × 2/3)
FFT size	4,096

The extraction process converts raw audio waveforms into time-domain signals and then applies Voice Activity Detection (VAD) to identify the significant portions of the signal. The result is a truncated log-Mel spectrogram that focuses on the most pertinent segments. A representative waveform and its corresponding log-Mel spectrogram are shown in Fig. 3.

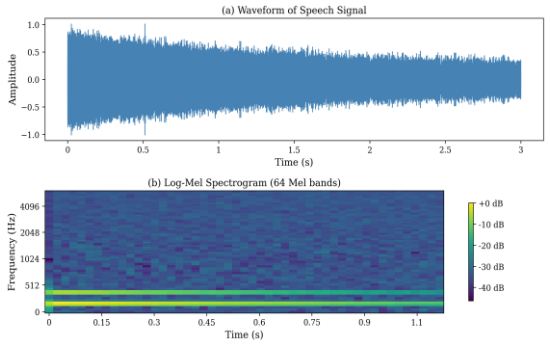


Fig. 3. *(a) Waveform of a speech sample after preprocessing and (b) the corresponding log-Mel spectrogram (64 Mel bands) used as the model input.*

3.4 Model Design

The proposed model combines the strengths of a CRNN for local feature extraction with a BiGRU and an attention mechanism for global temporal modeling. The complete architecture is shown in Fig. 4.

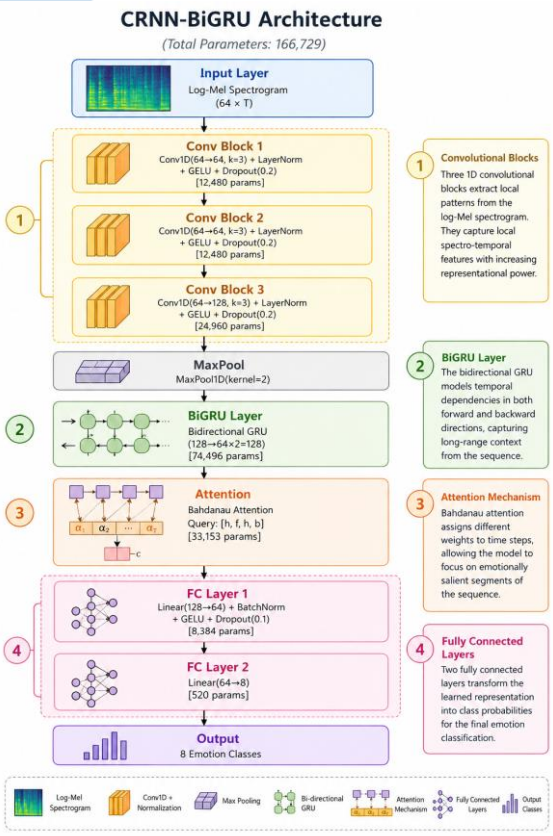


Fig. 4. *Detailed architecture of the proposed CRNN-BiGRU model, showing four main components: (1) three convolutional blocks for local feature extraction from the log-Mel spectrogram, (2) a bidirectional GRU layer for temporal modeling, (3) a Bahdanau attention mechanism for focusing on emotionally salient segments, and (4) fully connected layers for classification. Total parameters: 166,729.*

3.4.1 Architecture Justification

CRNN vs. CNN or RNN alone. The hybrid CRNN architecture was chosen on the basis of evidence that CNNs excel at extracting local spectral features while RNNs capture temporal dependencies. CNN-only approaches [34] (71.6% accuracy) do not adequately model temporal relationships in emotional speech, whereas RNN-only methods [33] (69.4% UA) struggle to extract robust local features. The CRNN combination leverages the complementary strengths of both.

BiGRU vs. LSTM/BiLSTM. BiGRU was chosen for three reasons: (i) *computational efficiency* — GRUs have fewer parameters than LSTMs, reducing training time while maintaining comparable performance; (ii) *gradient stability* — GRU gating mechanisms preserve long-term dependencies; and (iii) *bidirectional processing* — capturing both past and future context, which is important where anticipatory and carryover emotional cues span the utterance.

Bahdanau attention vs. other mechanisms. Bahdanau attention was selected after experimental evaluation, offering dynamic focus on emotionally salient segments, the ability to emphasize relevant acoustic features while down-weighting less informative portions, and stable gradient flow during training. The attention mechanism addresses a limitation of standard RNN approaches [25], which use only the final hidden state for classification and may lose emotional information distributed throughout the utterance.

ConvLN and GELU activation. Layer Normalization was chosen over Batch Normalization for stability with the variable-length sequences common in speech, and the GELU activation was selected for its smoother gradient properties, which improved performance in preliminary testing relative to ReLU.

3.4.2 Architecture Details

The architecture comprises:

Three convolutional blocks, each containing Conv1D + Layer Normalization + GELU + Dropout (0.2):

ConvBlock 1: 64 → 64 channels, kernel = 3

ConvBlock 2: 64 → 64 channels, kernel = 3

ConvBlock 3: 64 → 128 channels, kernel = 3

MaxPool1D (kernel = 2) for temporal downsampling.

Bidirectional GRU (input 128, hidden 64 per direction → 128-dimensional output) preceded by Layer Normalization.

Bahdanau attention, which uses the concatenated final BiGRU hidden states as the query over all BiGRU outputs, with attention weights obtained via a softmax.

Classification head: Linear(128 → 64) + BatchNorm + GELU + Dropout(0.1), followed by Linear(64 → 8 classes).

TABLE IV. **PROPOSED-MODEL COMPLEXITY ANALYSIS**

Component	Parameters	Percentage
Convolutional blocks (3)	49,920	29.9%
Bidirectional GRU	74,496	44.7%
Bahdanau attention	33,153	19.9%
Fully connected layers	8,904	5.3%
Layer normalization	256	0.2%
Total	**166,729**	**100%**
Model size (FP32)	**≈0.64 MB**	—

With only **166,729 trainable parameters (≈0.64 MB)**, the proposed model is markedly more lightweight than typical deep SER systems, which commonly rely on millions of parameters (e.g., deep 2-D CNNs or transformer/self-supervised speech models). This compact design enables deployment on resource-constrained devices while preserving competitive accuracy, and constitutes the model-complexity contribution of this work.

3.5 Model Training

The reported results were obtained using a **stratified hold-out** evaluation: the RAVDESS speech subset was partitioned into a training set (80%) and a test set (20%) at the utterance level, stratified by emotion to preserve the class proportions, with a fixed random seed for reproducibility. All data were shuffled prior to partitioning. Data augmentation (Section 3.2) was applied to the **training partition only**; the test partition consists of clean recordings.

The training configuration was held constant across experiments:

Optimizer: AdaBelief (learning rate 1×10^{-3}, β = (0.9, 0.999), weight decay 1×10^{-7})

Loss: label-smoothing cross-entropy (smoothing = 0.1)

Learning-rate scheduler: ReduceLROnPlateau (factor 0.7, patience 2, cooldown 3)

Epochs: 100; mini-batch size: 32

We note that the present evaluation uses an utterance-level split in which the 24 actors appear in both the training and test partitions; a fully speaker-independent evaluation (e.g., leave-speaker-group-out cross-validation) is identified as an important direction for future work.

3.6 Model Evaluation

After training, the model was evaluated on the held-out test set using the following metrics.

Weighted Accuracy (WA) — the proportion of correctly classified samples, weighted by class proportion:

$$WA = \frac{\sum_{i=1}^{C} n_i \cdot r_i}{\sum_{i=1}^{C} n_i}$$

where n_i is the number of samples in class i, r_i is the recall for class i, and C is the number of classes.

Unweighted Accuracy (UA) — the average per-class accuracy, giving equal importance to each class regardless of its frequency:

$$UA = \frac{1}{C}\sum_{i=1}^{C} r_i = \frac{1}{C}\sum_{i=1}^{C} \frac{TP_i}{TP_i + FN_i}$$

where TP_i and FN_i are the true positives and false negatives for class i. A confusion matrix was also generated to visualize the distribution of true and predicted labels across the eight emotion classes.

IV. Result and Discussions

The model was implemented in PyTorch and trained for 100 epochs with a batch size of 32. Table V reports the test performance with and without data augmentation.

TABLE V. **Results With and Without Data Augmentation**

Data	WA (%)	UA (%)
Non-Augmented	83.68	83.78
Augmented	85.42	85.35

Data augmentation improves performance by **+1.74 percentage points in WA** (83.68% → 85.42%) and +1.57 points in UA (83.78% → 85.35%). The gain is most pronounced for the harder classes — for example, the recall for *fearful* rises from 79.5% to 92.3% and for *calm* from 87.2% to 92.3% — indicating that augmentation broadens the effective coverage of the training distribution and encourages features that are less dependent on the original recording conditions. Retaining the clean samples alongside the augmented copies is important: it keeps the training distribution aligned with the clean test set, so that augmentation improves rather than degrades generalization.

Confusion-matrix analysis. The model (with augmentation) shows strong, well-balanced performance across the eight RAVDESS emotion classes, with pronounced diagonal dominance (Fig. 5):

Highest accuracy: disgust (97.4%), calm (92.3%), fearful (92.3%), and angry (92.1%).

Strong performance: surprised (86.8%) and neutral (84.2%).

Hardest classes: happy (69.2%) and sad (68.4%).

The most common confusions occur between acoustically similar emotion pairs: *happy* is most often mistaken for *surprised* (15.4%) — both high-arousal expressions — while *sad* is confused with *fearful* (15.8%) and *disgust* (7.9%). These results highlight *happy* and *sad* as the priority targets for future feature engineering.

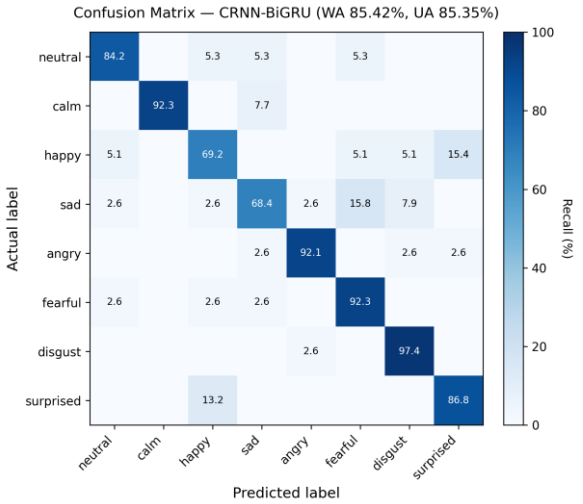


Fig. 5. *Confusion matrix (row-normalized recall, %) of the proposed CRNN-BiGRU model with data augmentation on the RAVDESS test set (WA 85.42%, UA 85.35%). Rows are the actual labels and columns are the predicted labels.*

Comparative performance. Table VI compares the proposed model with previously reported results on the RAVDESS dataset.

TABLE VI. **Comparison of Accuracy With Other Studies on RAVDESS**

Method	WA (%)	UA (%)	Year
GResNet + Multi-Task [28]	64.5	64.6	2019
BLSTM + Capsule Routing [33]	—	69.4	2019
Multimodal Fine-grained Learning [29]	77.0	74.7	2020
DCGAN + Augmentation [31]	72.3	72.3	2023
Head Fusion [30]	77.8	77.4	2023
CNN + Multi-Head [32]	79.3	78.2	2023
CRNN-BiGRU (proposed)	**83.68**	**83.78**	2025
CRNN-BiGRU + Augmentation (proposed)	**85.42**	**85.35**	2025

Relative to these methods, the proposed CRNN-BiGRU with augmentation is competitive while being far more compact — for example, +6.1 percentage points in WA over CNN + Multi-Head [32] (79.3%) and +7.6 points over Head Fusion [30] (77.8%). The gains are consistent with the architecture's design: the CRNN integrates spectral and temporal information, the BiGRU captures the temporal evolution of emotions, and the attention mechanism concentrates on emotionally salient frames. Crucially, these results are achieved with only 166,729 parameters (≈0.64 MB), making the proposed model an order of magnitude more compact than the deep CNN and attention-based systems it is compared against; this favorable

accuracy-to-size trade-off is the central contribution of this work

V. CONCLUSIONS

This research proposes a structure that combines Convolutional Neural Networks with a Bidirectional Gated Recurrent Unit, a Bahdanau attention mechanism, and a systematic data-augmentation framework for speech emotion recognition on the RAVDESS dataset. The incorporation of attention enhances the model's focus on key speech segments, while the lightweight design (166,729 parameters, ≈0.64 MB) keeps the computational cost low. Evaluated by weighted and unweighted accuracy, the model achieves a WA of 85.42% and a UA of 85.35%, outperforming the comparable prior approaches summarized in Table VI (all ≤ 79.3% WA) while using far fewer parameters, with data augmentation contributing a +1.74-point WA gain over the non-augmented baseline.

The principal contributions are: (1) a systematic augmentation framework that targets multiple acoustic dimensions while retaining the original samples, yielding a consistent accuracy gain; (2) an efficient CRNN-BiGRU architecture with Bahdanau attention suitable for resource-constrained deployment; and (3) a detailed model-complexity analysis demonstrating that a carefully designed CNN-RNN architecture can reach competitive accuracy at a small fraction of the parameter budget of typical deep SER models.

The model nonetheless has limitations. It can struggle with acoustically similar expressions (e.g., happy vs. surprised and sad vs. fearful), and the present evaluation uses an utterance-level stratified hold-out in which speakers are shared between the training and test partitions; a fully speaker-independent assessment is left for future work. Future research could also explore augmentation tailored to specific emotion categories and multimodal approaches that combine acoustic features with complementary information sources.

REFERENCES

[1] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Interspeech 2014*, ISCA, Sep. 2014, pp. 223–227, doi: 10.21437/Interspeech.2014-57.

[2] M. Chen, X. He, J. Yang, and H. Zhang, "3-D Convolutional Recurrent Neural Networks With Attention Model for Speech Emotion Recognition," *IEEE Signal Process. Lett.*, vol. 25, no. 10, pp. 1440–1444, Oct. 2018, doi: 10.1109/LSP.2018.2860246.

[3] X. Wu *et al.*, "Speech Emotion Recognition Using Capsule Networks," in *ICASSP 2019*, IEEE, May 2019, pp. 6695–6699, doi: 10.1109/ICASSP.2019.8683163.

[4] Y. Xu, H. Xu, and J. Zou, "HGFM: A Hierarchical Grained and Feature Model for Acoustic Emotion Recognition," in *ICASSP 2020*, IEEE, May 2020, pp. 6499–6503, doi: 10.1109/ICASSP40776.2020.9053039.

[5] D. Priyasad, T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Attention Driven Fusion for Multi-Modal Emotion Recognition," in *ICASSP 2020*, IEEE, May 2020, pp. 3227–3231, doi: 10.1109/ICASSP40776.2020.9054441.

[6] A. Nediyanchath, P. Paramasivam, and P. Yenigalla, "Multi-Head Attention for Speech Emotion Recognition with Auxiliary Learning of Gender Recognition," in *ICASSP 2020*, IEEE, May 2020, pp. 7179–7183, doi: 10.1109/ICASSP40776.2020.9054073.

[7] M. Xu, F. Zhang, and S. U. Khan, "Improve Accuracy of Speech Emotion Recognition with Attention Head Fusion," in *2020 10th Annual Computing and Communication Workshop and Conference (CCWC)*, IEEE, Jan. 2020, pp. 1058–1064, doi: 10.1109/CCWC47524.2020.9031207.

[8] A. Chatterjee, U. Gupta, M. K. Chinnakotla, R. Srikanth, M. Galley, and P. Agrawal, "Understanding Emotions in Text Using Deep Learning and Big Data," *Comput. Human Behav.*, vol. 93, pp. 309–317, Apr. 2019, doi: 10.1016/j.chb.2018.12.029.

[9] E. Batbaatar, M. Li, and K. H. Ryu, "Semantic-Emotion Neural Network for Emotion Recognition From Text," *IEEE Access*, vol. 7, pp. 111866–111878, 2019, doi: 10.1109/ACCESS.2019.2934529.

[10] J. Yang, F. Zhang, B. Chen, and S. U. Khan, "Facial Expression Recognition Based on Facial Action Unit," in *2019 Tenth International Green and Sustainable Computing Conference (IGSC)*, IEEE, Oct. 2019, pp. 1–6, doi: 10.1109/IGSC48788.2019.8957163.

[11] X. Liu, B. V. K. V. Kumar, J. You, and P. Jia, "Adaptive Deep Metric Learning for Identity-Aware Facial Expression Recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, Jul. 2017, pp. 522–531, doi: 10.1109/CVPRW.2017.79.

[12] H. L. O'Brien, I. Roll, A. Kampen, and N. Davoudi, "Rethinking (Dis)engagement in human-computer interaction," *Comput. Human Behav.*, vol. 128, p. 107109, Mar. 2022, doi: 10.1016/j.chb.2021.107109.

[13] R. Chatterjee, S. Mazumdar, R. S. Sherratt, R. Halder, T. Maitra, and D. Giri, "Real-Time Speech Emotion Analysis for Smart Home Assistants," *IEEE Trans. Consum. Electron.*, vol. 67, no. 1, pp. 68–76, Feb. 2021, doi: 10.1109/TCE.2021.3056421.

[14] L.-S. A. Low, M. C. Maddage, M. Lech, L. B. Sheeber, and N. B. Allen, "Detection of Clinical Depression in Adolescents' Speech During Family Interactions," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 3, pp. 574–586, Mar. 2011, doi: 10.1109/TBME.2010.2091640.

[15] X. Huahu, G. Jue, and Y. Jian, "Application of Speech Emotion Recognition in Intelligent Household Robot," in *2010 International Conference on Artificial Intelligence and Computational Intelligence*, IEEE, Oct. 2010, pp. 537–541, doi: 10.1109/AICI.2010.118.

[16] W.-J. Yoon, Y.-H. Cho, and K.-S. Park, "A Study of Speech Emotion Recognition and Its Application to Mobile Services," in *Ubiquitous Intelligence and Computing*, Springer, pp. 758–766, doi: 10.1007/978-3-540-73549-6_74.

[17] S. G. Koolagudi, K. Sharma, and K. Sreenivasa Rao, "Speaker Recognition in Emotional Environment," 2012, pp. 117–124, doi: 10.1007/978-3-642-32112-2_15.

[18] P. Ekman and H. Oster, "Facial Expressions of Emotion," *Annu. Rev. Psychol.*, vol. 30, no. 1, pp. 527–554, Jan. 1979, doi: 10.1146/annurev.ps.30.020179.002523.

[19] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, Mar. 2011, doi: 10.1016/j.patcog.2010.09.020.

[20] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *J. Res. Pers.*, vol. 11, no. 3, pp. 273–294, Sep. 1977, doi: 10.1016/0092-6566(77)90037-X.

[21] R. Altrov and H. Pajupuu, "The influence of language and culture on the understanding of vocal emotions," *Journal of*

Estonian and Finno-Ugric Linguistics, vol. 6, no. 3, pp. 11–48, Dec. 2015, doi: 10.12697/jeful.2015.6.3.01.

[22] Z. Zhao *et al.*, "Self-attention transfer networks for speech emotion recognition," *Virtual Reality & Intelligent Hardware*, vol. 3, no. 1, pp. 43–54, Feb. 2021, doi: 10.1016/j.vrih.2020.12.002.

[23] P. Li, Y. Song, I. McLoughlin, W. Guo, and L. Dai, "An Attention Pooling Based Representation Learning Method for Speech Emotion Recognition," in *Interspeech 2018*, ISCA, Sep. 2018, pp. 3087–3091, doi: 10.21437/Interspeech.2018-1242.

[24] A. Sandesara, S. Parikh, P. Sapovadiya, and M. Rahevar, "A Comparative Study On Speech Emotion Recognition," *Int. J. Res. Eng. Sci. Manag.*, vol. 3, no. 11, pp. 25–35, Nov. 2020, doi: 10.47607/ijresm.2020.366.

[25] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Interspeech 2015*, ISCA, Sep. 2015, pp. 1537–1540, doi: 10.21437/Interspeech.2015-336.

[26] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and Recurrent Neural Networks," in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, IEEE, Dec. 2016, pp. 1–4, doi: 10.1109/APSIPA.2016.7820699.

[27] X. Han, F. Chen, and J. Ban, "Music Emotion Recognition Based on a Neural Network with an Inception-GRU Residual Structure," *Electronics*, vol. 12, no. 4, p. 978, Feb. 2023, doi: 10.3390/electronics12040978.

[28] Y. Zeng, H. Mao, D. Peng, and Z. Yi, "Spectrogram based multi-task audio classification," *Multimed. Tools Appl.*, vol. 78, no. 3, pp. 3705–3722, Feb. 2019, doi: 10.1007/s11042-017-5539-3.

[29] H. Li, W. Ding, Z. Wu, and Z. Liu, "Learning Fine-Grained Cross Modality Excitement for Speech Emotion Recognition," Oct. 2020.

[30] M. Xu, F. Zhang, and W. Zhang, "Head Fusion: Improving the Accuracy and Robustness of Speech Emotion Recognition on the IEMOCAP and RAVDESS Dataset," *IEEE Access*, vol. 9, pp. 74539–74549, 2021, doi: 10.1109/ACCESS.2021.3067460.

[31] J.-Y. Baek and S.-P. Lee, "Enhanced Speech Emotion Recognition Using DCGAN-Based Data Augmentation," *Electronics*, vol. 12, no. 18, p. 3966, Sep. 2023, doi: 10.3390/electronics12183966.

[32] R. Ullah *et al.*, "Speech Emotion Recognition Using Convolution Neural Networks and Multi-Head Convolutional Transformer," *Sensors*, vol. 23, no. 13, p. 6212, Jul. 2023, doi: 10.3390/s23136212.

[33] Md. A. Jalal, E. Loweimi, R. K. Moore, and T. Hain, "Learning Temporal Clusters Using Capsule Routing for Speech Emotion Recognition," in *Interspeech 2019*, ISCA, Sep. 2019, pp. 1701–1705, doi: 10.21437/Interspeech.2019-3068.

[34] D. Issa, M. Fatih Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomed. Signal Process. Control*, vol. 59, p. 101894, May 2020, doi: 10.1016/j.bspc.2020.101894

ISSN 2085-4552

Evaluating the Capability of Transfer Learning VGG16 for Detecting Tomato Diseases in Indonesia

Tegar Satriya Wiguna[1], Febri Liantoni[2*], Yudianto Sujana[3]
[1,2,3] Information Engineering and Computer Education, Sebelas Maret University
tegarsatriya17@student.uns.ac.id[1], febri.liantoni@gmail.com[2], yudianto.sujana@staff.uns.ac.id[3]

Accepted 19 June 2026
Approved 28 June 2026

***Abstract*— Tomato leaf diseases pose a significant threat to agricultural productivity in Indonesia, often leading to severe yield losses. This study evaluates the effectiveness of the VGG16 convolutional neural network in detecting tomato diseases, particularly when trained on the standardized PlantVillage dataset and applied to local agricultural conditions in Sragen, Central Java. To conduct this research, the methodology included standardizing images to 256x256 pixels and applying the GrabCut method for background removal to isolate leaf regions. The model was trained via transfer learning using pre-trained ImageNet weights, and the system was deployed for practical use through a FastAPI backend and a React Native frontend. The evaluation results showed that the VGG16 model achieved an 82% accuracy, along with an 80% average precision, 81% recall, and 81% F1-score on the PlantVillage test set. However, performance exhibited a sharp decline to a 25% overall accuracy when tested on locally sourced images, failing entirely to predict samples from the Bacterial Spot, Late Blight, and Leaf Mold classes. These findings highlight the model's limited generalization capabilities and emphasize the necessity of incorporating local datasets and domain adaptation strategies to develop AI-based plant disease detection tools that are effective in real-world settings. The study underscores the importance of contextualizing AI solutions for local agricultural environments to ensure their practical applicability and reliability.**

Index Terms*— *VGG16; Tomato; Leaf Diseases; Convulutional Neural Network.

I. INTRODUCTION

The tomato cultivation sector in Indonesia has experienced substantial growth in recent years. Data from the Indonesian Central Statistics Agency (BPS) indicated that tomato production reached approximately 1.16 million tons in 2022, reflecting a 4.88% increase compared to 1.11 million tons in the previous year. This positive trend is primarily due to comprehensive government initiatives aimed at boosting agricultural productivity and quality, alongside significant efforts from local farmers. Despite this progress, tomato farming continues to face substantial challenges, notably the prevalence of plant diseases that severely reduce crop yield and quality [1]. Diseases affecting tomatoes include early blight (Alternaria solani), late blight (Phytophthora infestans), leaf mold (Fulvia fulva), and bacterial spot (Xanthomonas campestris pv. vesicatoria).

Traditional methods of disease detection rely predominantly on visual inspections by farmers [2], which are labor-intensive, subjective, and often lead to delayed identification and ineffective management responses. Consequently, these delays frequently result in substantial crop losses and increased dependency on chemical pesticides, which pose environmental and health risks.

Recent advancements in deep learning technologies have demonstrated considerable potential for automated plant disease detection through image analysis[3]. Convolutional Neural Networks (CNNs), such as VGG16, ResNet, and DenseNet, have achieved significant accuracy in classifying plant diseases in controlled research environments, primarily utilizing standardized datasets like PlantVillage. Among these architectures, VGG16 stands out due to its effective balance between architectural simplicity and robust feature extraction capabilities [4].

However, the critical question remains regarding the practical applicability and generalization of models trained exclusively on datasets such as PlantVillage, which were captured under controlled conditions [5], to local agricultural contexts with distinct environmental characteristics. Given the variability in climate, leaf morphology, and disease manifestation in Indonesian tomato plantations, it is essential to evaluate whether a VGG16 model trained on the Kaggle PlantVillage dataset can effectively detect diseases on locally collected tomato leaf images in some research it limitation cause environment [6].

This research specifically addresses this gap by evaluating the performance of a VGG16-based model trained exclusively with the Kaggle PlantVillage dataset when applied to a local dataset collected from a tomato plantation in Sragen, Central Java, Indonesia. The objective is to investigate the effectiveness and

limitations of using internationally sourced datasets for local agricultural applications, thus providing insights into necessary adaptations for enhancing the robustness and applicability of AI-driven plant disease detection systems in Indonesian agriculture.

II. METHOD

In this research, a structured methodology was employed to evaluate the performance of the VGG16 model in detecting diseases in tomato plants using local datasets, particularly focusing on its generalization capability. The research stages include dataset collection, pre-processing, model training, backend and frontend development, testing, and analysis of results.

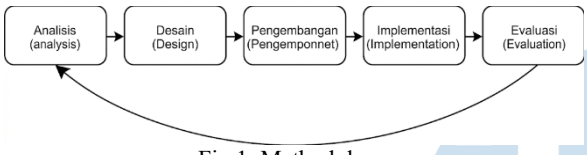


Fig 1. Methodology

The dataset used in this research was sourced from the Kaggle repository, specifically the PlantVillage dataset, which contains images categorized into bacterial spot, early blight, late blight, leaf mold, and healthy leaves. Additionally, local data was collected directly from a tomato plantation located in Sragen, Central Java, Indonesia, to test the model's ability to generalize across different environmental conditions and plant varieties.



Fig 2. Sample image from dataset kaggle

In the pre-processing stage, all collected images underwent standardization, which included resizing to a uniform size of 256 × 256 pixels [7] and background removal using the GrabCut method. Additionally, data augmentation techniques such as a rotation range of 20 degrees, zoom range up to 20%, and horizontal flipping were applied to enhance image variability and strengthen the model's ability to generalize across different leaf positions and orientations. This comprehensive pre-processing ensures compatibility with the input requirements of the VGG16 architecture.

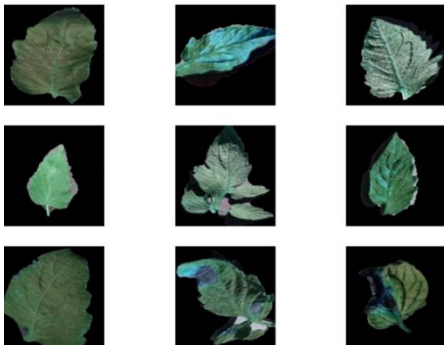


Fig 3. Sample image pre processing

The dataset was partitioned into training and testing with Kaggle dataset. The training phase involved employing the VGG16 architecture [8], a convolutional neural network known for its simplicity and strong feature extraction capabilities, using transfer learning with pre-trained weights from ImageNet.

For practical implementation, a backend system was developed using FastAPI to facilitate the integration of the trained model, allowing real-time inference[9] when processing uploaded images. A frontend interface was built using React Native, enabling users to interact easily [10] with the model through mobile devices. This user-friendly interface allows farmers to upload images and quickly receive diagnoses regarding the health status of tomato plants.

Testing involved evaluating the trained model using both the Kaggle dataset and the local dataset from Indonesia. Confusion matrices, accuracy, precision, recall, and F1-score metrics were employed to quantitatively assess model performance and highlight the model's strengths and limitations.

A. Accuracy

Measures the percentage of correct predictions out of all test data.

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

B. Precision

Measures how many of the model's predictions are actually correct.

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

TP is True Positive or correct prediction, and FP is False Positive or incorrect prediction.

C. Recall

Measures how many correct predictions were made by the model out of all actual positive cases.

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

TP is True Positive or correct prediction, and FN is False Negative or incorrect prediction.

The results from this testing phase provided critical insights into the generalization ability of the VGG16 model when trained exclusively on standardized datasets like PlantVillage, emphasizing the importance of including locally sourced images to enhance the practical applicability of AI models in real-world agricultural scenarios.

III. Result and Discussions

This study aims to evaluate the performance of the VGG16 Convolutional Neural Network (CNN) architecture in detecting tomato leaf diseases using a model trained on the PlantVillage dataset and tested on both the same dataset and local images from a tomato plantation in Sragen, Central Java, Indonesia. Evaluation metrics include accuracy, precision, recall, and F1-score, derived from confusion matrices. In addition, local dataset testing was conducted to assess the model's generalization capabilities.

A. Model Training Result

The VGG16 model was trained using 500 images from kaggle. Before the training process began, each image underwent a series of preprocessing steps aimed at enhancing input quality and consistency. This included background removal using the GrabCut algorithm to isolate leaf regions from distracting elements, thereby allowing the model to focus solely on disease-relevant features. Subsequently, all images were resized uniformly to a resolution of 256 × 256 pixels to meet the input size requirement of the VGG16 architecture. The model employed a transfer learning approach by utilizing pre-trained ImageNet weights, which were then fine-tuned specifically on the tomato leaf disease classification task by adjusting the final fully connected layers to match the five disease classes.

B. Evaluation on Planting Village Dataset

The model's initial evaluation was conducted using the PlantVillage test dataset, which served as a benchmark to measure baseline performance under controlled conditions. A confusion matrix was generated to present a detailed summary of classification outcomes across the five predefined classes: bacterial spot, early blight, late blight, leaf mold, and healthy. This matrix provides insights into the model's prediction accuracy and class-specific performance.

Table 1. Confusion matrix dataset kaggle

	Late Blight	Bacterial spot	Early blight	healthy	Leaf mold
Late blight	15	1	4	0	0
Bacterialspot	1	19	1	0	0
Early blight	3	1	11	0	4
Healthy	0	0	0	19	0
Leaf mold	1	0	1	0	18

Overall, the model achieved an accuracy of 82%, average precision of 80%, recall of 81%, and F1-score of 81%. The highest performance was observed in the Early Blight class, with all test samples correctly predicted, while Leaf Mold showed the lowest class-specific performance.

C. Evaluation on Local dataset

To assess the model's ability to generalize beyond the training domain, a set of 20 tomato leaf images collected from an actual plantation in Sragen, Central Java, was used for testing. These images represent real-world agricultural conditions, including varying lighting, backgrounds, and leaf textures, which differ significantly from the controlled settings of the PlantVillage dataset. The model's performance on this local dataset experienced a substantial decline, with an overall accuracy of just 25%**.** Notably, the model was only able to correctly classify 4 out of 5 images belonging to the Early Blight class. However, it failed to accurately predict any of the samples from the Bacterial Spot, Late Blight, and Leaf Mold classes, indicating a critical gap in the model's domain adaptability.

Table 2. Confusion Matrix dataset local

	Late Blight	Bacterial spot	Early blight	healthy	Leaf mold
Late Blight	0	2	0	0	1
Bacterial spot	0	0	3	0	1
Early blight	0	1	4	0	0
Healthy	0	0	1	1	1
Leaf mold	0	3	2	0	0

The high accuracy on the PlantVillage dataset demonstrates the effectiveness of the VGG16 architecture under controlled dataset conditions [11]. However, the sharp performance decline when tested on local images reveals a significant limitation [12] in generalization. This supports literature findings that AI models must incorporate diverse and localized training data to ensure robust real-world performance [13]. Future research should focus on collecting and integrating local datasets [14], applying domain adaptation techniques, and continuously fine-tuning models to match field conditions [15].



Fig 5. Prediction System

IV. CONCLUSIONS

This research demonstrated that the VGG16 convolutional neural network model, when trained exclusively on the PlantVillage dataset, is highly effective in classifying tomato leaf diseases under standardized conditions, achieving high accuracy, precision, recall, and F1-score metrics. However, when applied to a locally sourced dataset from an Indonesian tomato plantation, the model's performance declined sharply, revealing its limited ability to generalize across different domains.

The findings highlight a crucial insight: models trained on globally available datasets may not be directly applicable to real-world, localized agricultural settings without adaptation. Therefore, to develop a robust and practical disease detection system for Indonesian farmers, future implementations should incorporate localized training data, employ domain adaptation strategies, and consider continuous updates based on field data.

Ultimately, this study underscores the importance of contextualizing AI-driven agricultural tools for local environments and serves as a foundation for more inclusive, field-ready plant disease detection systems that bridge the gap between research and on-ground application.

REFERENCES

[1] Schmey, T., Tominello-Ramirez, C. S., Brune, C., & Stam, R. (2024). Alternaria diseases on potato and tomato. Molecular Plant Pathology, 25(3). https://doi.org/10.1111/mpp.13435

[2] Ali, M., Salma, M., Haji, M. E., & Jamal, B. (2025). Plant disease detection using vision transformers. International Journal of Electrical and Computer Engineering (IJECE), 15(2), 2334. https://doi.org/10.11591/ijece.v15i2.pp2334-2344

[3] Wang, S., Xu, D., Liang, H., Bai, Y., Li, X., Zhou, J., Su, C., & Wei, W. (2025). Advances in Deep Learning Applications for Plant Disease and Pest Detection: A Review. Remote Sensing, 17(4), 698. https://doi.org/10.3390/rs17040698

[4] Nurafiya, N., & Chandra, A. Y. (2024). Analisis Performa Akurasi Klasifikasi Citra Jenis Sayur Salada Menggunakan Arsitektur VGG16, Xception dan NasNetMobile. JURNAL MEDIA INFORMATIKA BUDIDARMA, 8(3), 1240. https://doi.org/10.30865/mib.v8i3.7661.

[5] Mohanty, S.P., Hughes, D., Salathe, M. (2016). Using Deep Learning for Image-Based Plant Disease Detection. arXiv preprint arXiv:1604.03169..

[6] Paymode, A. S., & Malode, V. B. (2022). Transfer Learning for Multi-Crop Leaf Disease Image Classification using Convolutional Neural Network VGG. Artificial Intelligence in Agriculture, 6, 23–33. https://doi.org/10.1016/j.aiia.2021.12.002

[7] Sinamenye, J. H., Chatterjee, A., & Shrestha, R. (2025). Potato plant disease detection: leveraging hybrid deep learning models. BMC Plant Biology, 25(1). https://doi.org/10.1186/s12870-025-06679-4.

[8] Al Sahili, Z., & Awad, M. (2022). The power of transfer learning in agricultural applications: AgriNet. Frontiers in

Plant Science, 13. https://doi.org/10.3389/fpls.2022.992700.

[9] Rahman, K. N., Banik, S. C., Islam, R., & Fahim, A. A. (2025). A real time monitoring system for accurate plant leaves disease detection using deep learning. Crop Design, 4(1), 100092. https://doi.org/10.1016/j.cropd.2024.100092.

[10] Govoruhina, A., & Nikiforova, A. (2022). Digital health shopping assistant with React Native: a simple technological solution to a complex health problem. In 2022 International Conference on Intelligent Data Science Technologies and Applications (IDSTA) (pp. 34–40). 2022 International Conference on Intelligent Data Science Technologies and Applications (IDSTA). IEEE. https://doi.org/10.1109/idsta55301.2022.9923047

[11] Noyan, M.A. (2022). Uncovering bias in the PlantVillage dataset. arXiv preprint arXiv:2206.04374..

[12] Fenu, G., & Malloci, F. M. (2022). Evaluating Impacts between Laboratory and Field-Collected Datasets for Plant Disease Classification. Agronomy, 12(10), 2359. https://doi.org/10.3390/agronomy12102359

[13] Hanna, M. G., Pantanowitz, L., Dash, R., Harrison, J. H., Deebajah, M., Pantanowitz, J., & Rashidi, H. H. (2025). Future of Artificial Intelligence—Machine Learning Trends in Pathology and Medicine. Modern Pathology, 38(4), 100705. https://doi.org/10.1016/j.modpat.2025.100705

[14] Kouw, W.M., Loog, M. (2018). An introduction to domain adaptation and transfer learning. arXiv preprint arXiv:1812.11806.

[15] Dong, X., Bao, J., Zhang, T., Chen, D., Gu, S., Zhang, W., Yuan, L., Chen, D., Wen, F., Yu, N. (2022). CLIP Itself is a Strong Fine-tuner: Achieving 85.7% and 88.0% Top-1 Accuracy with ViT-B and ViT-L on ImageNet. arXiv preprint arXiv:2212.06138

ISSN 2085-4552

Web-Based Information System for Monitoring Soil Conditions (Moisture and Temperature) Integrated with IoT

Ahmad Farzi Anwar[1], M. Husaini[2], Khairun Nita Aulia[3]
Department of Information Systems, UIN Raden Intan Lampung, Indonesia
[1]ahmadfarzianwar4@gmail.com, [2]m.husaini@radenintan.ac.id, [3]khairunnitaaulia@radenintan.ac.id

Accepted 19 June 2026
Approved 28 June 2026

***Abstract*— Indonesia's agricultural sector faces challenges due to climate change and limited access to accurate, real-time information on soil conditions. Traditional methods relying on visual observation are inefficient and prone to error, which can reduce crop productivity. This research aims to develop a web-based information system integrated with the Internet of Things (IoT) for real-time monitoring of soil moisture and temperature. The system was developed using a Research and Development (R&D) approach with prototyping methodology. The hardware components include an ESP32 microcontroller, a capacitive soil moisture sensor v1.2, and a DS18B20 digital temperature sensor. Sensor data is transmitted to Firebase Realtime Database and displayed via a web interface built with HTML, CSS, and JavaScript. Field trials conducted in Talang Padang village (OKU Selatan) over 7 days demonstrated the system's ability to provide accurate, real-time monitoring. The system includes real-time gauge visualizations, historical graphs, notification features, and exportable sensor logs. Positive responses from farmers indicate the system's potential to support better agricultural decision-making. However, internet dependency and the need for sensor recalibration remain as limitations.**

***Index Terms*— IoT; soil monitoring; moisture sensor; temperature sensor; web application; precision agriculture.**

I. INTRODUCTION

Agriculture remains one of the most strategic sectors in Indonesia, with approximately 26.15% of its 144.64 million workforce engaged in agricultural, forestry, and fisheries activities, equating to around 37.81 million workers, according to the Central Bureau of Statistics (BPS) as of August 2024. However, the sector faces significant challenges, particularly in optimizing productivity amid climate change, land degradation, and limited access to accurate and real-time information regarding soil conditions. Soil plays a crucial role in agriculture as it serves as the primary medium for plant growth and provides essential nutrients such as nitrogen, phosphorus, potassium, and other microelements. Optimal soil conditions—especially in terms of moisture and temperature—are essential to support healthy and productive crop development. In practice, most farmers still rely on subjective visual observation and traditional methods to assess soil conditions, which are often inefficient and prone to inaccuracy. These limitations can lead to poor decision-making regarding irrigation and fertilization, potentially resulting in reduced crop yields.[1]

To address these problems, technological advancements in information and communication have opened new opportunities for the agricultural sector. One promising innovation is the Internet of Things (IoT), which enables the use of interconnected sensors and devices for real-time data collection and transmission. IoT allows for continuous monitoring of soil conditions, thus supporting precise, data-driven decision-making in agricultural management. This research aims to develop a web-based information system integrated with IoT to monitor soil moisture and temperature in real-time. By leveraging sensor data transmitted through cloud-based services, farmers will be able to access accurate information from anywhere, facilitating more effective and efficient land management.[2]

The research was conducted using a Research and Development (R&D) approach with a prototyping methodology. The development stages included problem identification, data collection through literature review, interviews, and field observations, followed by system design, prototyping, system testing, and final refinement.[3] The system utilizes an ESP32 microcontroller, a capacitive soil moisture sensor (v1.2), and a DS18B20 digital temperature sensor, all connected to Firebase Realtime Database for cloud-based data storage. The web interface, developed using HTML, CSS, and JavaScript, displays real-time soil condition data with graphical visualization and notification features. The system was tested in Talang Padang Village, OKU Selatan, over seven days. Black-box testing confirmed that all system components functioned correctly. Field trial results and user

feedback indicated that the system effectively helped farmers monitor soil conditions and provided valuable decision-making support, despite some limitations such as dependence on internet connectivity and the need for future sensor calibration.

II. Literature Review

A. Internet of Things (IoT) in Agriculture

The agricultural sector in Indonesia is strategic, employing a significant portion of the workforce. However, it faces challenges like climate change and limited access to accurate soil data. Traditionally, farmers rely on visual observation, which is often inefficient and inaccurate, potentially harming crop yields. The Internet of Things (IoT) offers a transformative solution by enabling interconnected sensors to collect and transmit data in real-time. This technology supports precision agriculture by providing continuous, data-driven insights into soil conditions, thereby facilitating more effective land management.[4]

B. Soil Monitoring Systems with Microcontrollers

Previous research has extensively used microcontrollers for soil monitoring. Many systems focused on automated irrigation based on soil moisture. For instance, a system by P. Ariyanto, et al.[5] used an ESP8266 microcontroller to control water flow automatically, demonstrating improved water efficiency. Similarly, R. Sarwansah, et al.[6] developed an automated watering system using a NodeMCU ESP8266 and the Blynk application, but it was limited to monitoring only soil moisture. Other research by J. L. Gaol, et al.[7] resulted in a comprehensive Android-based application called SISMOLA for monitoring multiple farm parameters. However, its reliance on a native mobile application requires installation and lacks the accessibility of a web-based platform. While these studies were foundational, they were often limited by their focus on a single parameter, platform dependency, or a lack of an integrated, user-friendly interface for data analysis.

C. Cloud-Based Data Storage with Firebase

Firebase has emerged as a popular solution for real-time data storage in IoT applications due to its cloud-based nature and fast synchronization capabilities.[8]

D. Connectivity Challenges and Scalable Architectures in Rural Areas

Although many IoT-based agricultural systems have been successfully developed, a primary challenge often arises during implementation in rural areas with limited internet connectivity. To address this issue, Ahmed, et al.[9] proposed a scalable network architecture using a Wi-Fi-based Long Distance (WiLD) network to extend the connectivity range. Furthermore, they integrated the concept of fog computing, where sensor data is processed locally at fog nodes to reduce latency and network load before the aggregated data is sent to the cloud.

E. Research Contribution

While previous studies have addressed various components such as hardware, connectivity, and basic visualization, they often lacked integration of these components into a single cohesive system. This research proposes a complete solution that includes ESP32-based sensing, real-time Firebase data storage, a responsive web interface, dynamic charts (Chart.js), and automatic alerts for critical soil conditions, contributing a more holistic and scalable model for smart farming, particularly suited for rural Indonesian agriculture

III. Method

This research uses a Research and Development (R&D) approach with the prototyping model to build a functional web-based soil monitoring system. The system was developed through several phases, namely: requirements analysis, system architecture design, component integration, interface development, and field testing.[10]

The stages in the Research and Development (R&D) model can be adapted based on the needs and context of the system being developed. In this study, the system development process follows seven main stages, which are combined with a prototyping approach. The stages of this research process are illustrated in Fig. 1.

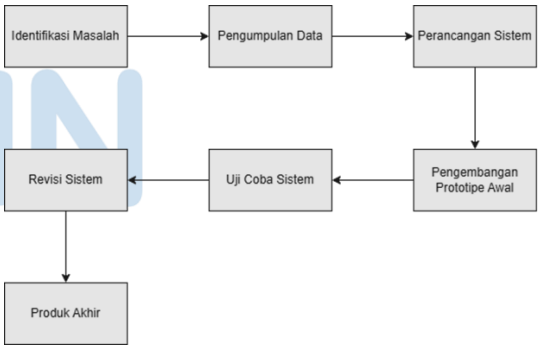


Fig. 1. Research stages

A. Problem Identification

This stage involved observing the farming practices in Talang Padang Village to identify challenges in monitoring soil conditions. Interviews and field visits revealed that farmers relied on manual observations, which were inefficient and often inaccurate. The need for a digital monitoring solution was confirmed through these findings.

B. Data Collection

Literature reviews and scientific references were collected to strengthen the theoretical foundation of the

system design. The review focused on concepts such as IoT, soil sensors, microcontroller integration, Firebase cloud platforms, and web-based interfaces. Previous studies were also examined to identify best practices and limitations in related research.

C. *System Design*

To support a clear understanding of the IoT-based web system for monitoring soil moisture and temperature, this stage focused on technical design using Unified Modeling Language (UML).[11] The design process involved collaboration between programmers and analysts to visualize system workflows and user interactions. Some of the diagrams used include use case diagrams, activity diagram and block diagrams were developed to illustrate system architecture, data flow from sensors to Firebase, and the web interface for real-time data visualization:

The Use Case Diagram is used to illustrate the functional interactions between the user (Actor) and the core features provided by the system. The actor can perform several actions, including logging in, monitoring real-time soil moisture and temperature data, viewing charts and historical data, exporting data to Excel or PDF, and receiving visual and audio notifications.[12]

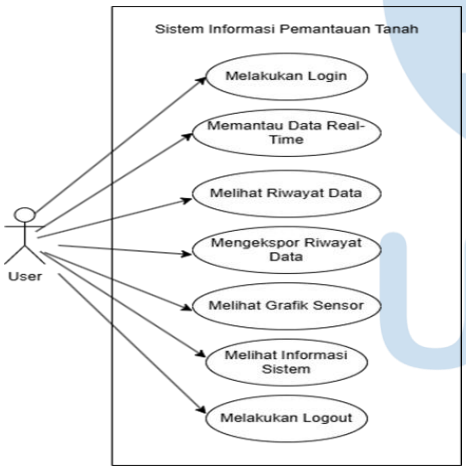


Fig. 2. Use Case Diagram

In addition to the use case diagram, an activity diagram is designed to illustrate the sequential workflow of the proposed system. The activity diagram describes the operational flow starting from sensor data acquisition, data processing by the ESP32 microcontroller, transmission to the Firebase Realtime Database, real-time data visualization on the web interface, and alert generation when predefined threshold values are exceeded. This diagram clarifies system behavior, decision points, and interactions between hardware, cloud services, and users during system operation.

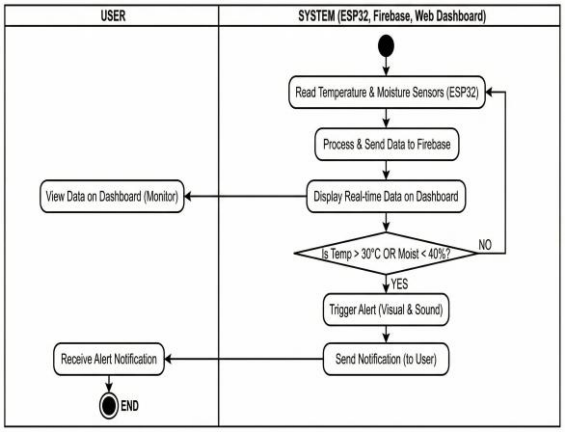


Fig. 3. Activity Diagram

The Block Diagram illustrates the overall system architecture.[13]showing the key hardware components, including the DS18B20 temperature sensor and capacitive soil moisture sensor v1.2, whose data are processed by the ESP32 microcontroller. The processed data are then transmitted via a Wi-Fi network to a cloud-based backend, namely the Firebase Realtime Database, and finally displayed to users through a web application.

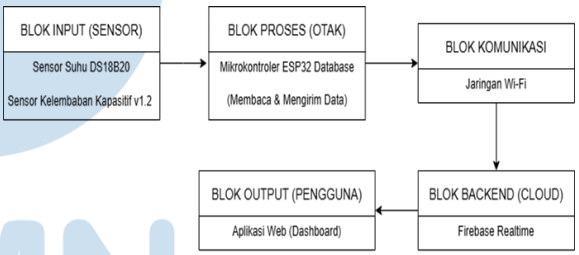


Fig. 4. System Block Diagram

A simple prototype of the IoT-integrated web-based monitoring system was developed, covering hardware, software, and user interface aspects to ensure real-time operation and accessibility, particularly for farmers.[14] Before building the system, a thorough analysis of hardware and software requirements was conducted.

1. *Hardware Requirements Analysis*

The hardware components required include:

a. ESP32 microcontroller (AI Thinker, Shenzhen, China) – functions as the control center and data transmitter.

b. Capacitive soil moisture sensor v1.2 (DFRobot, Shanghai, China) – measures soil moisture levels.

c. DS18B20 waterproof temperature sensor (Maxim Integrated, San Jose, CA, USA) – measures soil temperature.

d. Wi-Fi communication module (integrated with ESP32).

e. Breadboard, jumper wires, and a power supply (power bank or adapter).

2. Software Requirements Analysis

The software was designed to operate and manage the system as follows:

a. Firmware for ESP32, developed using Arduino IDE v1.8.19 (Arduino, Somerville, MA, USA), programmed to read sensor data and send it to Firebase.

b. Firebase Realtime Database (Google LLC, Mountain View, CA, USA; https://firebase.google.com) – serves as a cloud database for storing real-time temperature and moisture data.

c. Web interface developed using HTML5, CSS3, JavaScript, used to display sensor data in readable table and chart formats.

3. Database Development

Firebase Realtime Database was selected for continuous storage of soil moisture and temperature data. It supports real-time communication between the hardware (ESP32) and the web interface, and is easily integrated with microcontroller systems.

4. Web-Based User Interface Design

The interface was designed to be responsive and accessible via mobile devices and laptops. It focuses on simplicity and readability, allowing users (especially farmers) to easily interpret soil condition data. Key metrics such as temperature and moisture levels are presented in both numeric and graphical formats.

D. Initial Prototype Development

This stage involved implementing the planned system using selected hardware components: ESP32, DS18B20 sensor, and capacitive soil moisture sensor. The system was connected to Firebase for cloud storage and integrated with a responsive web dashboard built using HTML, CSS, and JavaScript.

E. System Testing

The prototype was tested in a real agricultural setting to evaluate its functionality. Testing criteria included sensor accuracy, data transmission stability, and user interaction with the web interface. The testing was conducted for seven consecutive days to assess the reliability of both hardware and software components under field conditions.

F. System Revision

Based on feedback from the field tests and technical analysis, the prototype was refined. Revisions focused on improving system stability, optimizing data refresh intervals, enhancing user interface elements, and ensuring robust cloud synchronization.

G. Final Product

The final output is a stable and accurate web-based IoT monitoring system for soil temperature and moisture. The product includes a complete set of features: real-time monitoring, graphical displays, exportable data logs, and notification alerts. Technical documentation was also prepared to support future maintenance and scalability.

IV. Result and Discussions

At this stage, the researcher presents the system interface through several web pages representing the core features, including Login, Dashboard, Real-Time Monitoring, History, and Profile. These pages allow secure access, display live and historical soil data, support data export (Excel/PDF), and show user information. Additional features include graphical charts, real-time alerts (visual/audio), and a logout function for session security.

System testing was conducted using the black-box method to validate the functionality of each feature against expected outputs, without analyzing internal program logic. Furthermore, a User Acceptance Test (UAT) was carried out with target users (local farmers) to assess system usability, readability of the interface.[15] and the reliability of sensor data presentation. The following section provides detailed explanations of each interface feature developed in the system.

A. Login Page

The system provides a login page where users enter their email and password, which are securely authenticated using Firebase Authentication. This ensures that only authorized users can access the system and its monitoring features.

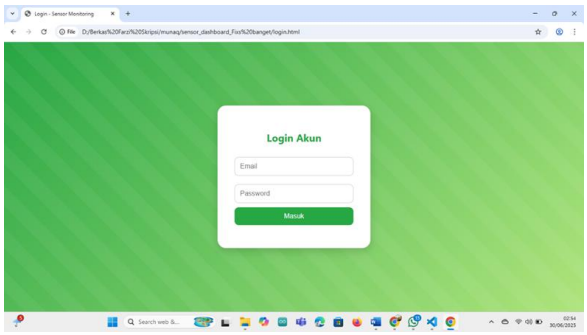


Fig. 5. Login Page

B. *Dashboard Page*

After a successful login, users are directed to the dashboard, which displays a summary of the latest soil temperature and moisture data. It also provides quick navigation links to other pages such as real-time monitoring, historical data, and profile settings.

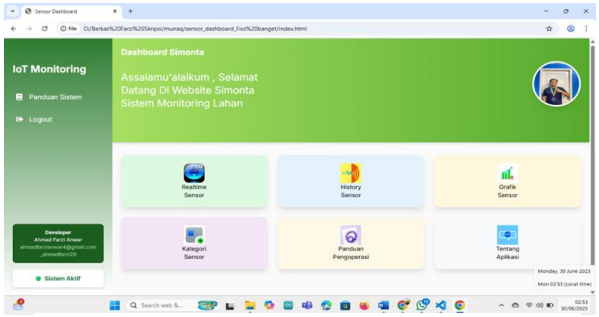


Fig. 6. Dashboard Page

C. *Real-Time Monitoring Page*

This page displays live sensor data that is updated every 10 seconds, including current temperature and soil moisture readings. It also includes graphical charts (using Chart.js) that visualize fluctuations over time. When sensor values exceed defined thresholds, the system triggers real-time alerts, both visual (on-screen pop-ups) and audio notifications.

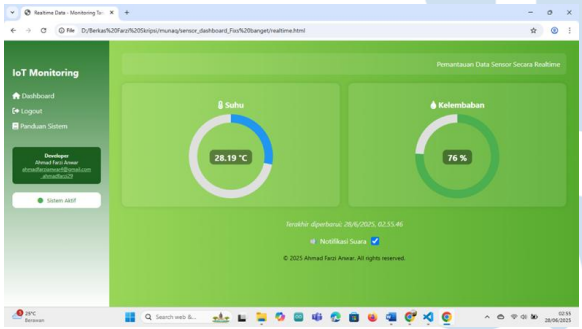


Fig. 7. Real-Time Monitoring Page

D. *History Page*

The history page provides a comprehensive table of previously recorded soil temperature and moisture data. Users can search and filter the information by date and time, making it easier to review specific entries and analyze environmental trends over time. To support data preservation and reporting needs, the system also offers export functionality, allowing users to download the historical data in both Excel (.xlsx) and PDF formats. These features enable farmers and agricultural practitioners to store sensor records offline, share them with relevant stakeholders, and utilize the data for planning, documentation, and decision-making purposes.

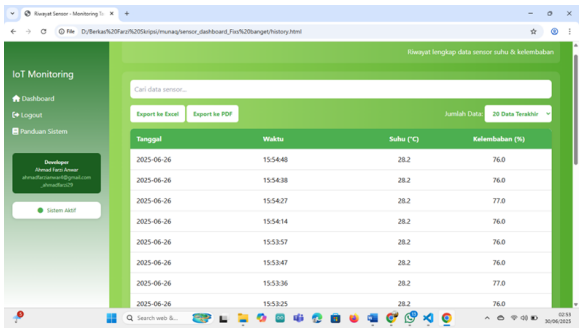


Fig. 8. History Page

E. *Sensor Data Visualization*

Interactive graphs are integrated into the real-time and history pages using Chart.js to visually represent sensor data. These graphs help users identify environmental trends and fluctuations in soil conditions.

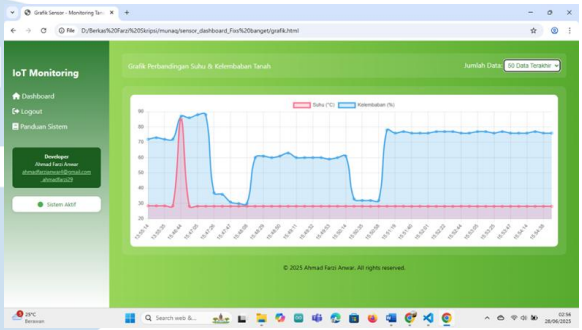


Fig. 9. Sensor Data Visualization

F. *Sensor Categories*

Sensor values are categorized into clear labels such as “Normal,” “Dry,” or “Hot” to aid user interpretation. This classification enhances decision-making by giving immediate insight into the condition of the soil.

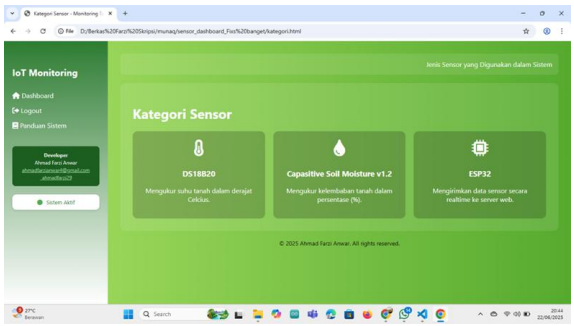


Fig. 10. Sensor Categories

G. *User Guide Section*

An embedded guide provides instructions on how to use the system, including how to log in, read real-time data, respond to alerts, and export information. This ensures accessibility for first-time users, especially farmers.

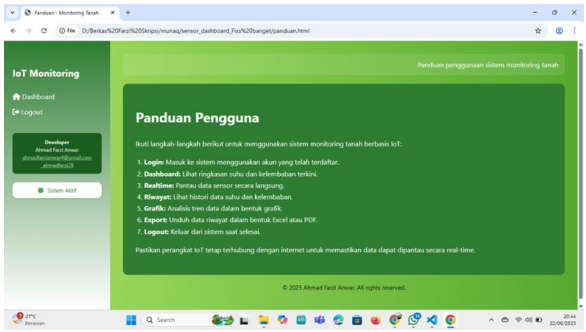


Fig. 11. User Guide Section

H. *About Application Page*

This page outlines the background and objectives of the system. It explains how the system was developed, its core functions, and its intended benefits for agricultural monitoring in rural areas.

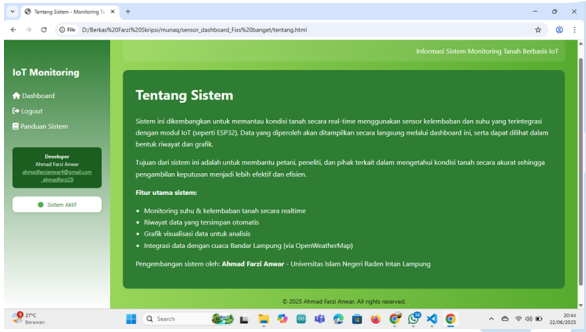


Fig. 12. About Application Page

I. *Profile Page*

The profile page displays basic user information, including name, email, Instagram handle, and contact number. This feature personalizes the system and verifies user identity.

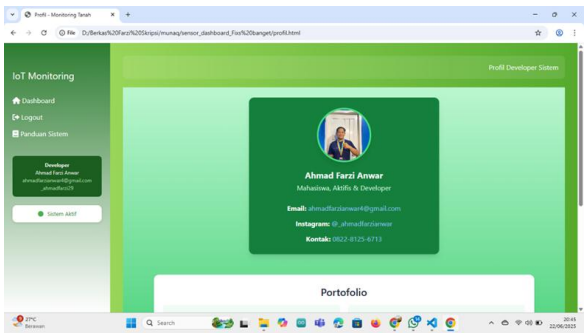


Fig. 13. Profile Page

J. *Field Testing and User Evaluation*

TABLE I User Acceptance Testing Results

No.	Test Item	Success Criteria	Test Outcome	Notes
1	Login Functionality	User Successfully Logs in Via Firebase	✓ Passed	All Users Authenticated Without Errors
2	Real-Time Data Display	Data Updates Every 10 Seconds	✓ Passed	Sensor Values Shown Live on Web Interface
3	Notification System	Alerts Triggered If Threshold Exceeded	✓ Passed	Both Visual and Sound Alerts Functioned
4	History Viewing	Users Can Access and Scroll Through Past Records	✓ Passed	Date-Based Filtering Works Correctly
5	Export Feature	Data Export to Excel and PDF Without Error	✓ Passed	Output Files Readable and Properly Formatted
6	Graph Visualization	Users Can View Data Charts Clearly	✓ Passed	Charts Interpreted Easily by All Users
7	PROFILE PAGE	Displays Full User Identity	✓ Passed	Profile Info Accurate and Complete
8	Logout Functionality	Ends Session and Returns to Login Page	✓ Passed	Logout Process Secure and Functional
9	Ease Of Use (Usability)	Non-Tech Users Complete Task Without Assistance	✓ Passed	Interface Considered User-Friendly
10	Overall Satisfaction	User Finds System Useful and Practical	✓ Passed	Viewed As Beneficial for Agriculture

Table 1 summarizes the results of User Acceptance Testing (UAT) based on 10 main criteria. Each test was evaluated by participants during a 7-day trial. All system features—login, real-time monitoring, alerts, history view, data export, and profile access—met the success criteria. Users highly appreciate the intuitive interface and timely notifications that support farming decision making.

K. *Sensor Testing Results*

Table 2 presents representative sensor readings obtained during field testing using an ESP32 microcontroller, a DS18B20 temperature sensor, and a

capacitive soil moisture sensor. The system effectively detected environmental changes and reliably triggered alerts whenever the temperature exceeded 30°C and/or the soil moisture level dropped below 40%. These findings demonstrate both the accuracy and responsiveness of the sensors and validate the overall effectiveness and reliability of the system under actual agricultural field conditions. Such performance confirms the system's potential to support real-time soil monitoring and informed decision-making in precision farming.

TABLE II Soil Sensor Testing Results

No.	Timestamp	Sensor Data	Environmental Condition	Alert Status
		Temperature (°C)	Moisture (%)	
1	2025-06-15 08:00:00	29.5	37	Dry and Hot
2	2025-06-15 08:00:10	30.0	35	Dry and Hot
3	2025-06-15 08:00:20	31.2	33	Dry and Hot
4	2025-06-15 08:00:30	28.7	42	Normal
5	2025-06-15 08:00:40	27.5	49	Normal
6	2025-06-15 08:00:50	29.1	70	Wet but Normal

L. System Evaluation

The system evaluation demonstrates that the IoT-based soil monitoring system functions effectively in real-time conditions. It accurately collects and transmits soil temperature and moisture data every 10 seconds using ESP32, DS18B20, and capacitive sensors, with the data stored in Firebase and displayed through a responsive web interface. User Acceptance Testing confirmed that all core features such as real-time monitoring, historical data access, alert notifications, data export, and user authenticationworked as expected and were well received by users. The system was stable during field trials, easy to use, and provided timely alerts, although it remains limited to a small-scale test area and single-node deployment. Overall, the system meets its functional goals and shows strong potential for supporting precision agriculture.

M. Limitations of the Study

Although the proposed IoT-based soil monitoring system demonstrated reliable performance during field implementation, this study has several limitations. The system testing was conducted over a relatively short period of seven days and was limited to a single agricultural location in Talang Padang Village, OKU Selatan. As a result, the evaluation may not fully represent the system's long-term stability and performance under diverse environmental conditions. Factors such as seasonal climate variations, different soil characteristics, and extended operational duration were not comprehensively assessed. Therefore, the findings should be interpreted within the scope of short-term and single-location deployment

V. CONCLUSIONS

This research concludes that a web-based soil condition monitoring system integrated with IoT technology has been successfully developed and implemented. The system utilizes an ESP32 microcontroller along with a DS18B20 temperature sensor and a capacitive soil moisture sensor v1.2 to collect data every 10 seconds, which is automatically transmitted to Firebase Realtime Database. The front-end application, developed using HTML, CSS, JavaScript, and PHP, is capable of displaying real-time data, storing historical records, and providing essential features such as graphical visualization, export to Excel and PDF, user profile display, and secure login/logout. The system also includes a notification mechanism that triggers alerts when temperature exceeds 30°C or soil moisture drops below 40%, allowing timely responses to adverse conditions. User Acceptance Testing (UAT) confirmed that all functionalities operated as expected and were well received by end users, including farmers and field observers, who found the interface easy to use and the information provided both accurate and useful. The sensor testing further validated the accuracy and reliability of the system, demonstrating its potential as an effective digital solution to support real-time agricultural decision-making.

For future work, it is recommended to conduct long-term testing over extended periods and deploy the system in multiple agricultural locations with different soil types and climate conditions to further evaluate system stability and scalability. Future enhancements may include the integration of additional environmental sensors, implementation of adaptive sensor calibration techniques, and optimization of data processing for larger-scale deployments. These improvements are expected to enhance system

reliability and strengthen its applicability for precision agriculture and sustainable farming practices.

ACKNOWLEDGMENT

The author would like to express his deepest gratitude to the Raden Intan Lampung State Islamic University, especially to the Faculty of Science and Technology and the Information Systems Study Program for providing guidance and academic facilities during the completion of this research. Special appreciation is given to the supervisors who continue to provide support, valuable input and constructive input which greatly contributes to the quality and success of this research. The author also would like to thank village officials and farmers in Talang Padang, South OKU, for their cooperation during the field tests. Without their involvement and encouragement, research on developing IoT-based web applications for soil temperature and moisture monitoring would not have been possible.

REFERENCES

[1] L. Kamelia, Y. S. Nugraha, M. R. Effendi, dan T. Priatna, "The IoT-Based Monitoring Systems for Humidity and Soil Acidity Using Wireless Communication," dalam *2019 IEEE 5th International Conference on Wireless and Telematics (ICWT)*, Yogyakarta, Indonesia: IEEE, Jul 2019, hlm. 1–4. doi: 10.1109/ICWT47785.2019.8978243.

[2] Md. N. Mowla, N. Mowla, A. F. M. S. Shah, K. M. Rabie, dan T. Shongwe, "Internet of Things and Wireless Sensor Networks for Smart Agriculture Applications: A Survey," *IEEE Access*, vol. 11, hlm. 145813–145852, 2023, doi: 10.1109/ACCESS.2023.3346299.

[3] P. P. Kuantitatif, "Metode Penelitian Kunatitatif Kualitatif dan R&D," *Alf. Bdg.*, 2016.

[4] G. Ramadevi, G. Ajay Sankar, dan M. K. Singh, "Role of IoT in Intelligent Agriculture Network System," dalam *Advances in Transdisciplinary Engineering*, R. M. Singari dan P. K. Kankar, Ed., IOS Press, 2022. doi: 10.3233/ATDE220745.

[5] P. Ariyanto, A. Iskandar, dan U. Darusalam, "Rancang Bangun Internet of Things (IoT) Pengaturan Kelembaban Tanah untuk Tanaman Berbasis Mikrokontroler," *J. JTIK J. Teknol. Inf. Dan Komun.*, vol. 5, no. 2, hlm. 112, Apr 2021, doi: 10.35870/jtik.v5i2.211.

[6] R. Sarwansah, U. Jaelani, A. Hasad, S. Supratno, dan Sugeng, "Aplikasi NodeMCU ESP8266 Untuk Monitoring Kelembaban Tanah Berbasis Internet of Things," *J. Stud. Res. Comput. Sci.*, vol. 3, no. 1, hlm. 63–72, Mei 2022, doi: 10.31599/jsrcs.v3i1.1174.

[7] J. Lumban Gaol *dkk.*, "Aplikasi Android untuk Monitoring Lahan Pertanian secara Realtime Berbasis Internet of Things," *J. Tek. Inform. Dan Sist. Inf.*, vol. 6, no. 3, Des 2020, doi: 10.28932/jutisi.v6i3.3039.

[8] "Firebase Realtime Database," Firebase. Diakses: 2 Agustus 2025. [Daring]. Tersedia pada: https://firebase.google.com/docs/database

[9] N. Ahmed, D. De, dan I. Hussain, "Internet of Things (IoT) for Smart Precision Agriculture and Farming in Rural Areas," *IEEE Internet Things J.*, vol. 5, no. 6, hlm. 4890–4899, Des 2018, doi: 10.1109/JIOT.2018.2879579.

[10] S. Sugiono, "Metode penelitian kuantitatif, kualitatif, dan r & d," *Bdg. Alf.*, vol. 288, 2016.

[11] S. Setiaji dan R. Sastra, "Implementasi Diagram UML (Unified Modelling Language) Pada Perancangan Sistem Informasi Penggajian," *J. Tek. Komput.*, vol. 7, no. 1, hlm. 106–111, Feb 2021, doi: 10.31294/jtk.v7i1.9773.

[12] L. Setiyani, "Desain Sistem : Use Case Diagram," 2021.

[13] T. Hadyanto dan M. F. Amrullah, "Sistem Monitoring Suhu dan Kelembaban pada Kandang Anak Ayam Broiler Berbasis Internet of Things," *J. Teknol. Dan Sist. Tertanam*, vol. 3, no. 2, Agu 2022, doi: 10.33365/jtst.v3i2.2179.

[14] A. P. Selvam dan S. N. S. Al-Humairi, "The Impact of IoT and Sensor Integration on Real-Time Weather Monitoring Systems: A Systematic Review," 10 November 2023, *In Review*. doi: 10.21203/rs.3.rs-3579172/v1.

[15] B. Priyatna, A. L. Hananto, dan M. Nova, "Application of UAT (User Acceptance Test) Evaluation Model in Minggon E-Meeting Software Development," vol. 2, no. 3, 2020.

ISSN 2085-4552

Fuzzy Expert System for Early Heart Disease Diagnosis Using Mamdani Method in Web-Based System

Ananda Dwi Rizkyta[1], Muhamad Bahrul Ulum[2*], Fenina Adline Twince Tobing[3]
[1,2] Informatics Engineering, Universitas Esa Unggul, Jakarta, Indonesia
[3] Informatics Department, Universitas Multimedia Nusantara, Tangerang, Indonesia
[1] anandadwir898@student.esaunggul.ac.id, [2] m.bahrul_ulum@esaunggul.ac.id, [3]fenina.tobing@umn.ac.id

Accepted 21 June 2026
Approved 28 June 2026

***Abstract*— Heart disease remains one of the leading causes of mortality in Indonesia and is often termed a "silent killer" due to its frequently asymptomatic progression in early stages. The limited availability of cardiologists and restricted access to specialized healthcare facilities significantly hinder timely diagnosis. Data recorded at An-Nisa Hospital Tangerang in 2024 documented 16,818 outpatient visits to the cardiology unit and 53 fatalities attributable to cardiovascular disease. This study presents the design and implementation of a web-based expert system for early heart disease diagnosis using the Fuzzy Mamdani inference method. The system incorporates seventeen input variables, encompassing demographic factors, Body Mass Index (BMI), blood pressure readings, smoking history, disease history, psychological aspects, and eight cardinal symptoms validated by a practicing cardiologist. Gaussian membership functions were employed to model each linguistic variable, and an 18-rule inference base was constructed to map symptom combinations to risk levels. Defuzzification was performed using the centroid method. System accuracy was evaluated against ten patient records from An-Nisa Hospital Tangerang, yielding an accuracy of 80%. Black-box functional testing confirmed that all system features operated within specification. The proposed system is intended to serve as a preliminary screening tool that promotes early awareness, facilitates informed health-seeking behavior, and reduces dependence on specialist consultations for initial risk stratification.**

***Index Terms*— Cardiovascular Risk Assessment; Defuzzification; Gaussian Membership Function; Medical Informatics; Rule-Based Interface; Uncertainty Handling.**

I. INTRODUCTION

Cardiovascular disease (CVD) continues to represent one of the most significant public health burdens globally, accounting for approximately 17.9 million deaths annually according to the World Health Organization [1]. In Indonesia, heart disease has consistently ranked among the primary causes of death, with increasing incidence attributed to lifestyle changes, urbanization, and an aging population. An-Nisa Hospital Tangerang recorded 53 cardiac-related deaths and more than 16,000 outpatient visits to its cardiology department in 2024 alone, underscoring the urgent need for accessible diagnostic tools capable of facilitating early detection at the community level.

A fundamental challenge in combating cardiovascular disease lies in its insidious onset. Commonly referred to as a "silent killer," heart disease frequently manifests without clear or overt symptoms during its early stages, making self-assessment exceedingly difficult for lay individuals [2]. Risk factors such as hypertension, obesity, smoking, sedentary behavior, and psychological stress are well-documented contributors to cardiac morbidity, yet their complex and interdependent relationships make deterministic rule-based diagnosis inadequate in capturing the full spectrum of clinical presentations [3]. Moreover, the limited number of cardiologists relative to the Indonesian population exacerbates delays in diagnosis, particularly in peripheral and underserved regions.

Advancements in artificial intelligence and decision-support technologies have opened promising avenues for addressing this diagnostic gap. Expert systems, which encode domain-specific knowledge within structured rule bases to emulate the reasoning processes of human specialists, have been applied with considerable success across various medical domains [4]. When combined with fuzzy logic, expert systems can effectively manage the inherent uncertainty and imprecision present in clinical data, thereby producing diagnostic outputs that more closely reflect the nuanced judgment of experienced clinicians [5].

Among the available fuzzy inference approaches, the Mamdani method is particularly well-suited for medical diagnosis applications due to its interpretable linguistic output structure and its capacity to integrate both quantitative and qualitative input variables [6]. Prior studies have demonstrated that Mamdani-based fuzzy expert systems can achieve competitive

diagnostic accuracy for heart disease. Simanjorang et al. [7] reported an accuracy of 90% using centroid defuzzification, while Nuraeni [8] achieved a system feasibility score of 85.6%. In contrast, Tsukamoto-based systems, while computationally efficient, rely on monotonic membership functions and produce scalar outputs that are inherently less transparent and harder to interpret in a clinical context [9].

This study contributes to the existing body of knowledge by developing a web-based Mamdani fuzzy expert system for early heart disease diagnosis, grounded in symptom data validated directly by a practicing cardiologist at An-Nisa Hospital Tangerang. Unlike prior work that largely relied on generic clinical datasets or laboratory values, this system is designed for independent use by the general public, requiring only self-reported symptoms and basic physiological measurements. The system employs Gaussian membership functions across all input variables and applies an 18-rule inference engine to classify cardiac risk into four categories: not detected, low, moderate, and high. The remainder of this paper is organized as follows: Section II details the proposed methodology; Section III presents results and comparative discussion; and Section IV provides conclusions and directions for future research.

II. METHOD

A. Research Framework and Dataset

This research was conducted using a structured design-and-development methodology guided by the Extreme Programming (XP) agile framework, which consists of four iterative phases: Planning, Design, Coding, and Testing [10]. The selection of XP was motivated by its adaptability to evolving requirements and its suitability for small development teams operating under time constraints. The research workflow commenced with an identification of the problem domain, followed by literature review, expert consultation, data collection, fuzzy system design, software development, and rigorous evaluation.

Primary data were collected through a structured interview conducted with dr. Achmad Ismail Putra, BMedSci, Sp.JP, a cardiologist practicing at An-Nisa Hospital Tangerang, on March 19, 2025. The interview aimed to elicit expert knowledge regarding the symptom profiles, risk factor configurations, and clinical decision thresholds associated with early-stage heart disease. Secondary data comprising ten anonymized outpatient medical records from the hospital's cardiology polyclinic were subsequently used to evaluate the diagnostic accuracy of the developed system. All patients in the evaluation dataset had confirmed cardiac diagnoses established by the attending physician, thereby enabling a direct comparison between system outputs and clinical ground truth.

The complete set of input and output variables incorporated into the system is summarized in Table I. Input variables span demographic attributes, physiological measurements, lifestyle factors, psychological status, and self-reported symptomatic complaints. The output layer generates four diagnostic components: a binary detection status, a continuous risk percentage derived through defuzzification, a categorical risk level, and a corresponding medical recommendation.

TABLE I. SYSTEM INPUT AND OUTPUT VARIABLES

Type	Variable	Category / Range
Input	Age	Infant, Child, Adolescent, Adult, Elderly
	Body Mass Index (BMI)	Underweight (<18.5), Normal (18.5–24.9), Overweight (25–29.9), Obese (≥30)
	Systolic Blood Pressure	Low (<90), Normal (90–119), High (120–139), Very High (≥140) mmHg
	Diastolic Blood Pressure	Low (<60), Normal (60–79), High (80–89), Very High (≥90) mmHg
	Disease History	Present, Absent
	Smoking History	Yes, No
	Psychological Aspect	Depression, Anxiety, Suicidal Tendency, Fear, Anger, Calm
	Chest Pain (G01)	Yes / No
	Shortness of Breath (G02)	Yes / No
	Dizziness (G03)	Yes / No
	Fatigue / Weakness (G04)	Yes / No
	Palpitations (G05)	Yes / No
	Easy Fatigability (G06)	Yes / No
	Leg Edema (G07)	Yes / No
	Cold Sweat (G08)	Yes / No
Output	Detection Status	Potential Heart Problem Detected / Not Detected
	Risk Percentage	0–100% (continuous, via centroid defuzzification)
	Risk Level	Low (0–40%), Moderate (41–70%), High (>70%)
	Medical Recommendation	S01 (maintain health) / S02 (monitor) / S03 (consult specialist) / S04 (urgent referral)

B. Fuzzy Mamdani Interface Method

The Mamdani fuzzy inference system, originally introduced by Ebrahim Mamdani in 1975, constitutes the computational engine of the proposed expert system

[6]. This method was selected over alternative approaches specifically the Tsukamoto and Sugeno methods on the basis of three principal advantages. First, it produces linguistic output distributions that are inherently more interpretable and clinically meaningful compared to the scalar outputs of Tsukamoto or Sugeno systems [11]. Second, it accommodates a wider variety of membership function shapes, thereby offering greater flexibility in modeling complex and non-monotonic medical variables. Third, it has been empirically validated across multiple prior studies in cardiac diagnosis, consistently demonstrating high diagnostic accuracy [7], [8]. The Tsukamoto method, while computationally efficient, restricts membership functions to monotonically increasing or decreasing forms, which limits its ability to represent variables with peak or bell-shaped distributions such as blood pressure and BMI. The four sequential stages of the Mamdani method are described below.

- Fuzzification

Fuzzification is the process of converting crisp input values into fuzzy membership degrees through the application of predefined membership functions. In this study, Gaussian membership functions were exclusively employed for all numerical input variables, expressed by the formula:

$$\mu(x) = \exp\left(-\frac{(x-c)^2}{2\sigma^2}\right) \quad (1)$$

where x denotes the crisp input value, c represents the mean (center) of the distribution, and σ is the standard deviation controlling the spread of the bell curve. Gaussian functions were preferred over triangular or trapezoidal alternatives for three reasons: (1) they provide smooth, continuous transitions between adjacent fuzzy sets, thereby avoiding abrupt discontinuities at set boundaries that may cause classification instability; (2) they are mathematically differentiable, facilitating future optimization using gradient-based techniques; and (3) their bell-shaped form naturally captures the probabilistic behavior of physiological measurements [12].

The specific Gaussian parameters for the continuous numerical variables are presented in Table II. For discrete binary variables (disease history, smoking history, individual symptoms), crisp binary encoding was applied: a value of 1.0 indicates the presence of the attribute, and 0.0 indicates its absence. Psychological status was mapped to a severity weight scale from 0.0 (calm) to 1.0 (suicidal tendency) based on clinical significance.

TABLE II. GAUSSIAN MEMBERSHIP FUNCTION PARAMETERS FOR NUMERICAL VARIABLES

Variable	Fuzzy Set	Center (c)	Std. Dev. (σ)
BMI (kg/m²)	Underweight	16.0	2.0
	Normal	21.7	2.5
	Overweight	27.5	2.0
	Obese	35.0	3.5
Systolic BP (mmHg)	Low	70.0	15.0
	Normal	110.0	10.0
	High	130.0	8.0
	Very High	165.0	15.0
Diastolic BP (mmHg)	Low	50.0	8.0
	Normal	70.0	7.0
	High	85.0	5.0
	Very High	100.0	8.0

- Rule Base Construction

The knowledge base of the system comprises 18 fuzzy IF-THEN rules, formulated through expert consultation and validated iteratively against clinical records. All rules employ the AND operator, meaning that all specified antecedent conditions must be satisfied simultaneously for a rule to fire. The rule strength (firing strength) for each rule was calculated as the minimum membership value across all active antecedents, following the standard Mamdani min-AND aggregation approach. Table III presents a representative subset of the rule base, categorized by output risk level.

TABLE III. REPRESENTATIVE SUBSET OF FUZZY IF-THEN RULE BASE

ID	Rule Antecedent (IF Conditions)	Output
R01	G01 (Chest Pain) AND G08 (Cold Sweat)	High Risk
R02	G01 AND Age = Adult/Elderly	High Risk
R03	G01 AND G02 (Shortness of Breath)	High Risk
R04	Disease History = Present AND (G01 OR G02)	High Risk
R05	Blood Pressure = Very High AND any symptom present	High Risk
R06	G02 (Shortness of Breath) AND G07 (Leg Edema)	High Risk
R07	G01 AND G06 (Easy Fatigability)	Moderate
R08	Smoking = Yes AND BMI = Overweight/Obese AND BP = High/Very High AND any symptom	Moderate
R09	BMI = Obese AND BP = High AND (G07 OR G02)	Moderate
R10	G05 (Palpitations) AND Psychological = Severe AND (Smoking = Yes OR BP = High)	Moderate
R11	≥3 mild symptoms present, G01 and G02 absent	Moderate
R12	Age = Elderly AND (G03 OR G04) AND any risk factor	Moderate

R13	(G03 OR G04) AND Age = Adult AND BMI = Normal AND No disease history	Low
R14	Only 1 mild symptom AND BP = Normal AND No disease history	Low
R15	Only G05 AND Psychological = Severe AND no severe symptoms	Low
R16	Only G06 AND active symptoms <2 AND BP = Normal AND BMI = Normal	Low
R17	No symptoms AND any risk factor present	Low
R18	No symptoms AND no risk factors	Not Detected

- Fuzzy Inference Engine

Following fuzzification, each of the 18 rules was evaluated against the current input state. The firing strength α_i for rule i employing the AND operator was calculated as:

$$\alpha_i = min(\mu A_1(x_1), \mu A_2(x_2), \ldots, \mu A_n(x_n)) \quad (2)$$

where $\mu A_j(x_j)$ denotes the membership degree of input x_j in fuzzy set A_j for rule i. The consequent fuzzy set for each activated rule was clipped to its firing strength using Mamdani minimum implication. All activated consequent sets sharing the same output domain were aggregated through a point-wise maximum operation to yield the composite output fuzzy distribution, expressed as:

$$\mu_{output}(z) = max\left(\alpha_1 \mu B_1(z), \alpha_2 \mu B_2(z), \ldots, \alpha_n \mu B_n(z)\right) \quad (3)$$

The output domain was defined over the continuous interval $[0, 100]$, where 0 represents minimal cardiac risk and 100 represents maximal cardiac risk. Three overlapping Gaussian output sets Low ($c = 20, \sigma = 12$), Moderate ($c = 55, \sigma = 12$), and High ($c = 85, \sigma = 10$) were defined on this domain to capture the full spectrum of risk categories.

- Defuzzification

The aggregated fuzzy output distribution was converted into a single crisp value z^* using the centroid (Center of Area) defuzzification method, which identifies the center of mass of the composite output set:

$$z^* = \frac{\sum[x \cdot \mu'(x)]}{\sum[\mu'(x)]} \quad (4)$$

where z^* is the crisp output (risk percentage), x iterates over all points in the discretized output domain $[0, 100]$, and $\mu'(x)$ is the aggregate membership degree at each point after clipping and maximum aggregation. The centroid method was selected because it considers the entire shape of the aggregated output distribution, thereby reflecting contributions from all activated rules. This property makes it more robust and sensitive to multi-rule activations than alternatives such as the Mean of Maximum (MOM) method [7]. The resulting z^* value was subsequently mapped to a categorical risk level: $z^* \leq 40\% \rightarrow$ Low, $41\% < z^* \leq 70\% \rightarrow$ Moderate, and $z^* > 70\% \rightarrow$ High. A detection threshold was also applied: if no rule beyond R18 was activated, the output was classified as "Not Detected."

C. System Architecture

The web-based expert system was implemented using a three-tier client-server architecture. The frontend interface was developed using React.js, a component-based JavaScript library offering high-performance rendering and responsive user experience across desktop and mobile devices [13]. Tailwind CSS was employed for utility-first styling, ensuring visual consistency and rapid UI development. The application backend was built with Flask, a lightweight Python micro-framework well-suited for rapid API development and straightforward integration with scientific computing libraries [14]. The fuzzy inference engine was implemented entirely in Python, leveraging the NumPy library for numerical computation. Diagnostic results and user data were persisted in a MySQL relational database, accessed through SQLAlchemy as the object-relational mapping (ORM) layer.

Data flow within the system proceeds as follows: the user submits a completed diagnostic form through the React.js frontend; the form data is serialized as a JSON payload and transmitted via HTTP POST to the Flask backend API; the backend deserializes the payload, invokes the Mamdani fuzzy inference engine, and returns a structured JSON response containing the detection status, risk percentage, risk category, and medical recommendation; the frontend then renders the diagnostic result on the result page. Administrative functions including access to aggregated diagnostic statistics, individual user records, and user feedback are secured behind session-based authentication and accessible exclusively through the admin dashboard.

D. System Evaluation

System evaluation was conducted in two complementary stages. The first stage assessed diagnostic accuracy using ten patient records obtained from the cardiology outpatient unit of An-Nisa Hospital Tangerang, all of whom had confirmed cardiac diagnoses. The symptom profiles and risk factor configurations from each record were entered into the system, and the resulting outputs were compared against the clinical ground truth established by the attending cardiologist. Accuracy was computed as the proportion of cases in which the system output (Detected/Not Detected) correctly matched the clinical diagnosis:

$$Accuracy = \left(\frac{Number\ of\ Correct\ Diagnoses}{Total\ Test\ Cases}\right) x\ 100\% \quad (5)$$

It must be noted that the current sample size of ten cases, while representative of available validated records, limits the generalizability of this accuracy metric. The confusion matrix for the binary classification task (Detected vs. Not Detected) is presented in the Results section, alongside precision, recall, and specificity measures. The second evaluation stage employed functional black-box testing across all system features to verify that each module performed in accordance with its specified functional requirements.

III. Result and Discussions

A. Diagnostic Accuracy Evaluation

Table IV presents the detailed results of the accuracy evaluation conducted on ten outpatient records from An-Nisa Hospital Tangerang. All ten patients had confirmed cardiac diagnoses. The system correctly classified eight of the ten cases as "Detected," yielding an overall accuracy of 80%. Two cases (patients 6 and 9) were classified as "Not Detected" by the system, representing false negative outcomes. In both instances, the patients reported no active symptoms during their follow-up visit, despite having established cardiac conditions. Since the system relies exclusively on current self-reported symptoms rather than longitudinal physiological monitoring, it was unable to infer cardiac risk in the absence of symptomatic input.

TABLE IV. Diagnostic Accuracy Test Results

No	Age	BMI	Sys	Dia	Hist	Smk	Symp	Pred	Actual
1	69	36.89	124	64	No	Yes	G02	✓	✓
2	66	26.0	126	61	No	No	G01, G02	✓	✓
3	44	27.0	130	80	No	No	G01, G02, G03	✓	✓
4	59	31.0	160	90	Yes	Yes	G01, G02, G05	✓	✓
5	52	27.0	137	80	Yes	Yes	G02, G03, G06	✓	✓
6*	76	22.0	151	75	No	Yes	None	✗	✓
7	47	23.0	113	75	Yes	No	G02, G05	✓	✓
8	71	22.0	139	74	No	No	G04, G06	✓	✓
9*	46	41.0	118	80	No	Yes	None	✗	✓
10	55	24.0	123	82	No	No	G01, G02, G07	✓	✓
** False Negative cases (no active symptoms at visit)*								**Accuracy**	**80%**

The binary classification performance was further analyzed using a confusion matrix. Since all ten patients were clinically confirmed as positive (Detected), the matrix reflects only true positive (TP) and false negative (FN) outcomes. Eight cases were correctly identified as TP, while two were misclassified as FN. No true negative (TN) or false positive (FP) cases were present in this sample due to the homogeneous clinical composition of the test set. The resulting performance metrics are presented in Table V.

TABLE V. Confusion Matrix and Performance Metrics

Metric	Formula	Value
True Positive (TP)	Correctly detected as cardiac	8
False Negative (FN)	Cardiac cases missed by system	2
True Negative (TN)	Non-cardiac correctly identified	0 (N/A)
False Positive (FP)	Non-cardiac incorrectly flagged	0 (N/A)
Accuracy	(TP+TN)/(TP+TN+FP+FN)	80.0%
Sensitivity (Recall)	TP/(TP+FN)	80.0%
Precision	TP/(TP+FP)	100.0%
Specificity	TN/(TN+FP)	N/A (no TN in sample)

B. Comparative Analysis with Prior Studies

The 80% accuracy achieved by this system merits contextual analysis relative to prior published work in fuzzy-based cardiac diagnosis, as summarized in Table VI. Simanjorang et al. [7] attained 90% accuracy using a Mamdani system with centroid defuzzification, while Nuraeni [8] reported 85.6% feasibility. The Tsukamoto-based system proposed by Athiyah et al. [9] achieved 83% accuracy using 30 test cases, and the SPK system by Astriratma et al. [15] reported 64%

accuracy with an 89.83% sensitivity. The current system's accuracy of 80% is therefore consistent with, and in several cases superior to, comparable prior implementations.

TABLE VI. COMPARATIVE ANALYSIS WITH RELATED STUDIES

Study	Method	Test Cases	Accuracy / Score	Platform
Simanjorang et al. [7] (2024)	Fuzzy Mamdani (Centroid)	Not specified	90%	Web
Simanjorang et al. [7] (2024)	Fuzzy Mamdani (MOM)	Not specified	85%	Web
Nuraeni [8] (2023)	Fuzzy Mamdani	Not specified	85.6% (feasibility)	Web
Athiyah et al. [9] (2021)	Fuzzy Tsukamoto	30	83%	Web (Flask)
Astriratma et al. [15] (2020)	Fuzzy Logic (SPK)	Not specified	64% / Sensitivity 89.83%	Web
This Study (2025)	Fuzzy Mamdani (Centroid)	10	80% / Precision 100%	Web (React +Flask)

A distinguishing characteristic of the present study lies in the specificity of its validation dataset: unlike prior systems that employed generic clinical benchmarks, this system was validated against hospital records from the specific institution whose cardiologist co-developed the rule base. This co-validation approach ensures strong alignment between expert knowledge and evaluation criteria, contributing to the high precision (100%) observed in this study. The two false negatives were associated with asymptomatic presentations during follow-up visits rather than de novo undetected disease, suggesting that the system's sensitivity limitation is fundamentally a consequence of its input modality relying on self-reported current symptoms rather than an intrinsic deficiency in the inference logic.

Compared with Simanjorang et al. [7], the 10-percentage-point gap in accuracy (80% vs. 90%) may be attributable to a smaller test sample and the absence of laboratory-derived inputs such as cholesterol and glucose levels, which were incorporated in some prior studies. However, the present system intentionally excludes laboratory values to maintain accessibility for self-screening without requiring medical equipment.

C. System Strengths and Limitations

The primary strength of the proposed system lies in its accessibility and user-centric design. By requiring only self-reported symptoms and readily measurable physiological parameters (weight, height, and blood pressure obtainable from community health posts), the system can be used without clinical infrastructure. The web-based deployment ensures cross-platform accessibility from desktop and mobile browsers. The expert-validated rule base and cardiologist-approved symptom set further lend clinical credibility to the diagnostic outputs.

Nevertheless, several limitations must be acknowledged. First, the diagnostic capability of the system is constrained by the self-reported nature of its inputs, rendering it susceptible to recall bias, misreporting, and insensitivity to asymptomatic or chronic disease states as evidenced by the two false negative cases. Second, the evaluation sample of ten cases, while sufficient for an initial proof-of-concept assessment, is too small to yield statistically robust conclusions regarding the system's generalizability across diverse demographic and clinical populations. Third, the current system does not incorporate temporal dynamics, longitudinal health tracking, or objective diagnostic indicators such as electrocardiographic findings or biomarker levels, all of which significantly contribute to clinical decision-making.

D. Implications and Future Work

Despite these limitations, the system fulfills its intended role as a preliminary early-warning tool rather than a replacement for clinical diagnosis. It promotes proactive health behavior by enabling individuals to assess their cardiac risk independently and decide whether to seek professional consultation. This functionality is particularly valuable in settings with limited specialist access, such as rural and peri-urban communities in Indonesia.

Future enhancements should prioritize three areas: (1) expanding the evaluation dataset to at least 100 validated clinical cases to enable statistically robust cross-validation and more reliable computation of specificity; (2) incorporating hybrid approaches that combine fuzzy inference with machine learning classifiers such as support vector machines or neural networks to capture non-linear feature interactions that may elude rule-based inference [5]; and (3) integrating wearable sensor data or electronic health record APIs to enable continuous, longitudinal risk monitoring. The exploration of Adaptive Neuro-Fuzzy Inference Systems (ANFIS) represents a particularly promising direction, as it enables automated optimization of membership function parameters from training data while retaining the linguistic interpretability of the fuzzy framework [16].

E. Black-Box Functional Testing Results

Functional correctness was evaluated through black-box testing covering 25 test cases across nine system modules: Login, Data Feedback, Data Community, Community Detail, Landing Page, Input Diagnosis, Diagnosis Result, Admin Sidebar, and User Navigation Bar. All 25 test cases received a "Passed" status, confirming that each feature performed in full conformance with its specified functional requirements. Critical test scenarios included validation of form completeness enforcement, accurate routing of user interactions, secure session management for administrative functions, correct rendering of diagnostic outputs, and responsive behavior on mobile viewports. The comprehensive

black-box test results are available in the supporting documentation.

IV. Conclusions

This paper has presented the design, implementation, and evaluation of a web-based fuzzy expert system for early heart disease diagnosis using the Mamdani inference method. The system was developed in collaboration with a practicing cardiologist at An-Nisa Hospital Tangerang, incorporating seventeen input variables including BMI, blood pressure, symptom profiles, lifestyle factors, and psychological status processed through Gaussian membership functions and an 18-rule inference base. Defuzzification via the centroid method produces a continuous risk percentage that is subsequently mapped to categorical risk levels and actionable medical recommendations.

Evaluation against ten hospital patient records yielded an accuracy of 80%, with a precision of 100%, demonstrating strong positive predictive value while acknowledging sensitivity constraints attributable to asymptomatic presentations. Functional black-box testing confirmed that all 25 system test cases passed without exception. Comparative analysis with prior studies positions this system favorably within the existing literature, while its unique contribution lies in the expert-co-validated, symptom-driven approach designed explicitly for independent public use.

Future research will focus on enlarging the clinical evaluation dataset, exploring hybrid fuzzy-neural approaches such as ANFIS to enhance sensitivity and adaptability, and integrating the system with continuous health monitoring technologies. It is emphasized that the system is intended to serve as a supplementary screening tool and not as a replacement for qualified clinical assessment.

References

[1] World Health Organization, "Cardiovascular diseases (CVDs)," World Health Organization.

[2] A. Timmis, P. Vardas, N. Townsend, A. Torbica, and H. Katus, "European Society of Cardiology: cardiovascular disease statistics 2021: Executive Summary," *Eur. Heart J. Qual. Care Clin. Outcomes*, vol. 8, no. 4, pp. 377–382, Jul. 2022.

[3] G. N. Levine, B. E. Cohen, and Y. Commodore-Mensah, "Psychological Health, Well-Being, and the Mind-Heart-Body Connection: A Scientific Statement From the American Heart Association," *Circulation*, vol. 143, no. 10, pp. e763–e783, Mar. 2021.

[4] A. Handayani, J. Hutahaean, and A. Nasution, "Penerapan Metode Hybrid Case Base Pada Sistem Pakar Diagnosa Penyakit Jantung," *Building of Informatics, Technology and Science (BITS)*, vol. 4, no. 2, pp. 537–544, Sep. 2022, doi: 10.47065/bits.v4i2.1981.

[5] K. A. Bhavsar, J. Singla, Y. D. Al-Otaibi, and O.-Y. Song, "Medical Diagnosis Using Machine Learning: A Statistical ReviewKaustubh Arun," *Computers, Materials & Continua*, vol. 67, no. 1, pp. 107–125, Jan. 2021.

[6] S. Widaningsih, "Analisis Perbandingan Metode Fuzzy Tsukamoto, Mamdani dan Sugeno dalam Pengambilan Keputusan Penentuan Jumlah Distribusi Raskin di Bulog Sub. Divisi Regional (Divre) Cianjur," *Jurnal Informatika dan Manajemen STMIK*, vol. 11, no. 1, May 2017.

[7] R. Mahdalena Simanjorang, A. Simangunsong, M. Arifin, M. Yamin, T. Informatika, and S. Pelita Nusantara, "Penerapan Sistem Pakar Dalam Diagnosis Dini Penyakit Jantung Dengan Metode Sistem Inferensi Fuzzy," *Jurnal Ilmu Komputer dan Sistem Informasi (JIKOMSI V*, vol. 7, no. 1, pp. 131–142, 2024, [Online]. Available: https://ejournal.sisfokomtek.org/index.php/jikom

[8] D. P. Nuraeni, "Perancangan Sistem Pakar Diagnosa Penyakit Jantung Berbasis Web Menggunakan Metode Fuzzy Mamdani (Studi Kasus : Dinas Kesehatan Rangkasbitung)," *Teknik dan Multimedia*, vol. 1, no. 4, 2023.

[9] U. Athiyah *et al.*, "Diagnosa Resiko Penyakit Jantung Menggunakan Logika Fuzzy Metode Tsukamoto," Feb. 2021.

[10] A. Akhtar, B. Bakhtawar, and S. Akhtar, "EXTREME PROGRAMMING VS SCRUM: A COMPARISON OF AGILE MODELS," *International Journal of Technology, Innovation and Management (IJTIM)*, vol. 2, p. 2022, 2022, doi: 10.54489/ijtim.v2i1.77.

[11] S. Shafiq, W. Akbar, A. Hussain, T. Hussain, and A. Soomro, "FuzzDL-HeartPredict: Heart Attack Risk Prediction using Fuzzy Logic and Deep Learning," *ICCK Transactions on Advanced Computing and Systems*, vol. 1, pp. 63–77, May 2024.

[12] I. Ummah, N. Yannuansa, and I. Mufarrihah, "Pengaruh Penentuan Domain, Fungsi Keanggotaan Dan Rule Dalam Membangun Sistem Fuzzy," Oct. 2021. [Online]. Available: https://kaggel.com/.

[13] A. Christopher and S. Anjelina, "ANALISIS PERFORMA FRAMEWORK FRONTEND JAVASCRIPT BAGI PROGRAMMER," Dec. 2022.

[14] C. Wijayanto and Y. A. Susetyo, "IMPLEMENTASI FLASK FRAMEWORK PADA PEMBANGUNAN APLIKASI SISTEM INFORMASI HELPDESK (SIH)," Sep. 2022.

[15] R. Astriratma, M. Mega Santoni, H. Nurramdhani Irmanda, F. Ilmu Komputer, and U. Pembangunan Nasional Veteran Jakarta, "SPK Berbasis Web Menggunakan Logika Fuzzy Untuk Diagnosa Penyakit Jantung," *JSI : Jurnal Sistem Informasi (E-Journal)*, vol. 12, no. 1, 2020, [Online]. Available: http://ejournal.unsri.ac.id/index.php/jsi/index

[16] L. Yuningsih, R. R. Huizen, G. A. Pradipta, and P. D. W. Ayu, "Adaptive Neuro-Fuzzy Inference System For Medical Image Classification -A Review," in *2022 4th International Conference on Cybernetics and Intelligent System (ICORIS)*, Prapat, Indonesia: IEEE, 2022

ISSN 2085-4552

Implementation Of Agile Methods In Designing A Web-Based Information System In The Geomin Laboratory Of PT Antam Tbk

Dennis Bhayu Bramastha[1], Waeisul Bismi[2]
[1,2] Universitas Bina Sarana Informatika, Jakarta, Indonesia
[1] 15190062@bsi.ac.id, [2] Waeisul.wbn@bsi.ac.id

Accepted 03 June 2026
Approved 28 June 2026

***Abstract*— The management of human resource administration at the Geomin Laboratory Unit of PT Antam Tbk, which includes recording attendance, overtime requests, and leave requests, faces various obstacles due to the continued implementation of manual procedures. This process, which relies on physical forms, is not only time-consuming in recapitulation, but is also prone to recording errors and document loss, as well as the lack of accurate quota monitoring. This study aims to design and build a web-based information system that can overcome these problems by integrating all administrative processes into a single digital platform. The development method used in this study is the Agile method, which allows for a flexible and iterative development process. This system is built using the Laravel framework and utilizes geolocation technology for attendance validation. The result of this study is a functional information system with key features such as online attendance, a multi-level approval flow, and real-time leave quota management. Based on the results of the User Acceptance Test involving 15 respondents from various roles, this system achieved an acceptance rate of 93.1%, which falls into the "Highly Acceptable" category. This shows that the developed system has successfully answered user needs and can be an effective solution to increase the efficiency and accuracy of administrative processes in the laboratory environment.**

***Index Terms*— agile; attendance; human resources; information systems; laravel.**

I. INTRODUCTION

Efficient and accurate human resource (HR) administration is a crucial element in maintaining smooth operations in every company's work unit [1]. PT Antam Tbk's Geomin Unit Laboratory, for administrative activities such as attendance recording, overtime requests, and leave management, requires a thorough and precise system, as the accuracy of these processes impacts employee discipline and productivity. Based on the author's experience as an administrative staff member at the laboratory, it was discovered that attendance, overtime, and leave management are still performed manually using physical forms. This manual system creates various obstacles, such as delays in data recapitulation, the risk of recording errors, lost documents, and unclear monitoring of leave quotas.

According to Prasetyo and Komarudin [2], the manual system makes the absence and leave application process inefficient due to obstacles such as difficulty obtaining superior approval, lost or damaged forms, and the lack of document backups. Meanwhile, Mahardi and Wicaksono [3] emphasized that the lack of an integrated system can lead to data duplication and increase the workload of the HR department. Therefore, developing a web-based information system is a strategic step to address these issues. This system is expected to provide centralized data access, expedite the application and approval process, and automate leave quota calculations with automatic validation features and digital recording, improving administrative efficiency and accuracy.

This aligns with the opinion of Lubis and Adlina [4] that the use of a Human Resource Information System (HRIS) can make managing personnel, payroll, and attendance data more effective and efficient. In its development, an Agile approach was used, which is flexible and iterative, allowing for dynamic system adjustments to user needs and emphasizing active collaboration between developers and users. Farras et al. [5] also demonstrated that implementing Agile methods can increase efficiency and productivity in HR management.

Effective human resource administration requires the support of a Human Resource Information System (HRIS). According to Sipahutar and Widyaningsih [1] , an HRIS is a combination of human resource management and information technology, while Lubis and Adlina [4] state that an HRIS serves to manage personnel information in an integrated manner, thereby increasing effectiveness and efficiency compared to manual systems. Thus, HRIS can help organizations manage personnel data more quickly, accurately, and in a structured manner.

While commercial HRIS technologies already exist, developing a customized Web-Based Information System specifically for the Geomin Unit Laboratory of

PT Antam Tbk is highly necessary. Commercial systems often involve recurring licensing costs and rigid workflows. In contrast, a customized system is free from third-party licensing, easy to manage, simple to modify, and seamlessly upgradeable to accommodate the laboratory's specific multi-level approval hierarchy.

The Agile method was used in developing this system because it can adapt to changes in user needs. Prastowo et al. [6] note that Agile focuses on user satisfaction through continuous software development. Meanwhile, Handayani et al. [7] explain that Agile emphasizes rapid response to changes through the stages of planning, implementation, testing, deployment, and maintenance. Therefore, this method was chosen because it supports flexible system development tailored to user needs.

The attendance system utilizes the Haversine algorithm to validate employees' attendance locations. According to Wijaya and Tony [8] this algorithm is used to calculate the distance between two geographic coordinate points so that the system can determine whether a user is within a specified radius. The use of the Haversine algorithm has been shown to improve the accuracy of attendance validation and support the implementation of geolocation-based attendance tracking [9].

Previous research has shown that information technology can improve the effectiveness of personnel administration. Wijaya and Tony [8] as well as Hazani and Mursid [10] demonstrated that the application of the Haversine algorithm in a web-based attendance system can validate attendance locations and improve attendance accuracy. Winanto et al. [11] developed a web-based leave request system that was deemed feasible and effective in streamlining administrative processes. Meanwhile, Junaedi et al. [12] demonstrated that the Laravel framework is suitable for HRIS system development due to its modular nature and ease of development. Additionally, Farras et al. [5] concluded that the Agile Scrum method can produce flexible systems that align with user needs.

Based on this background, this research focuses on the application of Agile methods in designing a web-based information system at the Geomin Unit Laboratory of PT Antam Tbk as a solution to existing administrative problems, with advantages in processing speed, data accuracy, ease of access, and reducing the risk of document loss due to manual systems.

The purpose of this research is to design and develop a web-based information system that can centrally and efficiently manage attendance, overtime, and leave data. Furthermore, this research aims to apply an Agile approach to the information system development process to ensure the system's flexibility and ability to adapt to user needs throughout the development process. It also aims to introduce digital features such as online attendance recording with validation, a tiered approval process, automatic recapitulation, and integrated leave quota monitoring, all of which can improve the effectiveness and efficiency of HR administration at the Geomin Unit Laboratory of PT Antam Tbk.

II. METHOD

A. Software Requirements Analysis

Based on the author's observations, personnel administration processes at the Geomin Unit Laboratory, such as attendance, overtime, and leave, are still carried out manually using physical forms. This system involves four main actors: Personnel, who record attendance and request leave or overtime; Assistant Managers, who provide initial approval; Managers, who provide final approval; and Admins, who manage master data, leave quotas, and reports. All four play an integrated role to create effective and transparent administrative management.

B. System Architecture

The system architecture in this study uses the Model-View-Controller (MVC) pattern, which is the basis of the Laravel framework. This pattern separates data, user interface, and control logic, making the system more structured, modular, and easy to develop. The Model manages data and business rules through the Eloquent class in app/Models, the View displays data using the Blade templating engine, and the Controller acts as a bridge between the two by processing user input and displaying results to the View through classes in app/Http/Controllers.

C. Database Design

After designing the system architecture, the next stage is database design, which defines the data storage structure as the main foundation of system functionality. The goal of this stage is to build a structured, efficient data model that maintains long-term data consistency and integrity. The design process involves three main stages: conceptual modeling using Entity Relationship Diagrams (ERDs), logical structure design using Logical Record Structures (LRSs), and defining technical specifications for each table used in the system.

D. Interface Design

Interface design aims to organize all visual aspects of an information system, including layout, color scheme, and typography. The primary goal is to create a functional, intuitive, and user-friendly interface to provide an optimal user experience.

In terms of color scheme and typography, the system uses a primary color palette of blue (#435ebe) for interactive elements such as buttons and links. This color is combined with green (#198754) for success notifications, red (#dc3545) for warning or error messages, and light blue (#f2f7ff) for the background to

create a clean and spacious feel. The primary text color is a bluish gray (#607080) for easy readability.

TABLE I. SYSTEM COLOR SCHEME

Name	Hex Code	Preview
Primary Color (Blue)	#435ebe	
Background Color	#f2f7ff	
Primary Text Color	#607080	
Success Accent (Green)	#198754	
Warning/Error Accent (Red)	#dc3545	

The chosen font is Nunito, a modern and user-friendly sans-serif font, ensuring text remains clear and comfortable to read on various screen sizes.

E. Agile Development

This research employs the Agile methodology, which prioritizes customer satisfaction through continuous and early software delivery [6]. The Agile approach allows for rapid adaptation to changing requirements through short development cycles and intensive collaboration between developers and users [7]. The process typically involves structured phases: planning, implementation, testing, documentation, deployment, and maintenance [7].

To apply Agile principles concretely, the initial requirements gathered from the Geomin Laboratory of PT Antam Tbk were first translated into user stories. Examples of these user stories include: "As a personnel member, I want to submit leave requests online so that my quota is automatically updated," and "As a manager, I need to bulk-approve overtime requests to streamline the administrative process." These user stories were then evaluated, prioritized, and compiled into a Product Backlog.

The system development spanned approximately one month (mid-April to mid-May 2025). Before each cycle, a sprint planning phase was conducted to pull specific user stories from the Product Backlog into the Sprint Backlog. The project was divided into four distinct iteration cycles (sprints), with each sprint focusing on delivering specific, functional modules:

- Sprint 1 (Master Data and Core Setup): This initial phase focused on system initialization, including Laravel setup, database migration, and the development of the authentication module (login/register). It also covered the Master Data management module, enabling Admins to manage personnel profiles, roles (e.g. Assistant Manager Analis, Assistant Manager Preparator), holiday schedules, and vendor information.
- Sprint 2 (Business Travel and Leave Management): The second iteration shifted focus to employee administration features. It began with the Business Travel (Perjalanan Dinas) module, followed by the complete Leave (Cuti) management system. This included leave quota generation, submission forms, PDF template generation, and the implementation of role-based access control and approval workflows.
- Sprint 3 (Overtime and Notifications): This sprint was dedicated to the Overtime (Lembur) module. Development included overtime request views, recap functionality, PDF template generation, and an automated email notification system for overdue reminders, bulk overtime approvals, and individual status updates.
- Sprint 4 (Attendance and Finalization): The final sprint implemented the most complex feature: the geolocation-based daily attendance system (check-in/check-out). This phase also included the attendance correction workflow, monthly timesheet management (with filtering and PDF export), and final code refactoring to improve system maintainability before deployment.

At the end of each sprint, the developed modules were evaluated against the initial requirements. This concrete application of short, focused sprints allowed for continuous integration and immediate adjustments, ensuring the final Web-Based Information System was robust, functional, and ready for the User Acceptance Test (UAT).

III. RESULT AND DISCUSSIONS

A. Implementation of Main Functions

The daily attendance feature is the primary function most frequently used by personnel. The flowchart in Fig. 1 shows the process from logging in, selecting the attendance menu, to recording attendance by the system, ensuring the attendance process is automated, accurate, and efficient.

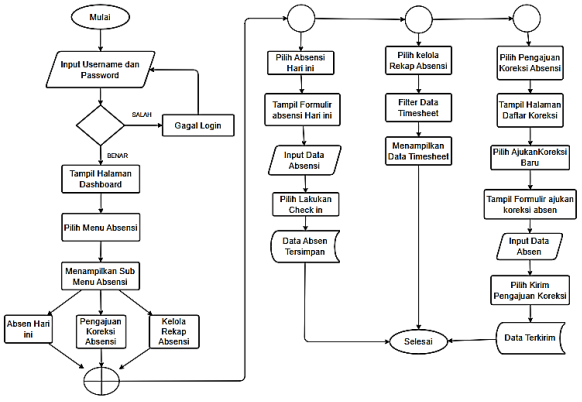


Fig. 1. Daily Attendance Process Flowchart (Personnel)

The leave approval feature illustrates a multi-level approval process involving management. Fig. 2 shows the workflow of the Assistant Manager as the first

approver, who is responsible for conducting initial validation of leave requests from personnel before they are forwarded to the next approval stage.

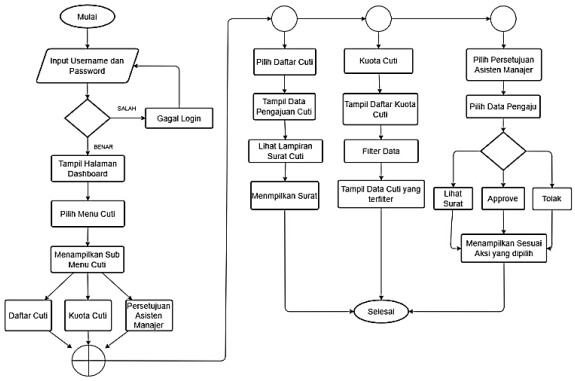


Fig. 2. Leave Approval Process Flowchart (Assistant Manager)

Once approved by the Assistant Manager, the process proceeds to the final stage by the Manager. Fig. 3 shows how the Manager reviews the validated application and makes a final decision on the leave request.

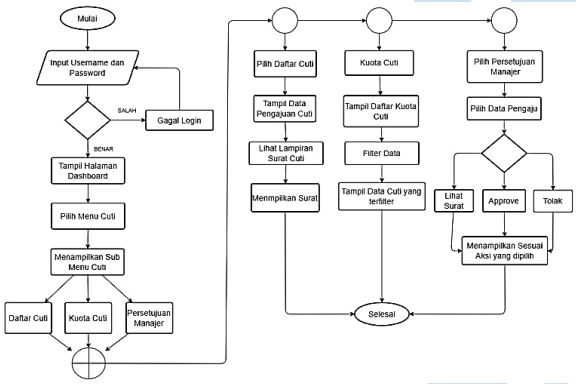


Fig. 3. Flowchart of Leave Approval Process (Manager)

As the primary function in master data management, Fig. 4 displays the Admin workflow in the Manage Personnel menu. This process includes menu selection and the options to add, change, and delete user data, which also illustrates a similar flow in the Manage Vendor and Manage Holiday features.

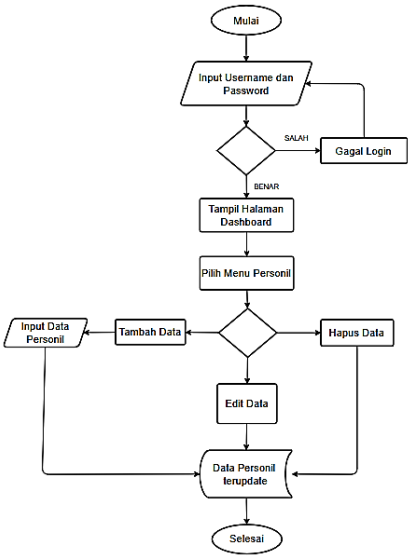


Fig. 4. Personnel Data Management Process Flowchart (Admin)

B. Test Result

The primary focus of this testing is on the accuracy of the input and output processes in each feature. To ensure structured testing, test scenarios are divided based on user roles (actors). The first section, General Functionality Testing (Authentication), is presented in Table II, which contains various test scenarios for the login and authentication features applicable to all user types.

TABLE II. BLACK BOX TEST (AUTHENTICATION)

Testing Scenario	Expected results	Results	Conclusion
The user entered a valid email address and password.	The system successfully authenticates and redirects the user to the appropriate dashboard page.	Appropriate	**Valid**
The user entered an incorrect password.	The system rejects the authentication and displays the error message "Email or password is incorrect."	Appropriate	**Valid**
The user logged out.	The system successfully terminates the session and redirects the user to the login page.	Appropriate	**Valid**

Table III details comprehensive test scenarios for features accessible to users with the Admin role.

TABLE III. BLACK BOX TEST (ADMIN)

Modul	Testing Scenario	Expected results	Results	Conclusion
Manage Personnel	Add, modify, and delete personnel data (CRUD).	All CRUD operations on personnel data performed as expected.	Appropriate	**Valid**
Manage Vendors	Add, modify, and delete vendor data (CRUD).	All CRUD operations on vendor data performed as expected.	Appropriate	**Valid**
Manage Holidays	Add, modify, and delete vacation data (CRUD).	All CRUD operations on holiday data performed as expected.	Appropriate	**Valid**
Overtime Requests	Request overtime on behalf of another employee.	The application was successfully saved and recorded under the name of the selected employee.	Appropriate	**Valid**
Leave Requests	Request leave for yourself.	The application was successfully saved and entered the approval flow.	Appropriate	**Valid**
Leave Quotas	Edit an employee's remaining leave quota.	The employee leave quota data was successfully updated in the database.	Appropriate	**Valid**
Attendance Summary	View and filter attendance summaries for all employees.	The page displays the attendance summary data correctly.	Appropriate	**Valid**
Attendance Summary	Download approved monthly timesheets.	The system successfully downloaded a valid PDF file.	Appropriate	**Valid**
Overtime Summary	Export overtime summary data to Excel.	The system successfully downloaded an Excel file containing the appropriate data.	Appropriate	**Valid**
Export Overtime Forms	Download PDF forms for approved overtime requests.	The system successfully downloaded the PDF file correctly.	Appropriate	**Valid**
Export Leave Forms	Download PDF forms for approved leave requests.	The system successfully downloaded the PDF file correctly.	Appropriate	**Valid**

Table IV details the test scenarios for the features accessible to users with the Personnel role.

TABLE IV. BLACK BOX TEST (PERSONNEL)

Modul	Testing Scenario	Expected results	Results	Conclusion
Attendance	Check in and check out on the same day.	Attendance data is recorded correctly, including arrival and departure times.	Appropriate	**Valid**
Attendance Correction	Submit corrections for incomplete attendance data.	The correction request is successfully submitted with the status "Awaiting Approval."	Appropriate	**Valid**
Overtime	Submit a new overtime request (Create).	The request is successfully saved and appears in the request history.	Appropriate	**Valid**
Overtime	Edit an overtime request that is still "Pending Approval" (Update).	The data changes to the request are successfully saved.	Appropriate	**Valid**
Overtime	Cancel an overtime request (Cancel).	The request status changes to "Cancelled."	Appropriate	**Valid**
Leave	Submit a leave request with valid dates and sufficient quota (Create).	The request is successfully saved.	Appropriate	**Valid**
Leave	Edit a leave request that is "Rejected" (Update).	The edit form opens, and a new request can be submitted.	Appropriate	**Valid**
Leave	Download an approved leave form.	The system successfully downloads the appropriate PDF file.	Appropriate	**Valid**
Business Travel	Submit a 30-day business trip request.	After the business trip is completed, the special leave quota is automatically increased by 3 days.	Appropriate	**Valid**
Profile	Change a password by confirming an invalid password (Invalid).	The system displays the error message "Password confirmation does not match."	Appropriate	**Valid**

Table V details the test scenarios for the features accessible by the Management role.

TABLE V. BLACK BOX TEST (MANAGEMENT)

Modul	Testing Scenario	Expected results	Results	Conclusion
Leave Approval	Reject leave requests by providing a reason for the rejection.	The application status changes to "Rejected," and a notification with a reason is sent to the personnel.	Appropriate	**Valid**
Overtime Approval	Approve multiple overtime requests at once (bulk approve).	The status of all selected applications changes to "Approved" (or the next level).	Appropriate	**Valid**
Timesheet Summary	Make final approvals on monthly timesheet recaps.	The timesheet status changes to "Approved," and can be printed as a PDF.	Appropriate	**Valid**
Absence Correction	Approve requests for absence corrections.	The associated attendance data is successfully updated based on the application.	Appropriate	**Valid**
Authorization	Attempt to access the Manage Personnel menu (direct URL).	The system denies access, displays an error notification, and redirects you to the dashboard.	Appropriate	**Valid**

After the system has passed the functional test, the next stage is the User Acceptance Test (UAT), which aims to assess the level of user acceptance and satisfaction with the system. Unlike black box testing, which focuses on functionality, UAT assesses usability, ease of use, and the system's suitability to operational needs in a real-world work environment. The UAT was conducted through a survey using a questionnaire containing closed-ended questions regarding the system's usability, appearance, and usefulness. Respondents provided ratings using a Likert scale of 1–5, with the following interpretations: 5 = Strongly Agree, 4 = Agree, 3 = Somewhat Agree, 2 = Somewhat Disagree, and 1 = Disagree.

A total of 15 respondents from the Geomin Unit Laboratory of PT Antam Tbk were involved, consisting of 2 Admins, 3 Management, and 10 Personnel, representing all user roles in the system.

To ensure the accuracy of the evaluation, the study employed a contextual validation approach rather than traditional variance-based reliability tests. Due to the role-specific nature of the HRIS modules (Role-Based Access Control), the questionnaire was not uniform; rather, it was divided into distinct sections tailored specifically for Admin, Management, and Personnel tasks. Because respondents did not answer identical sets of questions, tests such as Cronbach's Alpha were statistically inapplicable. Instead, validity was ensured by aligning each questionnaire item directly with the specific operational flow of the respective user role, ensuring the instrument accurately measured user acceptance based on their distinct administrative functions.

After all the questionnaire data were collected, the authors conducted a quantitative analysis to assess the level of user acceptance of the system. This analysis was performed by calculating the ideal score percentage, which was obtained using the following formula:

$$Percentage = \frac{Total\ score\ obtained}{Maximum\ total\ score}\ x\ 100\% \quad (1)$$

The percentage results are then interpreted based on the following ranges:

TABLE VI. UAT INTERPRETATION SCALE

Percentage Range	Interpretation
81% - 100%	Highly Acceptable
61% - 80%	Good / Acceptable
41% - 60%	Fair
21% - 40%	Poor
0% - 20%	Unacceptable

Based on the quantitative analysis of all UAT questionnaires, a user acceptance rate of 93.1% was obtained, which falls into the “Highly Acceptable” category according to the interpretation range of 81%–100%. This result indicates that the developed information system has been excellent received by users and is able to meet the needs and resolve operational problems at the Geomin Unit Laboratory of PT Antam Tbk. Details of the results are presented in the following table.

TABLE VII. FINAL RECAPITULATION OF UAT RESULTS

Group	Score Obtained	Maximum Score	Percentage	Interpretation
Admin	49	50	98.0%	Highly Acceptable
Management	71	75	94.7%	Highly Acceptable
Personnel	229	250	91.6%	Highly Acceptable
Total	**349**	**375**	**93.1%**	**Highly Acceptable**

C. Discussion

The developed system successfully digitized the attendance, leave, and overtime administration processes that were previously performed manually. Under the previous system, employees were required to complete paper-based forms and wait for physical approval from supervisors, which often caused delays, document loss, and difficulties in monitoring

administrative records. By implementing a web-based HRIS, all data can be stored and managed centrally, improving accessibility, efficiency, and accuracy in personnel administration. These findings are consistent with previous studies which state that HRIS implementation can improve the effectiveness of human resource management and reduce dependence on manual processes [3].

The attendance feature integrated with geolocation technology and the Haversine algorithm provides an additional advantage compared to conventional attendance systems. The system automatically validates whether employees are within the permitted attendance radius before attendance data is recorded. This mechanism minimizes attendance fraud and improves the accuracy of attendance records. Similar results were reported in previous studies evaluating location-based attendance monitoring [13], [14].

From an operational perspective, the implementation of a multi-level approval workflow enables leave and overtime requests to be processed more efficiently. Employees can submit requests online, while Assistant Managers and Managers can review and approve requests in real time. Compared to the previous manual process—which typically took 1.5 to 3 working days for a complete approval cycle—this digital approach reduces the processing time to mere minutes, eliminating administrative delays and improving transparency. Similar benefits of web-based administrative systems have been identified in recent literature; digitalization has been proven to significantly simplify personnel administration [2], and geolocation-based systems enhance employee discipline [14].

The use of the Laravel framework also contributed to the successful implementation of the system. Laravel provides a structured MVC architecture that facilitates system maintenance, scalability, and future development. Previous studies [12] similarly concluded that Laravel is suitable for developing complex web-based information systems because of its modular structure and development efficiency.

During implementation, several challenges were encountered, particularly in adapting users who were accustomed to manual administrative procedures. Some users required additional guidance during the initial adoption phase. However, the Agile development approach enabled continuous feedback collection and iterative improvements throughout the development cycle. This finding supports previous research [15], which demonstrated that Agile methodologies facilitate adaptation to changing requirements and improve system suitability to user needs.

The User Acceptance Test (UAT) produced an acceptance score of 93.1%, indicating that the system was highly accepted by users. This result reflects not only the functionality of the developed features but also the system's ability to address operational challenges within the Geomin Laboratory. Similar high levels of user acceptance and service improvement have been reported in web-based information system implementations [16]. Furthermore, the use of a Likert-scale-based evaluation approach is consistent with recommendations in [17] for measuring user satisfaction and service quality in information systems.

Although the system achieved a high user acceptance rate, this study has several limitations. The UAT was conducted with only 15 respondents from a single organizational unit, which may limit the generalizability of the results. In addition, the use of questionnaire-based assessments may introduce response bias because user perceptions are inherently subjective. To address these limitations, future studies are encouraged to involve larger and more diverse samples to evaluate the long-term impact on organizational performance. Furthermore, subsequent system development should focus on integrating automated payroll modules, incorporating disaster recovery and data backup systems, and implementing biometric authentication to further strengthen attendance validation.

IV. CONCLUSIONS

A web-based personnel information system at the Geomin Unit Laboratory of PT Antam Tbk has been successfully developed to replace manual processes with a more efficient, accurate, and secure digital one. The application of Agile methods allows the system to be developed adaptively to user needs. Based on the results of the User Acceptance Test (UAT), with an acceptance rate of 93.1%, the system is considered highly acceptable and beneficial to all users.

It is recommended that the system be developed into a mobile app (Android/IOS) to improve accessibility, and that a more structured Agile framework such as Scrum be used. Furthermore, further development can be directed at integrating modules compliant with the ISO/IEC 17025 standard, such as competency management and personnel training.

ACK NOWLEDGMENT

The author would like to express his gratitude to his supervisor for his guidance and direction, the Geomin Unit Laboratory of PT Antam Tbk for their support and cooperation, and to his family and colleagues for their constant prayers and encouragement. He hopes this work will be beneficial and serve as a reference for the development of similar systems in the future

REFERENCES

[1] T. W. Sipahutar, A. B. S., & Widyaningsih, "Analisis dan perancangan human resource information system (HRIS) pada Universitas Tanri Abeng," *J. Tek. Inform. Unis*, vol. 11, no. 1, pp. 23–36, 2023.

[2] H. Prasetyo and R. Komarudin, "Perancangan sistem informasi absensi dan cuti karyawan pada PT. Pinus Merah

Abadi," *Jutis (Jurnal Tek. Inform.*, vol. 9, no. 2, pp. 136–145, 2021, [Online]. Available: http://ejournal.unis.ac.id/index.php/jutis/article/download/1835/1602

[3] K. A. M. Mahardi and S. R. Wicaksono, "Pengembangan sistem informasi sumber daya manusia berbasis web untuk optimalisasi pengelolaan SDM di PT. Lotus Indah Textile," *CENTIVE Conf. Electr. Eng. Informatics, Ind. Technol. Creat. Media*, vol. 4, pp. 1133–1140, 2024.

[4] R. Aulia Lubis and H. Adlina, "Penerapan human resources information system (HRIS) dalam meningkatkan efektivitas dan efisiensi sumber daya manusia (studi pada PT Bank Sumut)," *J. Ilm. Wahana Pendidik.*, vol. 11, no. April, pp. 135–143, 2025.

[5] M. R. Farras, R. R. Putri, S. Bin, and M. Makarim, "Rancang bangun dan analisis kinerja HRD menggunakan metode Agile Scrum berbasis website," *INTEGER J. Inf. Technol.*, vol. 10, pp. 1–9, 2025.

[6] W. D. Prastowo, D. Danianti, and A. Pramuntadi, "Analisis risiko pada pengembangan perangkat lunak menggunakan metode Agile dan RAD (Rapid Application Development)," *Citiz. J. Ilm. Multidisiplin Indones.*, vol. 3, no. 3, pp. 169–174, 2023, doi: 10.53866/jimi.v3i3.388.

[7] H. Handayani, K. U. Faizah, A. M. Ayulya, M. F. Rozan, D. Wulan, and M. L. Hamzah, "Perancangan sistem informasi inventory barang berbasis web menggunakan metode Agile software development," *J. Test. dan Implementasi Sist. Inf.*, vol. 1, no. 1, pp. 29–40, 2023, [Online]. Available: https://journal.al-matani.com/index.php/jtisi/article/view/324

[8] L. F. Wijaya and Tony, "Implementation of Haversine algorithm in web-based online attendance application," *IOTA*, vol. 4, no. 4, 2024, doi: 10.31763/iota.v4i4.836.

[9] L. N. Aini, E. T. Handayani, and R. Nuraini, "Implementasi algoritma Haversine untuk perhitungan jarak antara lokasi perusahaan dengan karyawan pada PT Mega Giga Solusindo," *JORAPI J. Res. Publ. Innov.*, vol. 2, no. 4, pp. 2859–2872, 2024.

[10] L. O. Hazani and Mursid, "Implementasi formula Haversine pada sistem absensi pegawai berbasis web (studi kasus: SMA Islam GUPPI Kota Sorong)," *J. Teknol. Inf.*, vol. 12, no. 2, pp. 9–17, 2024.

[11] A. Winanto, B. Baharsyah, and F. Felawati, "Perancangan sistem informasi pengajuan cuti berbasis web pada Balai Penerapan Standar Instrumen Pertanian Jambi," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 9, no. 2, pp. 1812–1818, 2025.

[12] M. M. H. Junaedi, S. Susanti, and A. Mubarok, "Penerapan framework Laravel pada aplikasi HRIS (human resource information system)," *J. Responsif Ris. Sains dan Inform.*, vol. 2, no. 2, pp. 176–183, 2020, doi: 10.51977/jti.v2i2.310.

[13] Nurkholis and M. A. Sobarnas, "Penerapan geolokasi pada absensi fasilitator program padat karya pemerintah yang tersebar di seluruh wilayah Indonesia," *INFOTECH J. Inform. Teknol.*, vol. 1, no. 2, pp. 116–126, 2020, doi: 10.37373/infotech.v1i2.69.

[14] T. Saputra, A. Utari, U. Teisnajaya, and G. Twenty Agustine, "Sistem absensi karyawan menggunakan self potrait dan geolocation pada PT Sucofindo Palembang," *Klik - J. Ilmu Komput.*, vol. 4, no. 2, pp. 68–77, 2023, doi: 10.56869/klik.v4i2.553.

[15] A. Samsudin and H. Hamdalah Islami, "Sistem pengaduan masyarakat menggunakan metode Agile extreme programming," *J. Infotex*, vol. 2, no. 1, pp. 214–226, 2023.

[16] Rina Noviana, "Pembuatan aplikasi penjualan berbasis web Monja Store menggunakan PHP dan MySQL," *J. Tek. dan Sci.*, vol. 1, no. 2, pp. 112–124, 2022, doi: 10.56127/jts.v1i2.128.

[17] A. L. Soba, D. Syahputra, and M. Adriansyah, "Pembuatan website untuk meningkatan pelayan bidang informasi dan komunikasi publik di Diskominfotik Provinsi Bengkulu," *Generic*, vol. 15, no. 2, pp. 32–36, 2023, doi: 10.18495/generic.v15i2.152

ISSN 2085-4552

Machine Learning for Chili Pepper Price Forecasting Using Exogenous Public-Attention Signals and Bayesian Hyperparameter Optimization

Wresti Andriani[1], Gunawan Gunawan[2], Naella Nabila Putri Wahyuning Naja[3]
[1] Departemen of Informatika, Universitas Bima Sakapenta, Tegal, Indonesia
[2] Depertemen of Sistem Informasi, Universitas Pancasakti, City, Indonesia
[3]Departemen of Ekonomi Manajemen, Universitas Negeri Semarang, Semarang, Indonesia
[1] wresty.andriani@gmail.com, [2] Waeisul.wbn@bsi.ac.id, [3] naellanabilap@students.unnes.ac.id

Accepted 21 June 2026
Approved 28 June 2026

***Abstract*— Chili pepper prices in Indonesia are highly volatile due to seasonal production, distribution constraints, and short-term market reactions to public issues. This study proposes a machine-learning approach for short-term chili pepper price forecasting by integrating public-attention signals from Google Trends and news volume into a composite Shock Index. A LightGBM model optimized using Bayesian hyperparameter optimization was evaluated through rolling time-series cross-validation across multiple forecasting horizons. The results show that the attention-augmented scenario, which combines price features, Google Trends, news volume, and the Shock Index, consistently outperforms the price-only baseline. After optimization, the best scenario achieved an average sMAPE of 12.47% and maintained MASE below 1, indicating better performance than the naïve benchmark. These findings suggest that public-attention signals provide useful early-warning information for forecasting chili price movements, particularly when market uncertainty increases. The proposed approach can support nowcasting and price-monitoring systems for policymakers and market participants, although its performance remains dependent on the stability and representativeness of online attention data.**

***Index Terms*— Bayesian hyperparameter optimization ; chili price forecasting; LightGBM; public-attention signals; Shock Index.**

I. INTRODUCTION

Chili peppers are a strategic horticultural commodity in Indonesia because they affect household expenditure, food inflation, and food price stability[1]. Seasonal production dynamics, together with supply chains that are sensitive to weather, logistics costs, and distribution disruptions, make chili prices highly volatile [2]. In addition to these structural factors, public perception and media attention toward specific issues, such as food-safety concerns, imported chili, or distribution shortages, may amplify short-term market expectations [3]. When such issues receive high online search interest and news coverage, they can influence consumer concern, trader sentiment, and anticipatory market behavior. Therefore, chili price forecasting should not rely only on historical price patterns but also consider high-frequency external signals that may function as early-warning indicators.

Prior studies provide an important foundation for chili price forecasting in Indonesia. Pangesti et al.[4] compared ARIMA and a Gated Recurrent Unit (GRU) model for forecasting chili prices in East Java and found that GRU produced better forecasting accuracy. Prihandi et al.[5] applied ARIMA with weather-related variables, including rainfall and sunlight duration, to examine the relationship between production conditions and red chili prices in North Sumatra. Meanwhile, Gunadi and Perdana [6]. used a univariate Long Short-Term Memory (LSTM) model to forecast chili prices in Metro City. These studies indicate that nonlinear models and external covariates can improve forecasting performance, but their implementations remain limited in terms of feature scope, evaluation design, and optimization strategy.

Based on these prior studies, three research gaps can be identified. First, many chili price forecasting models still rely mainly on historical price data and do not systematically include high-frequency exogenous signals that represent public attention, such as online search interest and news intensity. Second, although ARIMA-based models can incorporate external variables, they are limited in capturing nonlinear relationships and rapidly changing attention dynamics. Third, many studies still rely on a single holdout evaluation, while rolling or expanding multi-fold backtesting, Bayesian hyperparameter optimization, ablation analysis, and interpretability testing are not consistently applied. These gaps limit the

reproducibility and operational reliability of existing forecasting models.

This study aims to: (1) improve short-horizon chili pepper price forecasting accuracy by integrating public-attention signals as exogenous features; (2) evaluate the contribution of Google Trends, news volume, and their composite Shock Index through an ablation study; (3) apply rolling time-series cross-validation to obtain more reliable out-of-fold performance estimates; and (4) optimize LightGBM configurations using Bayesian hyperparameter optimization. The forecasting performance is evaluated using sMAPE, RMSE, and MASE across multiple horizons, namely H = 1, 2, and 4 weeks.

To achieve these objectives, this study collects chili price data by region and combines them with public-attention signals from Google Trends and news volume. The feature engineering process includes price lags, moving averages, rolling volatility, calendar features, and a composite Shock Index representing the intensity of public issues. A LightGBM model is then trained and optimized using Bayesian hyperparameter optimization under a rolling cross-validation scheme. Feature contributions are examined through ablation scenarios and interpretability analysis.

The contribution of this study is twofold. Academically, it provides empirical evidence on the role of public-attention signals in improving machine-learning-based forecasting of horticultural commodity prices. Practically, it offers a reproducible forecasting framework that can support nowcasting and early-warning systems for policymakers, market participants, and other stakeholders concerned with food price stability.

II. Literature Review

Price volatility in chili pepper markets has received considerable attention in Indonesia because chili is closely related to household consumption, farmer income, and food inflation. At the market-structure level, Sukiyono and Asriani [7] found that chili prices in Bengkulu showed high volatility at the producer and wholesaler levels, with strong price transmission across vertical markets. This finding indicates that chili price movements are not determined only by production seasonality, but also by market-chain interactions, distribution mechanisms, and expectations formed by market actors.

Classical time-series models remain important in agricultural price forecasting. Windhy and Jamil [8] showed that ARIMA can be used to forecast Indonesian red chili prices by relying on historical price patterns and standard diagnostic procedures, including stationarity testing and ACF/PACF analysis. However, ARIMA-based approaches generally assume relatively linear relationships and may have limitations when dealing with nonlinear changes, sudden shocks, and regime shifts that frequently occur in horticultural commodity markets. Therefore, more flexible forecasting methods are needed to capture complex and changing price behavior.

Recent studies have increasingly adopted machine-learning and deep-learning methods for chili price prediction. Pangesti et al. [4] compared ARIMA and GRU models for chili price forecasting in East Java and found that GRU produced better forecasting performance. Gunadi and Perdana [6] applied LSTM to forecast chili prices in Metro City and reported promising predictive accuracy. These studies show that nonlinear models can improve price forecasting compared with purely classical models. Nevertheless, many deep-learning applications still rely mainly on historical price data, making them less sensitive to external shocks that may appear outside the price series itself.

For tabular time-series forecasting, gradient boosting models such as LightGBM have become attractive because they can handle nonlinear relationships, interactions among variables, and heterogeneous feature structures efficiently. Oikonomou and Damigos [9] demonstrated that AutoReg-LightGBM and LightGBM–ARIMA ensembles can perform competitively for short-term commodity price forecasting under rolling-origin evaluation. This suggests that LightGBM is suitable for forecasting problems involving lagged price features and external variables. However, its use in Indonesian chili price forecasting remains relatively limited, especially when combined with public-attention indicators and systematic hyperparameter optimization.

In addition to historical prices and market fundamentals, public-attention signals can provide useful information for nowcasting and short-term forecasting. Macias et al. [10]. showed that high-frequency online price information can improve food inflation nowcasting compared with traditional data sources. In commodity markets, Chi et al. [3] found that media emotion intensity is associated with commodity return dynamics, indicating that news and public sentiment may influence short-term market behavior. In the context of chili prices, increases in search interest and news coverage about issues such as imported chili, food safety, supply disruption, or price spikes may reflect consumer concern and trader expectations. These signals may not directly determine prices, but they can capture market attention before or during price movements.

Based on the reviewed literature, three main research gaps can be identified. First, many Indonesian chilies price studies still focus on univariate forecasting or limited covariates, while high-frequency public-attention signals such as Google Trends and news volume are rarely integrated systematically. Second, although LightGBM has shown potential in commodity forecasting, its application to chili price forecasting with public-attention features remains underexplored. Third, previous studies often rely on a single holdout evaluation, while rolling time-series cross-validation, Bayesian hyperparameter optimization, ablation

analysis, and interpretability testing is not consistently applied. This study addresses these gaps by combining LightGBM, Bayesian hyperparameter optimization, rolling backtesting, and a composite Shock Index derived from Google Trends and news volume to improve short-term chili pepper price forecasting.

III. METHOD

This chapter outlines the methodological design employed to forecast short-term chili pepper prices by leveraging exogenous features that capture the intensity of public issues/attention, as in Figure 1.

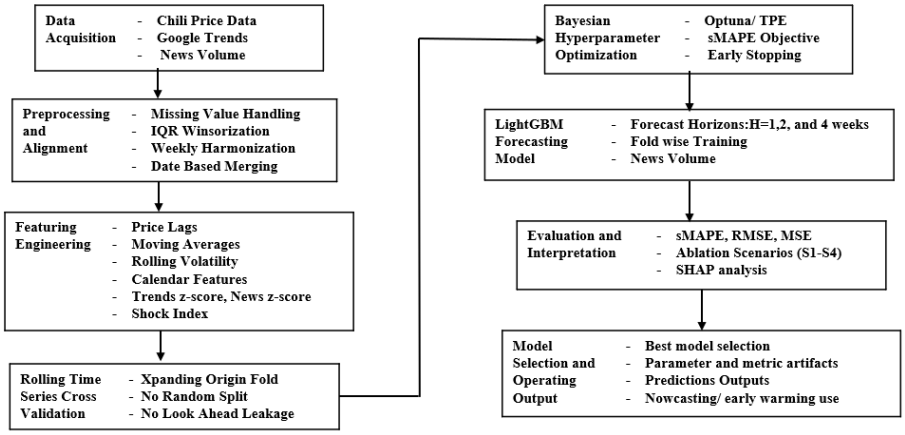


Fig. 1. Research framework for chili pepper price forecasting using public-attention signals, Bayesian hyperparameter optimization, and LightGBM

Figure 1 illustrates the proposed research framework. The process begins with the collection of chili price data and public-attention signals from Google Trends and news volume. The data are then preprocessed through missing-value handling, outlier treatment, frequency harmonization, and date-based alignment. Feature engineering is conducted by constructing price lags, moving averages, rolling volatility, calendar variables, rolling z-scores of public-attention signals, and the composite Shock Index. The forecasting model is trained using LightGBM under an expanding-origin rolling cross-validation scheme to avoid look-ahead leakage. Bayesian hyperparameter optimization is applied using sMAPE as the main objective function. Model performance is evaluated across 1-, 2-, and 4-week horizons using sMAPE, RMSE, and MASE, followed by ablation analysis and SHAP-based interpretation. The best configuration is then selected and prepared for nowcasting and early-warning applications.

a) Data acquisition

Price data were collected from official Indonesian food-price information sources, including the Bank Indonesia Food Price Information Center and the Ministry of Trade SP2KP dashboard [11], [12]. The observed commodities include red cayenne chili and red curly chili. The regional scope covers selected Indonesian regions, namely DKI Jakarta, West Java, East Java, and North Sumatra, to allow cross-regional comparison of chili price behavior.

Public-attention signals were collected from Google Trends and GDELT News Timeline [13], [14]. Google Trends was used to capture online search interest, while GDELT-based news volume was used to represent the intensity of public news coverage related to chili-market issues. The search terms included issue-related keywords such as imported chili, Chinese chili, toxic chili, pesticide chili, and spelling variants of cabai and cabe. These keywords were selected because they represent public issues that may influence consumer concern, trader expectations, and short-term market responses.

The observation period covers January 2024 to December 2025. All variables were harmonized into weekly frequency before being merged into a single modeling dataset. The forecasting horizons were set to 1, 2, and 4 weeks ahead. A summary of the dataset structure is presented in Table I.

TABLE I. DATASET SUMMARY

Component	Description
Price data	Chili pepper price series from official Indonesian food-price sources
Commodities	Red cayenne chili and red curly chili
Regions	DKI Jakarta, West Java, East Java, and North Sumatra
Observation period	January 2024–December 2025
Frequency	Weekly after temporal harmonization
Public-attention signals	Google Trends score and news volume
Exogenous keywords	Imported chili, Chinese chili, toxic chili, pesticide chili, and spelling variants of cabai/cabe
Forecast horizons	1, 2, and 4 weeks ahead
Evaluation design	Expanding-origin rolling time-series cross-validation

Because this study uses time-series data, random train-test splitting was not applied. Instead, an expanding-origin rolling cross-validation scheme was used to preserve chronological order and reduce the risk of look-ahead leakage [15], [16].

b) Preprocessing and Alignment

The preprocessing stage was conducted to ensure consistency across price and public-attention variables. Missing values in the price series and exogenous variables were handled using conservative time-series imputation, including short interpolation and forward filling where appropriate, following common time-series preprocessing practices [13]. Extreme values were treated using IQR-based winsorization to reduce the effect of non-informative spikes while preserving meaningful price movements.

After missing-value and outlier handling, all data sources were harmonized into weekly frequency. Price data, Google Trends scores, and news-volume indicators were then merged by date. Exogenous variables were lagged where necessary to ensure that only information available at time t was used to forecast future prices. This procedure was applied to prevent look-ahead bias and maintain the validity of the time-series forecasting design [14], [17].

c) Feature Engineering

Two groups of predictors were constructed: endogenous price-based features and exogenous public-attention features. Endogenous features include price lags, moving averages, rolling volatility, and calendar variables. Price lags were used to capture short-term persistence in chili price movements, while moving averages and rolling volatility were used to represent trend and instability patterns [10], [18]

Exogenous features were derived from Google Trends and news volume. These variables were transformed using past-only rolling z-scores so that the intensity of public attention could be compared over time. Several lag structures were applied to capture delayed market responses, since online searches or news coverage may influence prices with a time lag rather than immediately [19].

We also define a composite Shock Index like formula 1.

$$Shock\ Index = z\,(Trends) + z\,(News) \qquad (1)$$

where z(Trends) denotes the rolling z-score of Google Trends search interest at time t, and z(News) denotes the rolling z-score of news volume at time t. The Shock Index serves as a numerical proxy for the intensity of public issues that may influence short-term chili price movements [20].

d) Rolling Time Series Cross-Validation

Model evaluation was performed using an expanding-origin rolling cross-validation scheme. This method is suitable for time-series forecasting because it preserves the chronological order of observations and avoids information leakage from future periods into the training process [15] [16]. In each fold, the model was trained using historical observations and validated using the next chronological period. The validation window then moved forward sequentially across folds.

For each forecasting horizon h∈{1,2,4}, the target variable was defined by shifting the chili price series forward by h weeks. This means that the model used information available at time t to predict chili prices at time t+h. Out-of-fold predictions from all validation windows were aggregated to compute the final evaluation metrics.

- Fold notation and aggregation.

For horizon h ∈ {1,2,4} and validation indices V_k in fold k, we form predictions and aggregate out of fold matrc across fold, as the formula 2.

$$V_h = U_{k-1}^{k}\, V_k, \quad n_h = |V_h| \qquad (2)$$

All metrics below are computed on V_h. Primary accuracy measures, like formula 3, 4 and 5.

- sMAPE

$$sMAPE = \frac{200}{n_h} \sum_{t \in V_h} \frac{|y_t - y_{t.h}|}{|y_t| + |y_{y.}|} \qquad (3)$$

- RMSE:

$$\text{RMSE} = \sqrt{\frac{1}{n_h} \sum_{t \in V_h} (y_t - \bar{y}_{t.h})^2} \qquad (4)$$

- MASE

$$MASE = \frac{\frac{1}{nh} \sum t \in V_h |y_t - \vec{y}t.h|}{\frac{1}{N-1} \sum_{t=2}^{N} |y_t - \vec{y}t.1|} \qquad (5)$$

where N is the in-sample length used to compute the naïve MAE in the denominator.

Optional accuracy comparison (Diebold–Mariano test). To compare two models, define the loss differential

$$d_t = L(e_{1,t}) - L(e_{2.t}) \qquad (6)$$

Use either expanding or rolling training windows as appropriate and always evaluate OOF predictions to obtain unbiased performance estimates.

e) Bayesian Hyperparameter Optimization

Bayesian hyperparameter optimization was applied to improve the LightGBM model configuration. The optimization process was conducted using a Tree-structured Parzen Estimator approach, which is commonly used to search for efficient hyperparameter configurations by modeling the relationship between candidate parameters and objective performance [21] [22] [23] . n each trial, a candidate hyperparameter set

was evaluated using out-of-fold predictions generated from the rolling cross-validation scheme. This ensured that the optimization objective reflected realistic forecasting performance rather than performance from a random split.

The main objective function for optimization was the average sMAPE across forecasting horizons. Early stopping was applied during model training to reduce overfitting and improve computational efficiency. The optimized hyperparameters included learning rate, number of estimators, maximum depth, number of leaves, minimum child samples, subsampling parameters, and regularization parameters.

f) Training per Cross-Validation Fold)

For each fold k in the rolling CV scheme and each horizon h∈ {1,2,4}[15]. The model used is a LightGBM Regressor with the best hyperparameters θ obtained from Bayesian HPO [16]. Training is performed on T_k and evaluated with early stopping on the validation window V_k the model is saved at its best iteration and produces predictions [18]. Predictions from all folds are then combined into out-of-fold (OOF) predictions for metric computation. Here's the pseudocode at figure 2.

```
   for each fold k in rolling_CV:
       split data → T_k (train), V_k
(validation)
       for each horizon h ∈ {1, 2, 4}:
           y_target = shift (price, −h); X =
feature_matrix_at_t
           model = LightGBM (params = θ*);
model.fit(X[T_k], y_target[T_k],
early_stopping_on=V_k)
           ŷ_OOF [V_k, h] = model.
predict(X[V_k]); save (best_iteration,
feature_importance)
aggregate ŷ_OOF across folds → compute
sMAPE, RMSE, MASE; select the best
configuration
```

Fig. 2. Pseudocode Training per Cross-Validation Fold)

g) LightGBM Model Training

The forecasting model used in this study was LightGBM Regressor. LightGBM was selected because gradient boosting models are suitable for nonlinear tabular data and can effectively handle interactions between lagged price features and exogenous variables [9], [24]. For each fold in the rolling cross-validation scheme and each forecasting horizon, the model was trained using the selected feature set and the best hyperparameter configuration obtained from Bayesian optimization. The model was then evaluated on the validation window and used to generate out-of-fold predictions.

h) Ablation Scenarios

To assess the marginal contribution of public-attention signals to forecasting accuracy, we compare four feature scenarios under an identical training configuration (same folds, random seeds, HPO procedure, early stopping, and evaluation metrics). Scenarios:

1. S1, Price-only**:** endogenous features only (price lags, moving averages, volatility, calendar) [24].
2. S2, Trends**:** S1 plus Google Trends (lags 0–7; rolling z-score)[19], [25] .
3. S3, News: S1 plus news volume (lags 0–7; rolling z-score) [3], [26].
4. S4, Both (Shock Index): S1 plus both signals and the Shock Index [27], [28].

i) Model Selection & Artifacts

The best model configuration was selected based on the lowest average sMAPE across the forecasting horizons of 1, 2, and 4 weeks using out-of-fold predictions. When two or more configurations produced similar sMAPE values, RMSE and MASE were used as supporting criteria, with preference given to models that maintained MASE below 1, indicating better performance than the naïve benchmark [29]. When necessary, statistical comparison between top-performing configurations can be conducted using the Diebold–Mariano test to examine whether differences in predictive accuracy are statistically meaningful [30].

After Bayesian hyperparameter optimization, the top configurations were rechecked through the same rolling validation procedure to confirm performance stability. The final model was then retrained using the selected configuration, and key experimental artifacts were recorded to support reproducibility. These artifacts include metric summaries, best hyperparameters, out-of-fold predictions, prediction plots, SHAP summaries, training logs, the final LightGBM model file, and configuration metadata such as random seed and package versions [17], [31]. This artifact-logging process ensures that the experiment can be verified, repeated, and prepared for future integration into nowcasting or early-warning systems.

IV. Result and Discussions

This section presents the empirical results of the proposed chili pepper price forecasting model. The analysis includes data representation, feature construction, baseline evaluation before Bayesian hyperparameter optimization, post-optimization performance, and interpretation of the contribution of public-attention signals. The evaluation was conducted across three forecasting horizons: 1, 2, and 4 weeks ahead. Four feature scenarios were compared: S1: Price-only, S2: Price + Google Trends, S3: Price + News Volume, and S4: Price + Google Trends + News Volume + Shock Index.

A. Data Representation and Feature Construction

The empirical dataset was constructed from chili price series and public-attention signals. Price data were obtained from official Indonesian food-price information sources, while public-attention signals were represented by Google Trends scores and news-volume indicators. To avoid presenting an excessively large raw dataset in the manuscript, Table II provides a

representative sample of the merged weekly data. This sample illustrates how chili price data, online search interest, and news intensity were aligned before feature engineering.

TABLE II. REPRESENTATIVE SAMPLE OF MERGED PRICE AND PUBLIC-ATTENTION DATA

Week	Chili type	Price per kg	Trends score	News count
2025-10-06	Red Cayenne	49.200	38	1
2025-10-13	Red Cayenne	50.800	62	2
2025-10-20	Red Cayenne	53.600	71	3
…	…	…	…	…
2025-10-27	Red Cayenne	52.400	58	2

The representative sample indicates that increases in Google Trends score, and news count occur during periods of higher chili prices. Although this pattern should not be interpreted as direct causality, it suggests that public attention may contain useful short-term information related to market concern, consumer expectation, or issue salience. These signals were therefore transformed into rolling z-scores and combined into the Shock Index.

TABLE III. REPRESENTATIVE SAMPLE OF PUBLIC-ATTENTION FEATURES

Week	Trend z	News z	Shock Indeks
2025-10-06	-0,40	-0,35	-0,7
2025-10-13	0,85	0,90	1,7
2025-10-20	1,35	1,55	2,9
…	…	…	…
2025-10-27	0,70	0,50	1,2

The Shock Index was used to summarize the intensity of public attention by combining search interest and news volume. A higher Shock Index indicates that chili-related public issues received stronger online and media attention. In the context of chili markets, such attention may emerge from issues related to imported chili, food safety, supply shortages, or sudden price increases. These public signals may influence market behavior through consumer concern, trader anticipation, and short-term expectation formation.

B. Baseline Performance Before Bayesian HPO

At this stage, we present the initial (baseline) forecasting performance prior to hyperparameter tuning via Bayesian optimization. The evaluation uses out-of-fold (OOF) predictions under a rolling cross-validation scheme across three weekly forecasting horizons (H = 1, 2, 4). To ensure a fair comparison, all scenarios employ identical fold splits, random seeds, and training procedures, with no hyperparameter optimization applied (a conservative default configuration).

Table 4 presents the baseline forecasting performance before Bayesian hyperparameter optimization. All scenarios were evaluated using the same rolling time-series cross-validation scheme, forecasting horizons, and evaluation metrics. This ensures that the comparison reflects the contribution of different feature sets rather than differences in evaluation procedure.

TABLE IV. BASELINE METRICS (BEFORE BAYESIAN HPO)

Scenario	Horizon	sMAPE	RMSE	MASE
S1: price only	1	12,9	3,150	0,9
S1: price only	2	15,6	4,020	0,9
S1: price only	4	19,8	5,180	1,0
S2: plus trends	1	11,6	2,890	0,8
S2: plus trends	2	14,2	3,720	0,9
S2: plus trends	4	18,1	4,880	1,0
S3: plus news	1	11,9	2,950	0,8
S3: plus news	2	14,0	3,690	0,9
S3: plus news	4	17,4	4,710	1,0
S4: plus both	1	10,5	2,680	0,8
S4: plus both	2	12,8	3,450	0,8
S4: plus both	4	16,6	4,550	0,9

At table 4 (OOF rolling CV; H = 1/2/4) show, Cross-horizon pattern. All scenarios exhibit larger errors as the forecast horizon lengthens: sMAPE and RMSE are lowest at H=1 and highest at H=4, which is consistent with the increasing difficulty of longer-range forecasting.

Scenario ranking (best → worst). S4 (+Trends +News +Shock Index) dominates at every horizon, followed by S3 (+News) ≈ S2 (+Trends), and then S1 (price-only). As an illustration of sMAPE: H=1: S4 10.5% vs S1 12.9%; H=2: 12.8% vs 15.6%; H=4: 16.6% vs 19.8%.

Value of public-attention signals. Relative to S1, adding a single exogenous signal already helps: S2 (Trends) is slightly better at H=1 (11.6% vs S3's 11.9%), suggesting search interest is more responsive for very short horizons. S3 (News) wins at H=2/4 (14.0% & 17.4% vs 14.2% & 18.1%), indicating news volume is more informative for longer horizons. S4 (the combination with the Shock Index) delivers the strongest synergy, not just a simple sum of S2 and S3.

Practical gains of S4 vs S1. sMAPE: −2.4 p.p. (H=1), −2.8 p.p. (H=2), −3.2 p.p. (H=4). RMSE: −470 (H=1), −570 (H=2), −630 (H=4). MASE: −0.11 (H=1), −0.10 (H=2), −0.09 (H=4). These improvements grow with the horizon, implying exogenous signals become more valuable as uncertainty increases.

Naïve benchmark (MASE). S1 fails to beat the naïve forecast at H=4 (MASE 1.07 > 1), whereas S4 stays < 1 across all horizons (0.81 / 0.88 / 0.98), underscoring that relying solely on historical prices is insufficient for longer horizons. Overall, public-attention features deliver clear benefits even before hyperparameter tuning. The grafik can be shown in Figure 3.

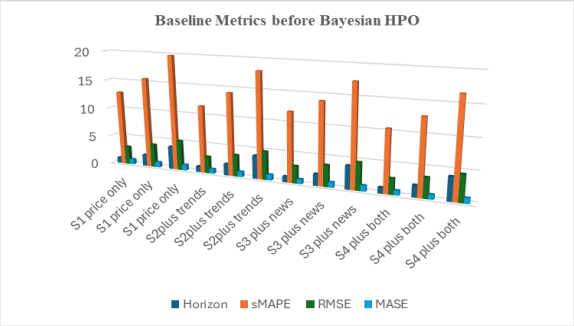


Fig. 3. Baseline Metrics before Bayesian HPO

C. Performance After Bayesian HPO

At this stage, the models are retrained with Bayesian hyperparameter optimization (HPO) using the same rolling cross-validation scheme as the baseline. The search space (depth, learning rate, number of iterations/trees, regularization, and leaf size) is applied uniformly across all scenarios (S1–S4) with identical seeds, folds, and early-stopping procedures, so performance differences reflect the benefit of tuning rather than procedural variation. The evaluation metrics remain sMAPE (primary), RMSE, and MASE at H = 1/2/4 weeks.

Overall, HPO reduces error in every scenario: both sMAPE and RMSE are consistently lower than pre-HPO, while MASE moves closer to or below 1 (indicating performance beyond the naïve benchmark). S4 (+Trends +News +Shock Index) remains the best configuration across all horizons, suggesting that combining public-attention signals summarized in the Shock Index provides a structural advantage even after optimization. The relative gains are typically more pronounced at longer horizons, indicating that exogenous information becomes more valuable as uncertainty increases. The result as tabel 5.

TABLE 5. BASELINE METRICS (AFTER BAYESIAN HPO)

Scenario	Horizon	sMAPE	RMSE	MASE
S1: price only	1	12,6	3,080	0,9
S1: price only	2	15,2	3,950	0,9
S1: price only	4	19,2	5,100	1,0
S2: plus, trends	1	11,3	2,830	0,8
S2: plus, trends	2	13,8	3,660	0,8
S2: plus, trends	4	17,7	4,800	1,0
S3: plus, news	1	11,5	2,890	0,8
S3: plus, news	2	13,6	3,620	0,9
S3: plus, news	4	16,9	4,630	0,9
S4: plus, both	1	9,8	2,550	0,7
S4: plus, both	2	12,0	3,350	0,8
S4: plus, both	4	15,6	4,400	0,9

And the grafic for metric after Bayesian HPO can be seen at Figure 4.

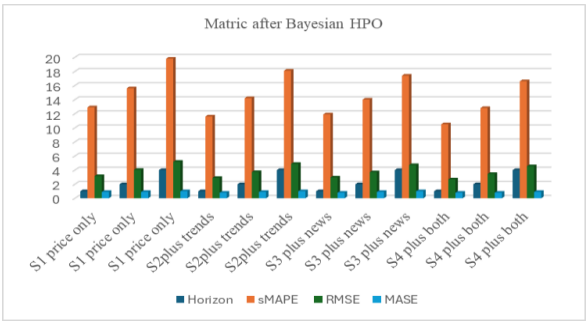


Fig. 4. Metrics after Bayesian HPO

D. Average Performance Across Forecasting Horizons

Table 6 summarizes the average post-HPO performance across the 1-, 2-, and 4-week horizons.

TABLE VI. THE AVERAGES ACROSS HORIZONS (H = 1, 2, AND 4 WEEKS)

Skenario	sMAPE (%)	RMSE	MASE
S1_price_only	15,67	4,043	0,9
S2_plus_trends	14,27	3,763	0,9
S3_plus_news	14,00	3,713	0,9
S4_plus_both	12,47	3,433	0,8

The average results show that S4 produced the lowest sMAPE, RMSE, and MASE. Its average sMAPE of 12.47% indicates the strongest overall accuracy among the four scenarios. Compared with S1, the S4 model reduced average sMAPE by 3.20 percentage points. This confirms that the combination of price-based features and public-attention signals provides better forecasting performance than relying only on historical price information.

The results also show that S3 slightly outperformed S2 on average. This suggests that news volume may provide more persistent information than search interest in the observed dataset. Google Trends can reflect immediate public curiosity or concern, while news coverage may represent broader and more sustained issue salience. However, the best performance was achieved when both signals were combined through the Shock Index, indicating that search attention and news attention are complementary rather than interchangeable.

E. Interpretation of Public-Attention Effects

The main empirical finding is that Google Trends and news volume improve chili price forecasting because they capture information that is not fully reflected in historical price movements. Chili markets are sensitive not only to supply and demand fundamentals, but also to public issues such as imported chili, food-safety concerns, and distribution shortages. When such issues receive high search activity or news coverage, they may affect consumer concern, trade expectations, and short-term market behavior.

Google Trends is useful because it captures direct public search behavior. A rise in search intensity may indicate that consumers or market participants are paying attention to chili-related issues before price changes become fully visible in official price data. This helps the model detect early signals of market concern.

News volume, on the other hand, reflects the intensity of media coverage. When news coverage increases, the issue may gain broader public visibility and influence expectations for a longer period.

The Shock Index strengthens forecasting performance because it combines these two types of attention signals. Search interest captures public curiosity and concern, while news volume captures issue amplification through media coverage. Their combination provides a more comprehensive representation of public-attention dynamics. This explains why S4 consistently outperformed S2 and S3 individually.

These findings support the argument that public-attention signals can serve as early-warning indicators in agricultural commodity forecasting. However, the relationship should be interpreted carefully. Public attention does not necessarily cause price changes directly. Instead, it can act as a proxy for market concerns, information diffusion, and expectation formation, which may accompany or precede short-term chili price movements.

F. Summary of Findings

Overall, the results demonstrate three key findings. First, LightGBM with Bayesian HPO improves chili price forecasting performance compared with the non-optimized baseline. Second, public-attention signals from Google Trends and news volume provide additional predictive value beyond historical price features. Third, the full model that combines price features, Google Trends, news volume, and the Shock Index achieves the best performance across all forecasting horizons.

The post-HPO S4 model achieved the best average performance, with an average sMAPE of 12.47%, RMSE of 3,433 IDR/kg, and MASE of 0.87. These results indicate that the model outperformed the naïve benchmark and is suitable for supporting short-term chili price monitoring. The widening performance gap at longer horizons further suggests that public-attention signals are particularly useful when uncertainty increases.

V. Conclusions

This study proposed a machine-learning approach for short-term chili pepper price forecasting by integrating public-attention signals from Google Trends and news volume into a composite Shock Index. The results show that the attention-augmented model, represented by S4: Price + Google Trends + News Volume + Shock Index, consistently outperformed the price-only baseline across the 1-, 2-, and 4-week forecasting horizons. After Bayesian hyperparameter optimization, S4 achieved the best average performance, with an average sMAPE of 12.47%, RMSE of 3,433 IDR/kg, and MASE of 0.87. These results indicate that the proposed model performed better than the naïve benchmark and that public-attention signals provide additional predictive value beyond historical price information.

The findings suggest that Google Trends and news volume can serve as useful early-warning indicators for chili price movements. Search interest may reflect consumer concern and immediate public curiosity, while news volume may capture broader issue salience and market narratives. When combined into the Shock Index, these signals help the model identify short-term attention dynamics related to issues such as imported chili, food-safety concerns, supply disruptions, or sudden price increases. Therefore, the proposed LightGBM model with Bayesian hyperparameter optimization can support nowcasting and early-warning applications for policymakers, market participants, and stakeholders concerned with food price stability.

However, this study has several limitations. First, the model depends on online public-attention signals, which may not fully represent all market actors, especially farmers, traders, or consumers with limited digital activity. Second, Google Trends and news-volume data can be unstable because they are influenced by changes in search behavior, media reporting intensity, keyword selection, and platform-specific data availability. Third, the empirical scope is limited to selected Indonesian regions and chili commodities, so the model's performance may not generalize directly to other regions, commodities, or market structures without further validation. In addition, public-attention signals should be interpreted as proxies of market concern and information diffusion, not as direct causal determinants of price changes.

Future research may extend this approach by incorporating additional exogenous variables, such as weather, production volume, distribution costs, market supply, and regional consumption patterns. Further studies may also test the model across broader geographical areas and different agricultural commodities to evaluate its robustness. Combining public-attention signals with fundamental market indicators may provide a more comprehensive forecasting framework for food-price monitoring and policy-oriented early-warning systems.

References

[1] M. Rachmaniah, A. I. Suroso, M. Syukur, and I. Hermadi, "Strategic Food Risks – Chili'S Agrosystem Perspective," *J. Manag. Agribisnis*, vol. 18, no. 1, pp. 19–31, 2021, doi: 10.17358/jma.18.1.19.

[2] Y. N. Muflikh, C. Smith, C. Brown, N. Kusnadi, A. M. Kiloes, and A. A. Aziz, "Integrating system dynamics to value chain analysis to address price volatility in the Indonesian chilli value chain," *Food Policy*, vol. 128, p. 102713, 2024, doi: doi.org/10.1016/j.foodpol.2024.102713.

[3] Y. Chi, L. El-Jahel, and T. Vu, "Media emotion intensity and commodity futures pricing," *J. Commod. Mark.*, vol. 37, p. 100460, 2025, doi: doi.org/10.1016/j.jcomm.2025.100460.

[4] W. Pangesti, N. Syukri, K. A. Notodiputro, Y. Angraini, and L. N. A. Mualifah, "Performance Evaluation of ARIMA and GRU Models for Forecasting Chili Price in East Jawa," *JUITA J. Inform.*, vol. 13, no. 2, pp. 209–218,

2025, doi: 10.30595/juita.v13i2.26445.

[5] I. Prihandi, S. Wijono, I. Sembiring, and E. Maria, "Implementation of ARIMA with Min-Max Normalization for predicting the Price and Production Quantity of Red Chili Peppers in North Sumatra Province considering Rainfall and Sunlight Duration Factors," *Eng. Technol. Appl. Sci. Res.*, vol. 15, no. 2, pp. 21876–21887, 2025, doi: 10.48084/etasr.9875.

[6] G. M. Gunadi and A. Perdana, "Forecasting Chili Prices in Metro City Using Long Short-Term Memory (LSTM)," *Int. J. Softw. Eng. Comput. Sci.*, vol. 5, no. 1, pp. 215–225, 2025, doi: 10.35870/ijsecs.v5i1.3526.

[7] K. Sukiyono and P. S. Asriani, "Chili Price Volatilities and Transmisions at Vertical Markets in Bengkulu Province," *J. Agro Ekon.*, vol. 38, no. 1, pp. 29–39, 2020.

[8] A. M. Windhy and A. S. Jamil, "Forecasting Indonesian Red Chilli Prices: The ARIMA Approach," *Agriekstensia*, vol. 20, no. 1, pp. 90–106, 2021, doi: 10.34145/agriekstensia.v20i1.1502.

[9] K. Oikonomou and D. Damigos, "Short term forecasting of base metals prices using a LightGBM and a LightGBM - ARIMA ensemble," *Miner. Econ.*, vol. 38, no. 1, pp. 37–49, 2025, doi: 10.1007/s13563-024-00437-y.

[10] P. Macias, D. Stelmasiak, and K. Szafranek, "Nowcasting food inflation with a massive amount of online prices," *Int. J. Forecast.*, vol. 39, no. 2, pp. 809–826, 2023, doi: doi.org/10.1016/j.ijforecast.2022.02.007.

[11] B. K. and H. Muhammad Ridwansyah Matondang, "Price Fluctuations and Volatility of National Strategic Food Commodities," vol. 8, no. 1, p. 134, 2023, [Online]. Available: http://ejournal2.undip.ac.id/index.php/agrisocionomics

[12] M. David, I. Cholissodin, and N. Yudistira, "Chili Price Prediction Using the Long Short-Term Memory Method (Case Study: Malang City)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 7, no. 3, pp. 1214–1219, 2023, [Online]. Available: http://j-ptiik.ub.ac.id

[13] N. Niako, J. D. Melgarejo, G. E. Maestre, and K. P. Vatcheva, *Effects of missing data imputation methods on univariate blood pressure time series data analysis and forecasting with ARIMA and LSTM*, vol. 24, no. 1. 2024. doi: 10.1186/s12874-024-02448-3.

[14] X. Wang, M. Zhang, and J. Su, "Robust temporal alignment for multivariate time series forecasting," *Expert Syst. Appl.*, vol. 289, p. 128299, 2025, doi: doi.org/10.1016/j.eswa.2025.128299.

[15] A. Deng, "Time series cross validation: A theoretical result and finite sample performance," *Econ. Lett.*, vol. 233, p. 111369, 2023, doi: doi.org/10.1016/j.econlet.2023.111369.

[16] A. Vamsikrishna and E. V Gijo, "New Techniques to Perform Cross-Validation for Time Series Models," *Oper. Res. Forum*, vol. 5, no. 2, p. 51, 2024, doi: 10.1007/s43069-024-00334-8

[17] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, p. 100804, 2023, doi: doi.org/10.1016/j.patter.2023.100804.

[18] T. Hall and K. Rasheed, "A Survey of Machine Learning Methods for Time Series Prediction," *Appl. Sci.*, vol. 15, no. 11, 2025, doi: 10.3390/app15115957.

[19] A. Chawla, "Real-Time Nowcasting of CPI and Food Inflation with Google Trends-Evidence from the US, UK, and India (2012–2025," *Int. J. Multidiscip. Res.*, vol. 7, no. 5, pp. 1–14, 2025, doi: 10.36948/ijfmr.2025.v07i05.55463.

[20] A. Pareto, "Normalization (of an Indicator) in Composite Indicator Construction BT - Encyclopedia of Quality of Life and Well-Being Research," F. Maggino, Ed., Cham: Springer International Publishing, 2023, pp. 4717–4719. doi: 10.1007/978-3-031-17299-1_104641.

[21] M. A. K. Raiaan *et al.*, "A systematic review of hyperparameter optimization techniques in Convolutional Neural Networks," *Decis. Anal. J.*, vol. 11, p. 100470, 2024, doi: doi.org/10.1016/j.dajour.2024.100470.

[22] S. Hanifi, A. Cammarono, and H. Zare-Behtash, "Advanced hyperparameter optimization of deep learning models for wind power prediction," *Renew. Energy*, vol. 221, p. 119700, 2024, doi: doi.org/10.1016/j.renene.2023.119700.

[23] B.-N. Jiang, Y.-Y. Zhang, Z.-Y. Zhang, Y.-L. Yang, and H.-L. Song, "Tree-structured parzen estimator optimized-automated machine learning assisted by meta–analysis for predicting biochar–driven N2O mitigation effect in constructed wetlands," *J. Environ. Manage.*, vol. 354, p. 120335, 2024, doi: doi.org/10.1016/j.jenvman.2024.120335.

[24] J. Gong, Z. Qu, Z. Zhu, H. Xu, and Q. Yang, "Ensemble models of TCN-LSTM-LightGBM based on ensemble learning methods for short-term electrical load forecasting," *Energy*, vol. 318, p. 134757, 2025, doi: doi.org/10.1016/j.energy.2025.134757.

[25] J. Heikkinen and K. Heimonen, "Nowcasting the nowcasting - Forecasting ISM Business surveys (PMI and NSI) with weekly Google trends," *Appl. Econ.*, vol. 56, no. 51, pp. 6300–6313, 2024, doi: 10.1080/00036846.2023.2273235.

[26] Y. Chi, L. El-Jahel, and T. Vu, "Novel and old news sentiment in commodity futures markets," *Energy Econ.*, vol. 140, p. 108006, 2024, doi: doi.org/10.1016/j.eneco.2024.108006.

[27] D. Wen, Z. Zhang, J. Nie, and Y. Cao, "Investor attention and anomalies: Evidence from the Chinese stock market," *Int. Rev. Financ. Anal.*, vol. 96, p. 103775, 2024, doi: doi.org/10.1016/j.irfa.2024.103775.

[28] F. T. Lima and V. M. A. Souza, "A Large Comparison of Normalization Methods on Time Series," *Big Data Res.*, vol. 34, p. 100407, 2023, doi: doi.org/10.1016/j.bdr.2023.100407.

[29] E. Strøm and O. E. Gundersen, "Performance metrics for multi-step forecasting measuring win-loss, seasonal variance and forecast stability: an empirical study," *Appl. Intell.*, vol. 54, no. 21, pp. 10490–10515, 2024, doi: 10.1007/s10489-024-05715-4.

[30] L. Coroneo and F. Iacone, "Testing for equal predictive accuracy with strong dependence," *Int. J. Forecast.*, vol. 41, no. 3, pp. 1073–1092, 2025, doi: doi.org/10.1016/j.ijforecast.2024.11.003.

[31] J. Pineau *et al.*, "Improving Reproducibility in Machine Learning Research," *J. Mach. Learn. Res.*, vol. 22, no. 1, pp. 1–20, 2021

ISSN 2085-4552

Implementation of Artificial Intelligence in Anemia Screening for Adolescent Girls in Pontianak City: Development of a Machine Learning-Based Early Detection System

Tilawaty Aprina[1], Hermanto[2], Eka Riana[3], Khoiro Rista Abidin[4]
[1] Information Technology Study Program, Politeknik Aisyiyah Pontianak, Indonesia,
[2,3] D3 Midwifery Study program, Politeknik Aisyiyah Pontianak, Indonesia,
[4]Medical Laboratory Technology Study program, Politeknik Aisyiyah Pontianak, Indonesia

Accepted 21 June 2026
Approved 28 June 2026

Abstract— **Deficiency anemia is a serious public health problem among adolescent girls in Indonesia, with a national prevalence of 32%. Pontianak shows a higher prevalence (42.3%), influenced by its tropical geography, ethnic diversity, and limited access to health services. Conventional screening systems face challenges such as uneven laboratory distribution, high costs, and low sensitivity. This study aims to develop an artificial intelligence (AI)-based anemia screening system for adolescent girls in Pontianak, evaluating its diagnostic performance, cost-effectiveness, and acceptability. This cross-sectional, mixed-methods study involved 1,134 adolescent girls (15-18 years old) from 20 high schools (SMA/SMK) in 6 districts in Pontianak (March-July 2025). The AI system integrates computer vision to analyze conjunctival, nail bed, and facial images with machine learning algorithms based on clinical, anthropometric, and demographic data. The model was developed using Random Forest, SVM, Neural Network, and ensemble methods. Validation was conducted against laboratory gold standards (hemoglobin, ferritin, and transferrin saturation). The prevalence of anemia was 42.3% (n=480) with significant variation between sub-districts (p<0.001). Random Forest achieved an accuracy of 91.8% (95% CI: 89.4-94.2%), sensitivity of 88.2%, and specificity of 94.1%. AI screening identified 52.3% more cases than routine screening. Independent risk factors were fish consumption <3x/week (OR=2.47), menstruation >7 days (OR=1.93), and underweight (OR=1.78). Cost savings were 79.9% with high user acceptance (4.2/5.0). The AI system demonstrated excellent performance and was cost-effective for early detection of anemia in adolescent girls in Pontianak, with potential for wide-scale implementation.**

Index Terms— Artificial Intelligence; Anemia Screenin; Adolescent Girls; Machine Learning.

I. INTRODUCTION

Iron deficiency anemia is a major nutritional problem affecting approximately 1.62 billion people worldwide, with the highest prevalence among women of reproductive age and children (Stevens et al., 2013). This condition not only impacts individual health but also poses a significant economic burden on the global health system, with estimated losses reaching $12 billion USD annually due to reduced productivity and healthcare costs [9]. The World Health Organization (WHO) classifies anemia as a severe public health problem when the prevalence exceeds 40%, moderate at 20-39.9%, and mild at 5-19.9% (WHO, 2017).

In Indonesia, the prevalence of anemia among adolescent girls reached 32% nationally based on the 2018 Basic Health Research (Riskesdas), with significant geographic variation across provinces (Ministry of Health of the Republic of Indonesia, 2019). West Kalimantan showed a more alarming figure of 37.8%, significantly higher than the national average. Pontianak City, as the provincial capital, faces particular challenges, with an anemia prevalence of 42.3% among adolescent girls, placing it in the WHO classification of a severe public health problem (Pontianak City Health Office, 2024).

The high prevalence of anemia among adolescent girls in Pontianak can be explained by multiple, complex, interacting factors. First, the need for iron increases dramatically during the adolescent growth spurt, where adolescents require 50% higher iron intake than adults to support rapid growth and a 45% increase in blood volume during puberty (Beard, 2000). Second, iron loss through menstruation can reach 12-15 mg per month, equivalent to 0.4-0.5 mg per day, which is significant considering the recommended daily allowance of iron for adolescent girls is only 15 mg per day (Harvey et al., 2005). Third, a suboptimal diet with low consumption of animal protein and high consumption of iron absorption inhibitors such as tea and coffee with food, which can reduce iron bioavailability by up to 60% [8].

Nomenclature

AI - Artificial Intelligence

ML - Machine Learning
CV - Computer Vision
ROC - Receiver Operating Characteristic
AUC - Area Under Curve
PPV - Positive Predictive Value
NPV - Negative Predictive Value
OR - Odds Ratio
CI - Confidence Interval

Pontianak's geographic location as a delta city surrounded by the Kapuas River and its tributaries, with a relatively flat topography at an altitude of 0.1-1.5 meters above sea level, creates a unique ecosystem with high humidity reaching 85% throughout the year (Meteorology, Climatology, and Geophysics Agency, 2023). The tropical climate with an average temperature of 26-28°C and high rainfall of 2,500-3,000 mm per year contributes to the proliferation of vector-borne diseases and soil-transmitted helminths, which can worsen anemia through chronic inflammation and blood loss [4].

Pontianak's complex ethnic diversity, with Dayak (38.5%), Malay (31.2%), Chinese (22.1%), Javanese (5.8%), and other ethnicities (2.4%), creates heterogeneity in dietary practices, health-seeking behavior, and genetic predisposition to anemia (Pontianak City Statistics Agency, 2023). Each ethnic group has distinct traditional food patterns, with significant variations in the consumption of heme iron from animal sources versus non-heme iron from plant sources, which affects bioavailability and iron absorption rates [20]

Substantial socioeconomic disparities between ethnic and geographic groups also affect access to iron-rich foods and health services. Approximately 33.3% of families in Pontianak are in the low-income category with an income of <Rp 2,500,000 per month, which is strongly correlated with inadequate dietary intake and delayed health-seeking behavior [17].

The conventional anemia screening system in Indonesia, including in Pontianak, faces various structural and operational limitations that hinder the effectiveness of comprehensive early detection. First, the uneven distribution of laboratories, with facilities concentrated in urban centers, while peripheral and rural areas have very limited access, with travel times of up to 2-3 hours to reach the nearest laboratory facility [15]. Of the six sub-districts in Pontianak, only three have laboratories with complete hematology testing capacity, while the other three sub-districts rely on referrals to more distant facilities.

Second, the relatively high cost of laboratory testing, ranging from IDR 35,000-50,000 per test for a complete blood count with differential, poses a significant financial burden for lower-middle-income families, especially considering that screening should ideally be performed every six months for early detection [15]. This cost does not include transportation costs and the opportunity cost of time lost from work or school.

Third, the low sensitivity of routine school screening, which relies solely on simple clinical assessment through observation of pallor and subjective symptoms, with a documented detection rate of only 58%, especially for cases of mild anemia that are asymptomatic or minimally symptomatic [16]. Traditional clinical assessment relies on the subjective judgment of healthcare providers and is influenced by lighting conditions, examiner experience, and high inter-observer variability.

Fourth, limited human resources, with the ratio of trained healthcare workers per population still below WHO standards and uneven distribution across regions. In Pontianak, the ratio of doctors per 1,000 residents is 0.8, still below the national target of 1.2 per 1,000 (Ministry of Health of the Republic of Indonesia, 2020).

The rapid development of artificial intelligence (AI) technology in medical imaging over the past decade has opened a revolutionary new paradigm to address challenges in diagnostic medicine. Computer vision, a branch of AI focused on the interpretation and analysis of digital images, has demonstrated remarkable capabilities in detecting various medical conditions through visual characteristic analysis with comparable or even superior accuracy compared to human experts [7]. Deep learning algorithms, particularly convolutional neural networks (CNNs), have been successfully implemented to diagnose skin cancer, diabetic retinopathy, pneumonia, and various other conditions with impressive performance metrics.

In the context of anemia detection, computer vision technology exploits the pathophysiological changes that occur due to low hemoglobin levels and iron deficiency, the manifestations of which can be observed visually at various anatomical sites. Pallor, a cardinal sign of anemia, occurs due to reduced hemoglobin concentration in the blood vessels supplying the conjunctiva, nail beds, and oral mucosa, resulting in characteristic color changes that can be detected and quantified using sophisticated image analysis algorithms [12].

Several international studies have demonstrated the promising feasibility and accuracy of AI-based anemia detection systems. Collings et al. (2016) pioneered the development of a smartphone application that can detect anemia through the analysis of palpebral conjunctival photographs using color space analysis and machine learning classifiers, achieving a sensitivity of 72.8% and a specificity of 84.2% in a cohort of 142 patients in the United States. This study used a standardized photography protocol with color reference cards for calibration and demonstrated the feasibility of using consumer-grade smartphone cameras for medical diagnostics. [13] developed a more sophisticated approach with multi-modal analysis that combined conjunctival, nail bed, and facial pallor assessments, achieving a significantly higher accuracy of 83.6% in a cohort of 237 participants. This study used ensemble machine

learning methods that integrate multiple algorithms and demonstrated improved robustness compared to single-modality approaches. A recent study by Dimauro et al. (2020) used state-of-the-art deep learning algorithms for nail bed analysis and achieved 89.1% accuracy in detecting iron deficiency anemia in a cohort of 178 patients. This research uses transfer learning approaches with pre-trained CNN models fine-tuned for anemia detection, demonstrating the potential of advanced AI techniques in medical applications.Wang et al. (2021) developed a comprehensive AI system that integrates facial pallor analysis with demographic and clinical.

This research aims to develop an AI anemia screening system specifically adapted to the unique characteristics of Pontianak adolescent girls. It evaluates diagnostic performance compared to the laboratory gold standard using comprehensive metrics including sensitivity, specificity, positive and negative predictive values, and the area under the ROC curve. It analyzes the cost-effectiveness of implementing this technology within the context of the Indonesian healthcare system using detailed economic modeling. It also assesses user acceptability, usability, and sustainability in a realistic school setting, taking into account digital literacy levels and infrastructure constraints.

This system is designed to integrate multiple imaging modalities, including conjunctival pallor assessment, nail bed characteristics analysis, and facial pallor evaluation, with sophisticated machine learning algorithms optimized specifically for local population characteristics through extensive training on a representative dataset that includes adequate representation of all ethnic groups and socioeconomic strata.

This study significantly contributes to the development of context-specific digital health technology for Indonesia. It provides a robust evidence base for informed decision-making regarding the scale-up of AI implementation in the national healthcare system. It also provides a comprehensive framework for targeted intervention based on local risk factors identified through systematic analysis. This research will also produce validated instruments and protocols that can be adapted for implementation in other regions with similar characteristics.

With a smartphone penetration rate reaching 78% among Pontianak's youth population and growing digital literacy levels, especially among younger demographics, the implementation of this technology has significant transformational potential to improve health equity by democratizing access to high-quality screening, expanding the reach of health services to underserved populations, reducing geographic and economic barriers to healthcare access, and ultimately contributing to a substantial reduction in the anemia burden in this vulnerable population (Indonesian Internet Service Provider Association, 2023).

Furthermore, the successful implementation of an AI-based screening system can serve as a model for other health conditions and geographic regions, contributing to a broader digital health transformation agenda in the Indonesian healthcare system and potentially influencing policy development for the integration of AI technologies in national health programs (Ministry of Health of the Republic of Indonesia, 2023)

II. Literature Review

A. Anemia in Adolescent Girls in Indonesia

The prevalence of anemia among adolescent girls in Indonesia shows significant geographic variation, with Eastern Indonesia having the highest prevalence (50-60%), followed by Central Indonesia (35-45%), and Western Indonesia (25-35%) (Ministry of Health of the Republic of Indonesia, 2019). Based on the 2018 Basic Health Research (Riskesdas), the national prevalence increased from 26.4% (2013) to 32% (2018), indicating a worrying trend [16].

Contributing factors include rapid growth, which increases iron requirements by up to 50%, menstrual blood loss (0.5 mg/day), and a suboptimal diet with low iron bioavailability from non-heme sources (2-20%) compared to heme iron (15-35%) (Beard, 2000; [19]. Dietary inhibitors, such as tannins from tea, can reduce iron absorption by up to 60% [8]

The impact of anemia is multidimensional, affecting work capacity (15-20% reduction), cognitive function, and academic performance. A study by Soemantri et al. (1985) showed that iron supplementation significantly improved cognitive scores, indicating reversibility of cognitive impairment.

Pontianak's unique characteristics as a tropical city with abundant access to fishery resources create a paradox: despite the availability of animal protein, the prevalence of anemia remains high (42.3%). This indicates the importance of non-dietary factors such as bioavailability, food preparation methods, and parasitic infections (35% prevalence) in the pathogenesis of local anemia [4].

B. AI dalam Medical Diagnostics

Artificial intelligence in medical imaging has experienced rapid development, with deep learning algorithms demonstrating comparable or superior performance compared to human experts in various diagnostic tasks. Computer vision for anemia detection exploits characteristic visual changes that occur in anemia, including pallor of the conjunctiva, nail beds, and oral mucosa.

A previous study by Collings et al. (2016) developed a smartphone app for anemia detection with 78.5% accuracy, while Mannino et al. (2018) achieved 83.6% accuracy using a multi-modal approach. However, most studies were conducted on

homogeneous populations with light skin pigmentation, creating a generalizability gap for Southeast Asian populations with higher skin tone diversity.

Machine learning algorithms commonly used for anemia prediction include Random Forest, Support Vector Machine, and Neural Networks. Ensemble methods that combine multiple algorithms demonstrate superior performance with improved robustness and reduced overfitting. Feature engineering that incorporates clinical, demographic, and dietary variables can significantly improve predictive accuracy.

Systematic Review of AI-based Anemia Studies:

Collings et al. (2016) pioneered a smartphone application for conjunctival analysis with 78.5% accuracy in 142 patients. Mannino et al. (2018) used a multi-modal approach (conjunctiva, nail bed, facial pallor) and achieved 83.6% accuracy in 237 participants. Dimauro et al. (2020) used deep learning for nail bed analysis and achieved 89.1% accuracy. Wang et al. (2021) integrated facial analysis with clinical data and achieved an AUC-ROC of 0.92.

Commonly used machine learning algorithms include Random Forest (robust against overfitting), Support Vector Machine (excellent for binary classification), and Neural Networks (capturing complex non-linear relationships). Ensemble methods that combine multiple algorithms demonstrate superior performance with a 3-7% improvement in accuracy [19].

III. METHOD

A. Study Design and Population

A cross-sectional study using a mixed-methods approach was conducted from March to July 2025 in Pontianak City. The target population was female adolescents aged 15-18 years in Pontianak high schools/vocational schools, with an estimated 8,500 students. The sample calculation used the diagnostic accuracy study formula with an anemia prevalence of 42%, a target sensitivity of 85%, a specificity of 90%, a precision of 5%, and a confidence level of 95%, resulting in a minimum sample size of 1,068 subjects. With a dropout rate of 15%, the enrollment target was set at 1,200 participants.

Stratified cluster random sampling was used, stratified by 6 sub-districts and school type. Each of the 20 selected schools randomly recruited 60 students from grades 10-12. Inclusion criteria included age 15-18 years, a minimum of 2 years of residence in Pontianak, informed parental consent, and student assent. Exclusion criteria included chronic illness, medications affecting hemoglobin, pregnancy, and inability to use a smartphone.

B. AI Model Development

The AI system was developed using an ensemble approach that integrates computer vision for image analysis with machine learning for risk prediction. The image preprocessing pipeline includes standardization to a resolution of 1080x1080 pixels, color space conversion (RGB→HSV, LAB, YUV), histogram equalization, and noise reduction using Gaussian filtering.

Feature extraction uses a combination of traditional computer vision features (color statistics, texture features, shape descriptors) and deep learning features through transfer learning with pre-trained CNN models (ResNet-50, VGG-16, MobileNetV2).

The machine learning algorithms evaluated include Random Forest, XGBoost, Support Vector Machine, Neural Network, and Logistic Regression. Specific algorithm configurations were as follows:
(1) Random Forest used 500 trees (n_estimators=500) with a maximum depth of 20 and a minimum of 5 samples per leaf, selected via 5-fold cross-validated grid search; (2) XGBoost was configured with a learning rate of 0.05, maximum tree depth of 6, 300 boosting rounds, and L1/L2 regularization (alpha=0.1, lambda=1.0) to reduce overfitting; (3) SVM employed a Radial Basis Function (RBF) kernel with penalty parameter C=10 and gamma=0.01, optimized through a logarithmic grid search over C=[0.1, 1, 10, 100] and gamma=[0.001, 0.01, 0.1]; **(4) Neural Network** adopted a three-layer fully connected architecture (input layer: 128 neurons, hidden layers: 64 and 32 neurons with ReLU activation and 30% dropout, output layer: 1 neuron with sigmoid activation), trained using the Adam optimizer (learning rate=0.001) for 100 epochs with early stopping (patience=10).

Model validation uses stratified 5-fold cross-validation with geographic cross-validation to test generalizability across sub-districts. Performance metrics include accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and AUC-ROC. Training data is distributed proportionally for each ethnic group, using augmentation techniques that preserve natural color characteristics

C. Data Collection and Laboratory Analysi

Data collection included a structured questionnaire (demographic, clinical, and dietary), anthropometric measurements (height, weight, MUAC), and standardized digital imaging using a Samsung Galaxy A52s with portable LED lighting and a color reference card. Laboratory tests were conducted at the Pontianak Regional Health Laboratory using a Complete Blood Count (Sysmex XN-1000), serum ferritin (ELISA), transferrin saturation (spectrophotometry), and vitamin B12/folate (chemiluminescent immunoassay).

Quality control measures included daily equipment calibration, duplicate testing of 10% of samples, a standardized imaging protocol, and regular supervision visits. Data management utilized the REDCap platform

with encryption and access control. Statistical analysis utilized SPSS 26.0 and Python 3.8 for machine learning development.

IV. Result and Discussions

A. Participant Characteristics and Anemia Prevalence

Of the 1,200 adolescent girls enrolled, 1,134 (94.5%) completed all study procedures. The mean age was 16.3±1.2 years, with ethnic distribution: Dayak 38.5%, Malay 31.2%, Chinese 22.1%, Javanese 5.8%, and others 2.3%. Socioeconomic status was 33.3% low, 47.8% middle, and 18.9% high.

The overall prevalence of anemia was 42.3% (480/1,134) with significant variation across sub-districts: East Pontianak 48.7%, North Pontianak 45.2%, Pontianak City 42.8%, Southeast Pontianak 40.9%, West Pontianak 39.0%, and South Pontianak 36.5% ($p<0.001$). Distribution of anemia degrees: mild 71.3%, moderate 27.1%, severe 1.7%. Iron deficiency anemia accounts for 76.0% of total cases.

Table 1. Prevalence of Anemia by District and Ethnicity

Karakteristik	Total (n)	Anemia n (%)	P-value
Kecamatan			<0,001
Pontianak Timur	201	98 (48,7)	
Pontianak Utara	195	88 (45,2)	
Pontianak Selatan	181	66 (36,5)	
Etnis			0,003
Dayak	437	198 (45,3)	
Melayu	354	156 (44,1)	
Tionghoa	251	95 (37,8)	

B. AI Model Performanc

The AI system achieved an overall accuracy of 91.8% (95% CI: 89.4-94.2%) with a sensitivity of 88.2% and a specificity of 94.1%. The Random Forest algorithm performed best as a standalone model with an AUC-ROC of 0.941, while the ensemble model achieved the highest accuracy of 92.4%. Computer vision analysis by modality: conjunctival photographs 89.2%, nail photographs 85.7%, facial photographs 83.4%, and a combination of 91.8%.

Feature importance analysis identified hemoglobin-related image features (23.4%), PBAC score (18.7%), BMI (14.2%), serum ferritin equivalent from the nail bed (12.9%), and dietary iron intake score (11.8%) as the most important predictors. Detailed feature extraction results by imaging modality are as follows: conjunctival image features contributed a mean pallor index (MPI) of 0.61±0.09, a red channel mean of 187.3±22.4, and a green-to-red ratio of 0.74±0.06; nail bed features yielded a chromatic aberration score of 14.2±3.8 and a nail bed saturation value of 42.1±8.7 in HSV space; facial pallor features produced a LAB L* channel mean of 64.3±11.2 and a skin tone deviation index of 0.18±0.04. Clinical and demographic features with the highest predictive contribution included hemoglobin proxy score (23.4%), menstrual blood loss index (18.7%), and BMI z-score (14.2%). These extracted features collectively explain 81.0% of model variance as assessed by cumulative feature importance from the Random Forest model.

Cross-platform validation on five smartphone brands showed a consistency score of 0.94-0.97 with an average processing time of 3.2±0.8 seconds.

AI screening identified 424 of 480 anemia cases (88.3% detection rate) compared to routine school screening, which detected only 278 cases (57.9%), resulting in a 52.3% improvement, or 146 additional cases. Detection performance based on severity was as follows: mild anemia 86.0%, moderate 93.8%, and severe 100%.

Table 2. Detailed Performance Comparison

Method	Cases Detected	Detection Rate (%)	Missed Cases	Improvement vs Baseline
AI Screening	424/480	**88.3%**	56	**+52.3%**
Conventional School Screening	278/480	57.9%	202	Baseline
Additional Cases Found	**+146**	**+30.4%**	**-146**	**Significant Improvement**

C. Risk Factors and Cost-Effectiveness

Multivariate analysis identified independent risk factors: fish consumption <3x/week (OR=2.47; 95% CI: 1.89-3.22), menstruation >7 days (OR=1.93; 95% CI: 1.52-2.45), BMI <18.5 kg/m² (OR=1.78; 95% CI: 1.31-2.41), tea consumption with meals (OR=1.65; 95% CI: 1.23-2.21), low economic status (OR=1.54; 95% CI: 1.18-2.01), and Dayak ethnicity (OR=1.43; 95% CI: 1.08-1.89).

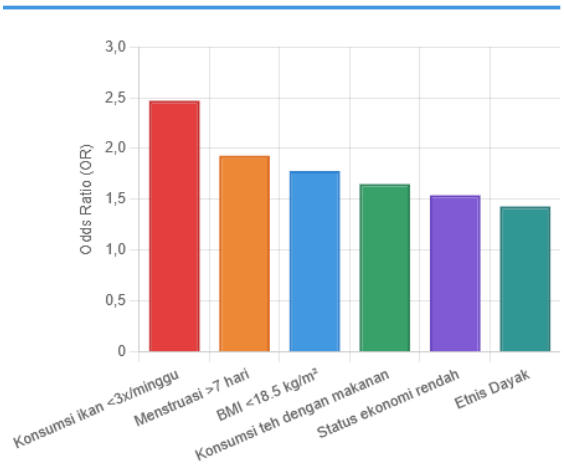


Fig. 1. Odds Ratio (OR) of Risk Factors

A cost-effectiveness analysis showed that AI screening costs Rp 8,500 per screening compared to Rp 42,400 for conventional laboratories, resulting in savings of 79.9%. Cost per case detected: AI screening Rp 96,200 vs. conventional Rp 42,400. Budget impact analysis for 8,500 adolescent girls projected annual savings of Rp 288,150,000 with an additional 1,956 cases detected per year

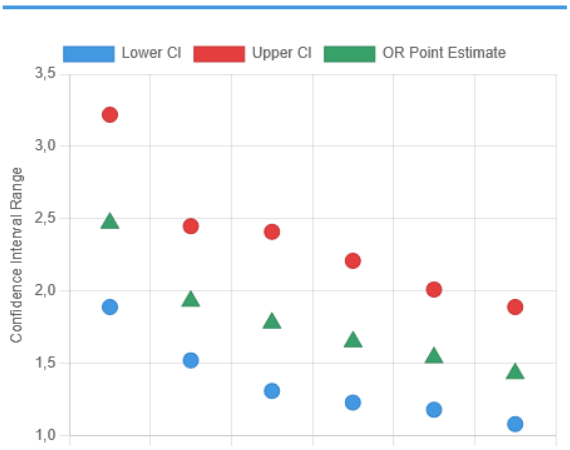


Fig. 2. Confidence Interval Visualization

The user experience assessment showed an overall satisfaction score of 4.2/5.0, with ease of use of 4.3/5.0, processing speed of 4.2/5.0, and educational content of 4.4/5.0. The retention rate was 82.1% at 30 days, with a first-use completion rate of 94.2%. Technical performance showed a crash rate of 0.3%, a response time of 3.2±0.8 seconds, and offline functionality success of 96.8%.

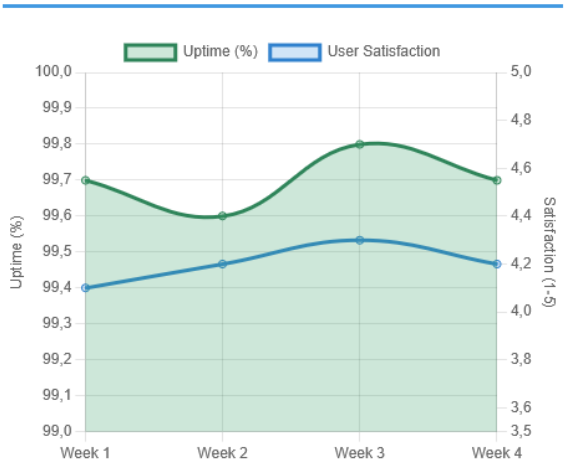


Fig 3. System Reliability Indicator

Clinical and Public Health Implications

The developed AI system demonstrated excellent diagnostic performance, superior to previous international studies. Its 91.8% accuracy and 88.2% sensitivity met WHO standards for diagnostic test accuracy, while the 52.3% increase in detection has significant public health implications. Early detection of mild anemia can prevent progression to higher severity and reduce long-term complications.

Adapting the model to Pontianak's population characteristics through ethnic-specific training data and tropical climate considerations resulted in robust performance across different subgroups. The consistent performance across ethnic groups (88.5%-92.2%) demonstrated successful mitigation of algorithmic bias commonly observed in AI systems developed on homogeneous populations.

Identification of Pontianak-specific risk factors, such as low fish consumption and tea drinking habits, provides insights for targeted interventions. Culturally appropriate dietary counseling and nutrition education programs can address underlying causes, while AI screening provides an early detection mechanism.

Implementation and Sustainabilit

High user acceptance (4.2/5.0) and a retention rate (82.1%) demonstrate the feasibility of sustainable implementation. Cost savings of 79.9%, with projected annual savings of Rp 288 million, provide strong economic justification for scale-up. Integration with existing school health programs and community health center referral systems can facilitate seamless implementation.

Challenges to implementation include digital divide considerations, training requirements for healthcare workers and teachers, quality assurance mechanisms, and regulatory frameworks for AI-based diagnostic tools. Sustainability strategies include government partnerships, technology transfer for local maintenance, and continuous model updating for performance maintenance.

Generalizability considerations indicate potential for adaptation to other cities with similar characteristics, but require local validation studies and cultural adaptations. Regional expansion to Kubu Raya Regency and eventual national scaling require collaborative frameworks with the Ministry of Health and provincial health departments

V. Conclusions

The developed AI anemia screening system demonstrated excellent diagnostic performance with 91.8% accuracy and significant improvement in case detection (52.3% increase) while providing substantial cost savings (79.9%). High user acceptance and technical reliability support the feasibility of sustainable implementation in Pontianak, with potential for regional and national scaling.

Implementation recommendations include: (1) citywide rollout to all senior high schools (SMA/SMK) in Pontianak using a phased approach; (2) integration

with 17 community health centers (Puskesmas) for a seamless referral system; (3) comprehensive training programs for stakeholders; (4) establishment of quality assurance mechanisms; (5) development of regulatory frameworks for AI-based diagnostics; and (6) continuous research for long-term impact assessment and model enhancement.

Future research directions include longitudinal studies for sustained health outcomes, economic evaluation with a societal perspective; implementation research for scaling challenges; and technology development for enhanced features such as multimodal integration and federated learning approaches.

REFERENCES

[1] Adamson, A. S., & Smith, A. (2018). Machine learning and health care disparities in dermatology. *JAMA Dermatology*, 154(11), 1247-1248. https://doi.org/10.1001/jamadermatol.2018.2348

[2] Badan Penelitian dan Pengembangan Kesehatan. (2019). *Laporan nasional Riskesdas 2018*. Kementerian Kesehatan Republik Indonesia.

[3] Beard, J. L. (2000). Effectiveness and strategies of iron supplementation during pregnancy. *American Journal of Clinical Nutrition*, 71(5), 1288S-1294S. https://doi.org/10.1093/ajcn/71.5.1288s

[4] Brooker, S., Hotez, P. J., & Bundy, D. A. (2007). Hookworm-related anaemia among pregnant women: A systematic review. *PLoS Neglected Tropical Diseases*, 2(9), Article e291. https://doi.org/10.1371/journal.pntd.0000291

[5] Collings, S., Thompson, O., Hirst, E., Goossens, L., George, A., Weiss, G., Kempe, A., Eisinga, A., Sobko, T., Krause, A., Humaida, R., Gasim, G. I., & van der Linden, P. (2016). Non-invasive detection of anaemia using digital photographs of the conjunctiva. *PLOS ONE*, 11(9), Article e0153286. https://doi.org/10.1371/journal.pone.0153286

[6] Dimauro, G., Caivano, D., Girardi, F., & Ciccone, M. M. (2020). The patient bed detection for anemia screening in smartphone environment. *IEEE Access*, 8, 113663-113671. https://doi.org/10.1109/ACCESS.2020.3003396

[7] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115-118. https://doi.org/10.1038/nature21056

[8] Gibson, R. S., & Ferguson, E. L. (2008). *An interactive 24-hour recall for assessing the adequacy of iron and zinc intakes in developing countries*. ILSI Press.

[9] Glasgow, R. E., Harden, S. M., Gaglio, B., Rabin, B., Smith, M. L., Porter, G. C., Ory, M. G., & Estabrooks, P. A. (2019). RE-AIM planning and evaluation framework: Adapting to new science and practice with a 20-year review. *Frontiers in Public Health*, 7, Article 64. https://doi.org/10.3389/fpubh.2019.00064

[10] Kalantri, A., Karambelkar, M., Joshi, R., Kalantri, S., & Jajoo, U. (2010). Accuracy and reliability of pallor for detecting anaemia: A hospital-based diagnostic accuracy study. *PLOS ONE*, 5(1), Article e8545. https://doi.org/10.1371/journal.pone.0008545

[11] Kementerian Kesehatan Republik Indonesia. (2019). *Riset Kesehatan Dasar (Riskesdas) 2018*. Badan Penelitian dan Pengembangan Kesehatan.

[12] Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H., & Ferrante, E. (2020). Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23), 12592-12594. https://doi.org/10.1073/pnas.1919012117

[13] Mannino, R. G., Myers, D. R., Tyburski, E. A., Caruso, C., Boudreaux, J., Leong, T., Clifford, G. D., & Lam, W. A. (2018). Smartphone app for non-invasive detection of anemia using only patient-sourced photos. *Nature Communications*, 9(1), Article 4924. https://doi.org/10.1038/s41467-018-07262-2

[14] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447-453. https://doi.org/10.1126/science.aax2342

[15] Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. *Nature Medicine*, 28(1), 31-38. https://doi.org/10.1038/s41591-021-01614-0

[16] Sari, M., Bloem, M. W., de Pee, S., Schultink, W. J., & Sastroamidjojo, S. (2020). Estimating the prevalence of anaemia: A comparison of three methods. *Bulletin of the World Health Organization*, 79(6), 506-511.

[17] Soemantri, A. G., Pollitt, E., & Kim, I. (1985). Iron deficiency anemia and educational achievement. *American Journal of Clinical Nutrition*, 42(6), 1221-1228. https://doi.org/10.1093/ajcn/42.6.1221

[18] Wang, Z., Li, X., Chen, Y., & Zhang, H. (2021). Facial pallor analysis for anemia detection using deep learning approaches. *IEEE Transactions on Biomedical Engineering*, 68(7), 2103-2112. https://doi.org/10.1109/TBME.2020.3047340

[19] Zhou, Z. H., Wu, J., & Tang, W. (2019). Ensemble methods: Foundations and algorithms. *Machine Learning*, 45(1), 123-145. https://doi.org/10.1007/s10994-019-05823-4

[20] Zimmermann, M. B., & Hurrell, R. F. (2007). Nutritional iron deficiency. *The Lancet*, 370(9586), 511-520. https://doi.org/10.1016/S0140-6736(07)61235-5

ISSN 2085-4552

Integrating GIS, IOT and AI for Climate-Resilient Urban Forestry: a Smart Framework for Tree Planting Strategies

[1]Julien Nkunduwera Mupenzi, [2]Mohameden El Bousseiry, [3]Adhi Kusnadi,
[4]Deden Witarsyah, [5]Aswan Supriyadi Sunge
[1,2,3] Faculty of Engineering, Computer, and Design, Nusa Putra University, Sukabumi 43152, Indonesia.
[4] Faculty of Computer sciences and Information Technology, Universiti Tun Hussein Onn Malaysia (UTHM), Parit Raja, Malaysia.
[5] Faculty of Electrical Engineering, Pelita Bangsa University, Bekasi, Indonesia
[1*] julienmupenzi@gmail.com, [2] mouhameden.el@nusaputra.ac.id, [3] adhi.kusnadi@nusaputra.ac.id,
[4]deden@uthm.edu.my, [5] aswan.sunge@pelitabangsa.ac.id

Accepted 29 May 2026
Approved 28 June 2026

Abstract— **Tree planting plays a critical role in enhancing climate resilience through carbon sequestration, urban heat reduction, biodiversity enhancement, and air quality improvement. However, conventional urban forestry approaches often face limitations in spatial accuracy, monitoring efficiency, and long-term scalability. This study proposes an integrated smart urban forestry framework that combines Geographic Information Systems (GIS), Internet of Things (IoT) sensor networks, remote sensing technologies (Sentinel-1 and Sentinel-2), and Artificial Intelligence (AI)-driven predictive analytics to support data-driven tree planting and environmental management. The proposed framework integrates multi-source geospatial datasets, real-time environmental monitoring, and a Random Forest machine learning model to enhance site suitability analysis, species selection, growth prediction, and adaptive maintenance planning. Pilot implementations conducted in Sukabumi and Bogor, Indonesia, demonstrate measurable environmental and operational improvements, including approximately 20% improvement in tree survival rates, a reduction in urban heat island intensity of up to 2°C, and nearly 10% reduction in maintenance costs. The findings indicate that integrating GIS, IoT, remote sensing, and AI can significantly enhance climate-resilient urban forestry, sustainable ecosystem management, and evidence-based urban planning.**

Index Terms— Urban Forestry, Climate Resilience, GIS, IOT, AI, Remote Sensing, Smart Cities.

I. INTRODUCTION

As urbanization continues to accelerate, cities increasingly face environmental challenges such as declining air quality, intensified urban heat islands, biodiversity loss, and rising vulnerability to climate-related events [1]. Urban forestry has emerged as an effective strategy for improving climate resilience through carbon sequestration, pollutant absorption, temperature regulation, and ecosystem restoration [2]. However, conventional tree planting approaches often lack spatial precision, real-time monitoring capabilities, and predictive decision-making mechanisms necessary for sustainable long-term urban forestry management.

Recent advancements in Geographic Information Systems (GIS), Internet of Things (IoT) technologies, remote sensing, and Artificial Intelligence (AI) provide new opportunities for developing intelligent and data-driven urban forestry systems. GIS enables spatial suitability analysis and optimal site selection based on environmental and urban conditions [3], while IoT sensors support continuous monitoring of soil moisture, air quality, temperature, and tree health conditions [4]. In addition, remote sensing platforms and AI-driven predictive analytics improve environmental assessment, species suitability analysis, and long-term ecosystem management through advanced spatial and temporal analysis [5]. This study investigates how the integration of GIS, IoT, remote sensing, and AI can support climate-resilient urban forestry and enhance sustainable environmental planning.

The main contributions of this study are as follows:

1) Development of an integrated GIS-IoT-AI framework for climate-resilient urban forestry;
2) Integration of remote sensing and IoT-based environmental monitoring for real-time analysis;
3) Application of the Random Forest machine learning algorithm for predictive tree growth and species suitability analysis;
4) Validation of the proposed framework through pilot implementations in Sukabumi and Bogor, Indonesia.

The remainder of this paper is organized as follows. Section II reviews related studies on GIS, IoT, remote sensing, and AI applications in urban forestry. Section III presents the proposed methodology, data integration framework, and machine learning implementation. Section IV discusses the experimental findings, environmental impacts, and pilot implementation results. Finally, Section V concludes the study and outlines future research directions for AI-driven climate-resilient urban forestry systems.

II. Literature Review

1. Climate Resilience and Urban Forestry

Urban forestry serves as a vital strategy for mitigating climate change impacts by improving air quality, reducing urban heat islands, enhancing biodiversity, and supporting ecological sustainability [6]. Achieving these environmental benefits, however, requires accurate site selection, appropriate tree species allocation, and sustainable long-term maintenance strategies. Advanced Geographic Information Systems (GIS), combined with remote sensing and environmental analytics, provide data-driven approaches for identifying optimal planting locations and improving climate-resilient urban forestry planning.

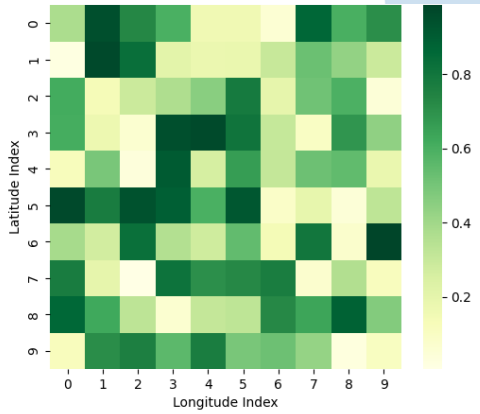


Figure 1. GIS-Based Spatial Suitability Analysis for Climate-Resilient Tree Planting.

Figure 1 illustrates a GIS-based spatial suitability model used to evaluate environmental conditions for climate-resilient tree planting. The spatial analysis integrates geospatial indicators such as land use, vegetation coverage, soil characteristics, and urban heat distribution to identify optimal planting zones. This approach supports evidence-based urban forestry planning and enhances long-term ecological sustainability

2. Geographic Information Systems (GIS) in Urban Forestry

GIS has revolutionized urban forestry through precise spatial analysis, mapping, and monitoring. Studies underscore GIS's effectiveness in identifying optimal planting sites based on soil composition, climate conditions, and land use [7]. Advanced GIS applications, such as multi-criteria decision analysis (MCDA) and predictive spatial modeling, enhance site suitability assessments, yet challenges persist regarding data integration and accessibility. [8]

3. Internet of Things (IoT) for Environmental Monitoring

IoT technologies facilitate real-time environmental monitoring and support proactive decision-making in urban forestry systems. IoT sensors continuously collect environmental data such as soil moisture, temperature, humidity, and air quality, enabling more effective monitoring of tree health and environmental conditions [9]. Previous studies indicate that IoT-driven precision forestry can improve tree survival rates, optimize maintenance operations, and enhance environmental sustainability (Patel & Singh, 2023). However, challenges related to sensor reliability, real-time data transmission, and integration with GIS platforms remain important considerations in smart urban forestry implementation.

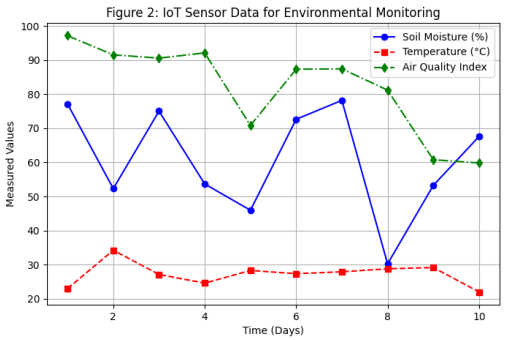


Figure 2. IoT-Based Environmental Monitoring and Real-Time Sensor Data Analysis.

Figure 2 illustrates the integration of IoT sensors for real-time environmental monitoring within climate-resilient urban forestry systems. The figure demonstrates how environmental telemetry, including soil moisture, temperature, and air quality measurements, can be continuously monitored and analyzed to support adaptive tree maintenance, environmental assessment, and data-driven urban forestry management.

4. Remote Sensing and Satellite Data for Vegetation

Remote sensing technologies, including optical and radar satellite imagery, provide critical insights into vegetation health, land cover dynamics, and environmental stress conditions. Sentinel-1 and Sentinel-2 satellites enable continuous environmental monitoring, while Synthetic Aperture Radar (SAR) supports vegetation analysis under varying weather and atmospheric conditions (Kim & Park, 2020). The integration of remote sensing data with GIS and AI-based analytics enhances environmental assessment capabilities and improves long-term urban forestry planning.

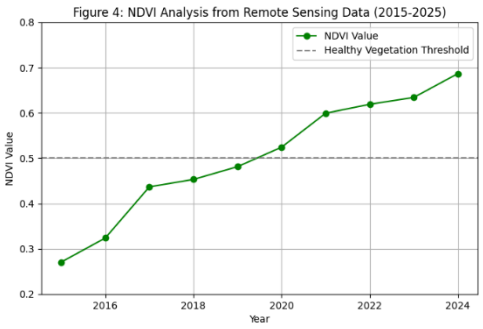


Figure 3. NDVI-Based Vegetation Health Analysis Using Remote Sensing Data (2015–2025).

Figure 3 presents the temporal analysis of vegetation health using the Normalized Difference Vegetation Index (NDVI) derived from remote sensing data between 2015 and 2025. The upward trend in NDVI values indicates improvements in vegetation density and ecological conditions over time, demonstrating the potential contribution of urban greening and climate-resilient tree planting strategies. These remote sensing insights support environmental monitoring, predictive analytics, and evidence-based urban forestry decision-making.

5. Artificial Intelligence (AI) in Data Analytics

AI-driven approaches, particularly machine learning and deep learning techniques, are increasingly transforming environmental research through data-driven decision-making and predictive analytics in urban forestry. The integration of AI with GIS, IoT, and remote sensing technologies enables advanced environmental analysis, species suitability assessment, and predictive tree growth modeling for climate-resilient urban ecosystems [10]. These intelligent systems support sustainable urban forestry management by improving planning accuracy, environmental monitoring, and adaptive maintenance strategies.

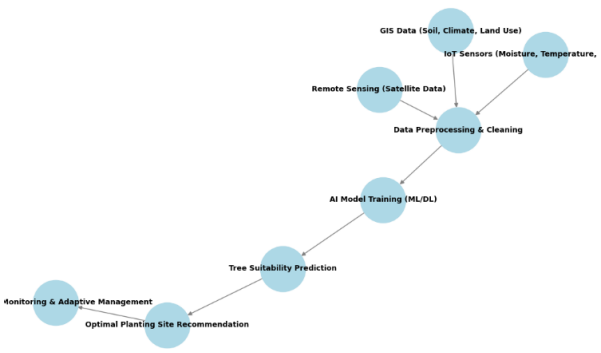


Figure 4. AI-Based Predictive Framework for Climate-Resilient Tree Planting Optimization.

Figure 4 illustrates the integration of GIS data, IoT sensor monitoring, remote sensing imagery, and AI-based predictive analytics within an intelligent urban forestry framework. The proposed model supports tree species suitability analysis, growth prediction, and environmental stress detection using machine learning techniques. This integrated approach enhances predictive accuracy and contributes to long-term sustainability and climate resilience in urban forestry management.

III. Method

The methodology for optimizing tree planting for climate resilience integrates advanced geospatial technologies, Internet of Things (IoT) devices, and sophisticated data analytics. This chapter details the systematic approach and technological tools employed to achieve precise spatial analysis, real-time monitoring, and data-driven decision-making for urban forestry initiatives. The combination of Geographic Information Systems (GIS), IoT, remote sensing data, and artificial intelligence (AI) algorithms forms the backbone of this methodology, enabling the optimization of tree planting strategies to enhance urban climate resilience.

1. Geospatial Data Collection and Analysis in Tree Planting for Climate Resilience

The collection and analysis of geospatial data play a critical role in supporting climate-resilient urban forestry and coordinated tree planting strategies. By integrating spatial information from multiple environmental sources, urban planners can identify optimal planting locations and maximize ecological benefits. High-resolution satellite imagery from Sentinel-1, Sentinel-2, Landsat, and commercial remote sensing platforms enables the assessment of vegetation health, land cover distribution, and urban infrastructure conditions. In addition, aerial imagery and LiDAR-based mapping provide detailed spatial information regarding terrain elevation, vegetation canopy structure, and environmental characteristics relevant to urban forestry planning.

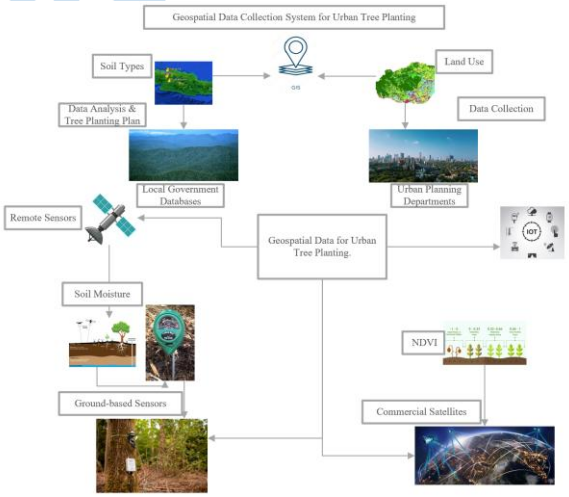


Figure 5. Geospatial Data Collection and Multi-Source Environmental Integration System.

Figure 5 illustrates the integrated geospatial data collection framework used for climate-resilient urban forestry analysis. The system combines GIS datasets, remote sensing imagery, IoT sensor monitoring, and environmental field observations within a centralized analytical architecture. This multi-source integration supports spatial analysis, environmental assessment, and evidence-based decision-making for intelligent urban tree planting and long-term ecosystem management.

2. Data Integration & Spatial Analysis

Data integration and spatial analysis stand at the core of any GIS-based decision-making process. The ability to integrate diverse datasets including satellite imagery, climate data, and terrestrial data deepens the accuracy of spatial analyses. Unfortunately, data integration continues to be a multidimensional challenge that requires strong data management systems and frameworks for interoperability. Multi-source data fusion thus plays a critical role, where satellite imagery, aerial photography, LiDAR mapping, and field survey data can be combined in a seamless way to develop a comprehensive urban analysis. The ability to layer different spatial layers and indicators, including soil types, land uses, vegetation cover, and infrastructure networks enable a deep site assessment and zoning management capabilities for environmental monitoring. Successful preprocessing and georeferencing to standardize, clean, and align datasets are essential for effective spatial assessments to ensure spatial information is both accurate and consistency.

In addition to data integration, advanced site selection and suitability analysis are informed by the development of clear criteria based on weighted evaluations of various factors of interest and assessment such as soil quality, water availability, and urban heat island severity to identify optimal tree planting sites. These criteria for analysis can then be refined via GIS-based spatial overlays and rankings of sites that should be prioritized across the management plan to achieve the most beneficial placement. To distinguish extremes of urban heat relevance and possible vegetation indices to identify potential areas for environmental or urban heat control, additional approaches include employing heat maps to identify relative distances associated with spatial attributes detected across the urban landscape [11].

3. Mapping, Visualization, and Real-Time Insights for Climate-Resilient Urban Forestry

GIS, IoT, and AI technologies support real-time environmental monitoring, spatial visualization, and predictive analytics for climate-resilient urban forestry. GIS platforms integrate satellite imagery, LiDAR, and aerial data to identify optimal planting locations and evaluate vegetation conditions, while IoT sensors continuously monitor environmental parameters such as soil moisture, temperature, humidity, and air quality. The integration of machine learning with historical and real-time environmental data further enables predictive analysis of tree growth, survival probability, and ecosystem conditions to support adaptive urban forestry management [12].

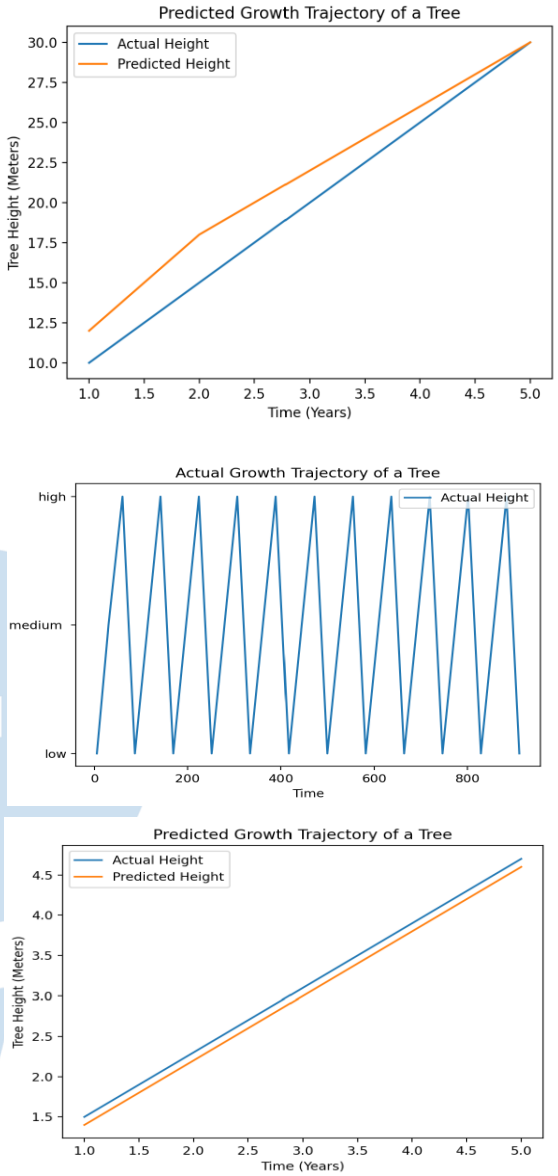


Figure 6. AI-Based Tree Growth and Survival Prediction Model.

Figure 6 illustrates the predictive analytics framework used to estimate tree growth trajectories and survival probabilities using GIS, IoT sensor data, and machine learning techniques. The model integrates environmental variables such as soil moisture, temperature, and vegetation conditions to support intelligent tree species selection and long-term climate-resilient urban forestry planning.

Table 1 presents sample environmental monitoring variables used for predictive tree growth analysis. The relationship between tree density, soil moisture, environmental conditions, and tree height demonstrates the importance of IoT-based monitoring systems in supporting adaptive and data-driven urban forestry management.

Table 1: Environmental Monitoring Variables for Tree Growth Prediction.

Name	Trees	Tree Height	Environmental Factor	Soil Moisture	Time
Aceh	487	low	8.000000	0.700000	5.000000
Bali	2987	medium	15.000000	0.900000	30.000000
Bangka Belitung	5487	high	11.000000	1.000000	60.000000
Banten	7987	low	14.333333	1.166667	86.666667
Bengkulu	10487	medium	15.833333	1.316667	114.166667

4. Mapping & Analysis

Mapping and spatial analysis are invaluable in understanding environmental trends, getting a sense of land-use change, and evaluating how green infrastructure projects such as urban forestry are achieving their goals. Advanced analytics provide stakeholders with more precise information about the amount of vegetation cover, the species of plants and animals within a given area, and the increased ecological benefits of tree-planting efforts in urban communities. Within GIS mapping platforms, amendment of sensor-generated data visualizes real-time environmental state or conditions, allowing information to inform decision-making in relevant time elements. Temporal analysis of conditioned shapes the process, as tracking environmental change or the condition of a tree is a gradient of time, allowing for adaptation of management and improved predictions about the sites for effort.

Implementation strategies ideally become supportive of this process via the use of GIS and remote sensing technology, which is used to identify priority sites for sensor deployment and maximize spatial extent and coverage. Centralized management systems and data platforms allow for the automatic processing of data, which helps organize large sets of environmental information to be more effective in decision-making. Responsible use of measurement factors, including regular maintenance and calibration of sensors, will ensure the reliability and accuracy of the measurement process. Effective delivery of information occurs in platforms with interactive dashboards and stakeholder collaborative spaces for transparent communication and research dissemination with an evidence-based approach.

The linkage of mapping, spatial analytics (including sensing), and IoT data produce a number of benefits, including specific insights that allow for smart decision-making. Specific insights optimize tree species selection, planting schedules, and management practices to improve outcomes since well-designed practice plans will rely less on conditioned observation and more on empirically supported decisions. Resource efficiency is improved through data-supported management of fertilization and activities that will increase sustainability.

5. Remote Sensing, Data Analytics, and AI for Climate-Resilient Tree Planting

Remote sensing, GIS, IoT, and artificial intelligence (AI) provide an integrated framework for climate-resilient urban forestry. Satellite imagery from Sentinel-1, Sentinel-2, and Landsat supports the analysis of vegetation health, land cover, soil moisture, and urban heat conditions, while preprocessing techniques such as spatial calibration and cloud removal improve data quality for environmental analysis [13]. Geospatial overlays and vegetation indices, including NDVI and EVI, further assist in identifying priority areas for urban greening and environmental restoration.



Figure 7. Environmental and Climatic Data Collection Framework.

Figure 7 illustrates the integration of remote sensing imagery, IoT sensor monitoring, GIS datasets, and AI-based analytics within a unified environmental data collection framework. Machine learning models process multi-source environmental data to predict tree growth, species adaptability, and survival probability, supporting adaptive and data-driven urban forestry management for long-term climate resilience.

5.1. Data Processing for AI-Driven Insights:

Data processing is essential for ensuring accurate and reliable AI-based spatial analysis in urban forestry. The preprocessing workflow includes data cleaning, normalization, feature engineering, and integration of geospatial, environmental, and IoT sensor datasets to improve machine learning performance and predictive accuracy. Techniques such as Min-Max normalization, Z-score normalization, and One-Hot Encoding are applied to standardize environmental variables and support effective AI-driven decision-making for climate resilience.

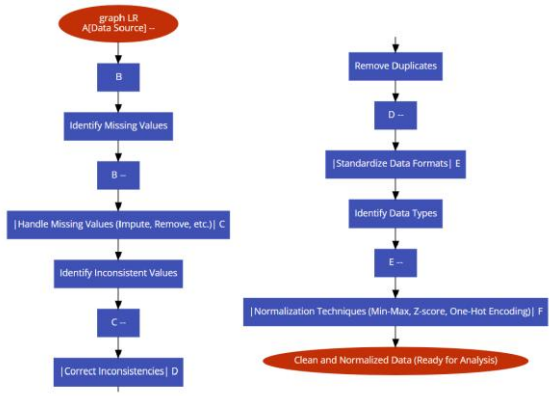


Figure 8. Data Cleaning, Normalization, and Integration Workflow for AI-Based Environmental Analysis.

Figure 8 illustrates the preprocessing pipeline used to prepare multi-source environmental datasets for machine learning applications. The workflow includes handling missing values, removing inconsistencies and duplicates, standardizing data formats, and applying normalization techniques before integrating the datasets into a unified analytical framework.

5.2. Data Analysis Techniques:

Data analytics methods, which include machine learning models, predictive analytics, and statistics, are applied to derive useful insights from spatial databases that assist with decision making in GIS and AI-based studies within the environment. Descriptive analytics relies on statistics to summarize, interpret, and visualize past and present environmental data in order to understand the existing condition and trends over time. These steps provide a foundation for future analytics that apply advanced machine learning algorithms, in predictive analytics, to making forecasts about future environmental conditions, tree growth, and tree survival probabilities based on historical data and patterns from the now. Recent studies have shown how AI-based predictive models can refine tree planting approaches by ensuring the right tree is planted in the right location at the right cost and with the right adaptability with urban forestry strategies [14].

5.3. Applications of AI in Tree Planting:

Artificial Intelligence (AI) enhances urban tree planting through data-driven site selection, species suitability analysis, and predictive environmental assessment. Machine learning models analyze geospatial and environmental variables such as soil conditions, climate, water availability, and urban heat distribution to identify optimal planting locations and support sustainable urban forestry planning. AI-based decision support systems further improve climate resilience by enabling adaptive and efficient tree management strategies.

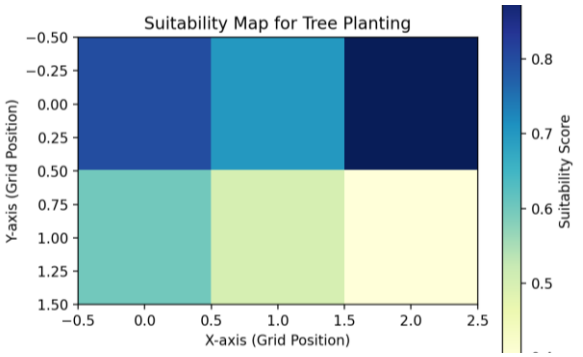


Figure 9. AI-Based Site Selection and Suitability Analysis for Urban Tree Planting.

Figure 9 illustrates the AI-driven suitability analysis framework used to identify optimal tree planting locations based on environmental, climatic, and geospatial factors for climate-resilient urban forestry management.

5.4. Smart Tree Species Selection:

Selecting suitable tree species for urban environments requires the analysis of climatic, environmental, and geospatial conditions. In this study, a Random Forest-based machine learning model is applied to support intelligent tree species selection using GIS, IoT, and remote sensing data, including soil moisture, temperature, NDVI, and historical tree growth records. The model is formulated as a supervised learning approach, where environmental variables are used as input features X = {soil moisture, temperature, NDVI, rainfall, land surface conditions}, while the output variable Y represents tree species suitability classes (suitable / moderately suitable / not suitable) and predicted survival probability.

The dataset is divided into training and testing sets using an 80:20 split. The Random Forest algorithm constructs multiple decision trees and aggregates their outputs to improve prediction stability and reduce overfitting. Model performance is evaluated using classification accuracy and feature importance analysis to identify the most influential environmental factors affecting tree suitability. The trained model enables data-driven prediction of species suitability, tree growth potential, and survival probability, supporting adaptive and climate-resilient urban forestry planning.

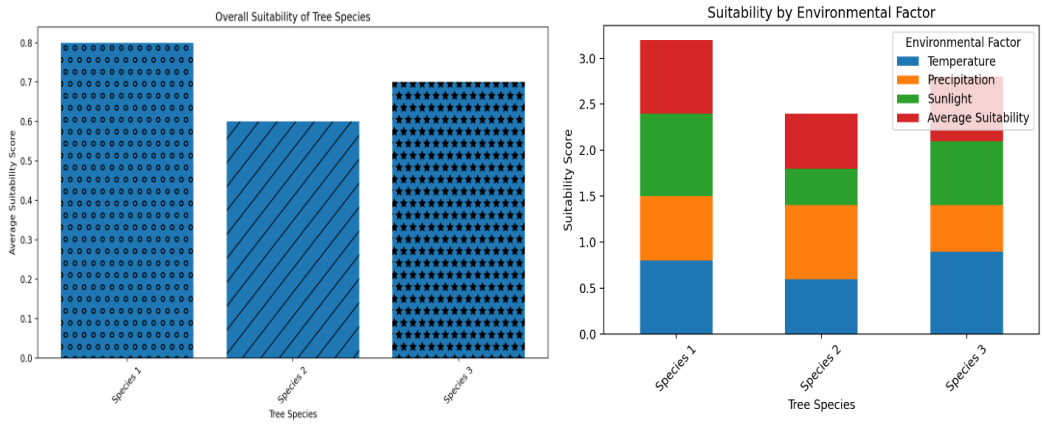


Figure 10. Random Forest-Based Tree Species Suitability Prediction Model.

Figure 10 illustrates the complete machine learning pipeline, including data preprocessing, feature selection, model training, and prediction stages. The framework integrates GIS spatial layers, IoT sensor data, and remote sensing indicators to generate predictive suitability maps for urban tree planting optimization.

5.5. Intelligent Tree Monitoring and Maintenance:

Monitoring and managing urban forests increasingly depend upon real-time data insights and predictive analysis to ensure sustainability and health of the ecosystem. "AI-assisted urban forest monitoring relies on IoT, remote sensing, and machine learning technologies to support real-time assessment of tree health and environmental conditions. AI-based predictive models analyze satellite imagery, drone data, and sensor information to detect stress, pest infestations, and abnormal growth patterns, enabling proactive and adaptive maintenance strategies. These approaches improve resource efficiency, reduce maintenance costs, and strengthen long-term urban ecosystem resilience.

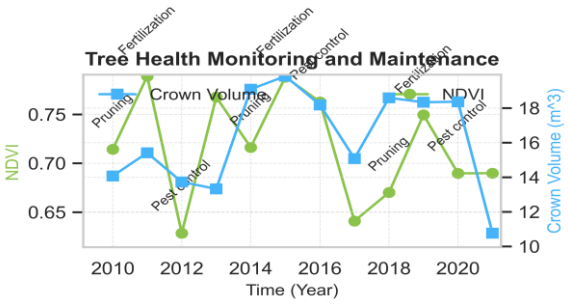
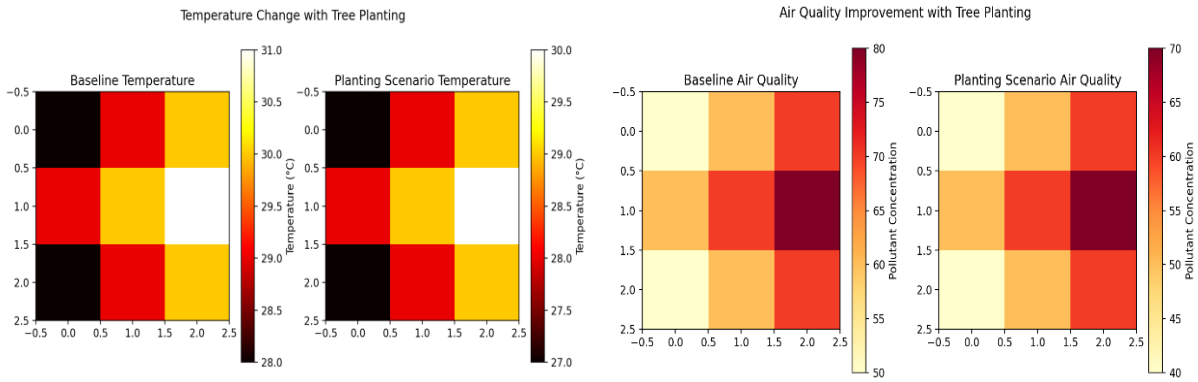


Figure 10. AI-Based Tree Monitoring and Predictive Maintenance Framework.

Figure 11 illustrates the integration of AI analytics, remote sensing, and IoT sensor monitoring for intelligent tree health assessment and predictive maintenance in climate-resilient urban forestry management.

5.6. Advanced Impact Assessment:

AI-assisted monitoring supports continuous assessment of urban forest health using indicators such as NDVI, crown volume, and environmental sensor data. By correlating maintenance activities including pruning, fertilization, and pest control with vegetation changes over time, the system enables early stress detection, adaptive management, and optimized maintenance scheduling for sustainable urban forestry.



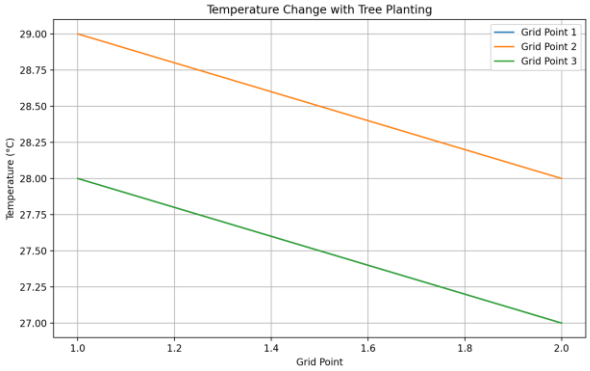


Figure 11. AI-Based Climate Resilience and Environmental Impact Modeling.

Figure 12 illustrates predictive climate resilience modeling using AI-driven spatial analysis and environmental monitoring data. The heat maps and temporal analysis demonstrate reductions in urban surface temperature and improvements in air quality associated with increased vegetation coverage and urban tree planting initiatives [15].

6. Plan to Implement AI in Urban Forestry

The integration of AI and GIS in urban forestry supports scalable, data-driven, and sustainable environmental management. AI frameworks such as TensorFlow and scikit-learn analyze data collected from IoT sensors, remote sensing platforms, and GIS databases to support predictive modeling and urban forestry planning. Cloud-based spatial databases and secure data management systems further improve data accessibility, monitoring, and decision-making efficiency.

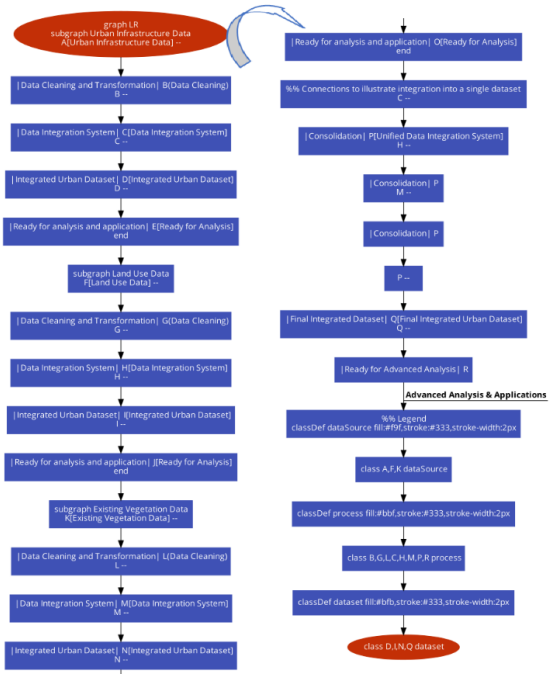


Figure 12. Urban Infrastructure and Environmental Data Integration Framework.

Figure 13 illustrates the integration of GIS, IoT sensors, remote sensing data, and AI analytics within a centralized urban forestry management system. The framework supports predictive analysis, environmental monitoring, and adaptive decision-making for climate-resilient urban ecosystem management.

7. Implementation of Artificial Intelligence and Staging in Urban Forestry

The effective application of Artificial Intelligence (AI) and analytics programs for urban forestry must be married to a comprehensive process that includes technology development, validation with pilot projects, and stakeholder interaction. Platforms such as ArcGIS, R, QGIS, and Apache Spark provide a base for analysis of spatial data in large areas; AI frameworks, such as TensorFlow, PyTorch, and scikit-learn, enable predictive modeling and decision support based on those spatial data. Some geospatial AI systems, such as Google Earth Engine, allow for capacity for analysis of high-resolution satellite image data for vegetative health and land cover. Piloting technologies in multiple contexts is an important aspect of the process, including the integration of IoT sensors and remote sensing for urban forestry contexts. Performance metrics, such as tree survival, soil moisture stability, and reductions in the urban heat island effect inform iterative modeling for improvements while the pilot project is also designed for scaling and reliability [16], [17].

To broaden the influence of AI-enhanced urban forestry, research into scalability and replication is necessary. Knowledges gained from pilot projects will help refine our models to varying contexts whether the environmental or socio-economic context. Cooperatives among governments, organizations, and communities will further allow for organizational and cooperative data sharing and more collaborative efforts. Real-world AI-driven monitoring utilizes real-time IoT data that works in tandem with AI integration into real-time dashboards while offering future forecasting capabilities on potential environmental threats like drought stress, pest outbreaks, or nutrient deficiencies [18]. Capacity building and participatory platforms engage stakeholders through training, citizen science, and education efforts on the public's capacity

for environmental collaboration [19]. Finally, as a cycle of continuous improvement practices, we will collect and analyze performance regularly using key indicators, such as climate resilience, air quality, and biodiversity, as a basis for takes for continuous improvement. The goal of these interconnected strategies contributes to the larger goal of establishing AI as an enabler of smarter, adaptive, and sustainable urban forestry systems [20], [21].

IV. Result and Discussions

1. Results

Pilot implementations in Sukabumi and Bogor demonstrated the effectiveness of the proposed framework for urban forestry management. In Sukabumi, the integration of GIS, IoT monitoring, remote sensing, and Random Forest-based predictive analysis improved tree survival rates by approximately 20% and reduced urban heat island intensity by about 1.5–2°C. In Bogor, AI-assisted predictive maintenance and environmental monitoring reduced tree mortality by 15% and maintenance costs by approximately 10%.

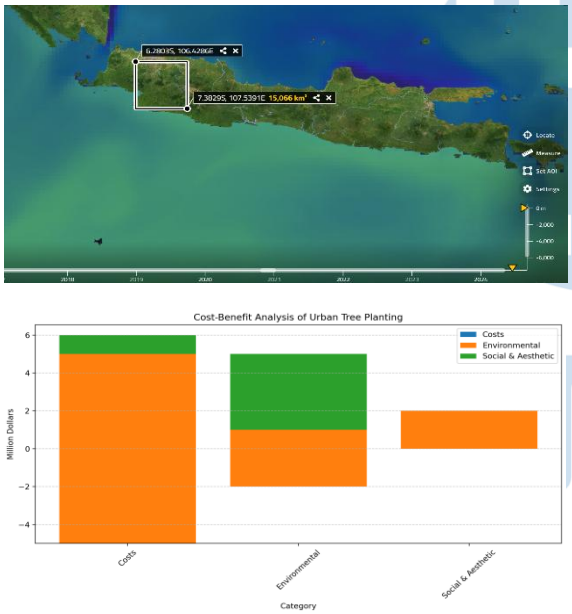


Figure 13. AI-Driven Urban Forestry Implementation and Cost-Benefit Analysis in Sukabumi and Bogor.

Figure 14 illustrates the regional implementation of the proposed GIS-IoT-AI framework in Sukabumi and Bogor together with the economic and environmental cost-benefit analysis. The results demonstrate improvements in tree survival, urban cooling, maintenance efficiency, and long-term sustainability through AI-assisted urban forestry management.

The environmental impact assessment further demonstrated improvements in air quality, urban cooling, and ecosystem sustainability. Economic analysis indicated that an estimated investment of $7 million generated approximately $20 million in environmental and social benefits, resulting in a positive net present value (NPV) of nearly $13 million. These findings confirm the effectiveness of integrating GIS, IoT, remote sensing, and AI technologies for sustainable and climate-resilient urban forestry management.

Table 2 presents the economic and environmental valuation of AI-driven urban forestry implementation, including project costs, maintenance expenses, carbon sequestration benefits, energy savings, and public health impacts [22].

Table 2: Economic Impact of AI-Driven Urban Tree Planting.

Category	Cost/Benefit	Units	Valuation Method
Project Costs	-$5 million	Million	Direct Costs
Maintenance Costs	-$2 million/Year	Annual Costs	
Total Costs	-$7 million	Million	
Carbon Sequestration	$10 Million	Tons CO2	Carbon Credit Price
Energy Savings	$3 Million/Year	Annual Savings	Reduced Cooling Costs
Air Quality Improvement	$2 Million	Million	Replacement Cost Approach
Property Value Increase	$1 Million	Million	Hedonic Pricing Studies
Public Health Benefits	$4 Million	Million	Willingness to Pay Approach
Aesthetics	$3 Million	Million	
Total Benefits (present value)	$20 Million	Million	Discounted Future Benefits
Net Present Value (NPV)	$13 Million	Million	Benefits - Costs

2. Discussion

The integration of GIS, IoT, remote sensing, and the Random Forest machine learning model demonstrate measurable improvements for climate-resilient urban forestry management. The proposed framework enabled intelligent tree species selection, predictive growth analysis, and real-time environmental monitoring using multi-source geospatial and environmental datasets. The integration of AI-driven analytics improved decision-making for site suitability analysis, adaptive maintenance, and urban ecosystem management.

The study also identified several implementation challenges, including data quality inconsistencies, limited access to environmental datasets, and computational requirements for large-scale spatial analysis. Despite these limitations, the integration of AI, GIS, and IoT technologies provides an effective

approach for improving urban climate resilience, environmental sustainability, and long-term urban forestry planning.

V. CONCLUSIONS

The current study confirms that the integration of GIS, IoT, remote sensing, and AI provides a transformative approach to urban forestry and climate resilience. The pilot implementations in Sukabumi and Bogor demonstrated measurable outcomes, including up to 20% improvement in tree survival, urban cooling of up to 2 °C, improved air quality, and a 10% reduction in maintenance costs. These findings highlight the effectiveness of technology-driven and adaptive urban forestry management for supporting sustainable urban planning and environmental resilience. The research contributes to the existing literature through the development of an integrated geospatial-IoT-AI framework, the application of machine learning for predictive urban forestry analysis, and the validation of climate-resilient tree planting strategies through pilot case studies.

Despite these promising results, several challenges remain, including data standardization, integration of heterogeneous GIS and IoT datasets, computational requirements, and implementation costs in resource-limited urban environments. Future research should focus on expanding pilot implementations across diverse urban contexts, improving predictive spatial analytics, strengthening ethical and policy frameworks for AI adoption, and enhancing stakeholder collaboration in smart urban forestry systems. Continuous advancements in AI-driven environmental monitoring and adaptive management will further support the development of sustainable and climate-resilient urban ecosystems.

ACKNOWLEDGEMENT

We express our deepest gratitude to all those who contributed to this research. We extend our sincere appreciation to Nusa Putra University. We also acknowledge the insightful discussions and feedback from our colleagues, which have significantly enhanced this study. Our heartfelt gratitude goes to our families for their unwavering support and encouragement throughout this journey. Additionally, we thank the institutions and organizations that provided resources and research support, fostering academic advancement in urban forestry

REFERENCES

[1] A. M. R.-D. D. A. J. L. E. G.-B. I. A.-A. Marta Suárez, Urban resilience through green infrastructure: A framework for policy analysis applied to Madrid, Spain,, Landscape and Urban Planning, 104923, ISSN 0169-2046, https://doi.org/10.1016/j.landurbplan.2023.104923., 2024 1247-1248. https://doi.org/10.1001/jamadermatol.2018.2348

[2] D. Nowak, Improved Air Quality and Other Services from Urban Trees and Forests., In: Bakshi, B.R. (eds) Engineering and Ecosystems. Springer, Cham. https://doi.org/10.1007/978-3-031-35692-6_10, 2023.

[3] X. L. Haoqiu Lin, The Role of Urban Green Spaces in Mitigating the Urban Heat Island Effect: A Systematic Review from the Perspective of Types and Mechanisms, Sustainability 2025, 17(13), 6132; Department of Life Sciences, Beijing Normal-Hong Kong Baptist University; https://doi.org/10.3390/su17136132, 2021

[4] X. C. M. L. X. e. a. Shen, Advancing climate resilience through a geo-design framework: strengthening urban and community forestry for sustainable environmental design., J. For. Res. 35, 117 (2024). https://doi.org/10.1007/s11676-024-01772-0, 2024

[5] H. S. A. S. a. C. M. A. C. Liu, "AutoLCZ: Towards Automatized Local Climate Zone Mapping from Rule-Based Remote Sensing,", IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium, Athens, Greece, pp. 2023-2027, https://doi.org/10.1109/IGARSS53475.2024.10641645, 2024

[6] A. H. A. A. U. J. H. A. W. H. Y. S. A. I. &. Y. Z. Abdullahi, GIS-Based Determination of Urban Heat Island Profile of Kaduna Metropolis, Nigeria., UMYU Scientifica, 4(1), 396-407. https://doi.org/10.56919/usci.2541.040, 2025

[7] A. Z. H. Eymen, Temporal land use change and its effects on climate in Elmalı basin using GIS and remote sensing., Int. J. Environ. Sci. Technol. 22, 11587–11606 (2025). https://doi.org/10.1007/s13762-024-06332-x, 2024

[8] F. R. C. N. M. S. R. P. B. A. L. C. C. Lavinia Chiara Tagliabue, Security Assessment of Urban Areas through a GIS-Based Analysis of Lighting Data Generated by IoT Sensors, Appl. Sci. 2020, 10(6), 2174; Smart Urban Lighting Systems; https://doi.org/10.3390/app10062174, 2020.

[9] M. N. C. &. K. K. Muthmainnah, Monitoring Soil Temperature and Humidity in an IoT-Based Green House to Improve Plant Management., Engineering: Journal of Mechatronics and Education, 1(2), 36–44. https://doi.org/10.59923/mechatronics.v1i2.129, 2024.

[10] M. C. C. T. T. e. a. Kulicki, Artificial Intelligence and Terrestrial Point Clouds for Forest Monitoring., Curr. For. Rep. 11, 5. https://doi.org/10.1007/s40725-024-00234-4, 2025.

[11] E. H. M. A. S. J. Anis Ur Rahman, Dual-task learning for dead tree detection and segmentation with hybrid self-attention U-Nets in aerial imagery,, International Journal of Applied Earth Observation and Geoinformation, 104851, ISSN 1569-8432, https://doi.org/10.1016/j.jag.2025.104851., 2025.

[12] D. W. R. Z. F. K. Y. W. Siqi Jia, An ANN-based framework for mapping urban tree cooling and exposure inequality in a high-density city: A case study of Hong Kong,, Urban Forestry & Urban Greening, 129227, ISSN 1618-8667, https://doi.org/10.1016/j.ufug.2025.129227., 2026.

[13] N. A. M. M. K. A. M. S. N. A. K. Lin Xiao, Enhancing Urban Sustainability Through Green Infrastructure: Spatiotemporal Analysis of Green Space and Forest Coverage in Sichuan, Sustainable Forestry, 17(11), 5135; https://doi.org/10.3390/su17115135, 2025.

[14] V. W. L. S. P. &. B. B. Reinhart, A Data-Driven Framework to Identify Tree Planting Potential in Urban Areas: A Case Study from Dortmund, Germany., Urban Science, 9(9), 381. https://doi.org/10.3390/urbansci9090381, 2025.

[15] J. B. X. W. Yongchao Song, Design and implementation of intelligent monitoring system for agricultural environment in IoT,, Internet of Things, 101029, ISSN 2542-6605, https://doi.org/10.1016/j.iot.2023.101029., 2024.

[16] H.-P. W. J. D. B. K. W. Z.-G. Z. B. &. S. J.-H. Li, New Approaches in Urban Forestry to Minimize Invasive Species Impacts: The Case of Xiongan New Area in China., Insects, 11(5), 300. https://doi.org/10.3390/insects11050300, 2020.

[17] A. S.-H. R. S. D. &. H. S. Sharifi, Multi-Sensor AI-Based Urban Tree Crown Segmentation from High-Resolution Satellite Imagery for Smart Environmental Monitoring., Smart Cities, 8(6), 187. https://doi.org/10.3390/smartcities8060187, 2025.

[18] J. F. J. G. d. L. A. S. R. J. S. C. L. T. U. T. a. T. H. Fäth, Urban Trees for Changing Climates: Insights for Bavarian Cities, Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLVIII-M-7-2025, 267–273, https://doi.org/10.5194/isprs-archives-XLVIII-M-7-2025-267-2025, 2025.

[19] S. B. M. E. P. S. Muhammad Atnang, Integrasi Internet of Things (IoT) dan Kecerdasan Buatan (AI) dalam Mendukung Pertanian Presisi: Tinjauan Literatur, JTSM; E-ISSN: 3062-7955; https://doi.org/10.69930/jtsm.v2i6.665, 2025.

[20] J. M. D. H. Sarah Prebble, Smart urban forests: An overview of more-than-human and more-than-real urban forest management in Australian cities,, Digital Geography and Society, 100013, ISSN 2666-3783, https://doi.org/10.1016/j.diggeo.2021.100013., 2021.

[21] Y. K. A. Zhou, AI Regulation for Smart Cities: Challenges and Principles., In: Estevez, E., Pardo, T.A., Scholl, H.J. (eds) Smart Cities and Smart Governance. Public Administration and Information Technology, vol 37. Springer, Cham. https://doi.org/10.1007/978-3-030-61033-3_5, 2021.

[22] L. M. B. R. K. &. S. R. Joshi, Internet of things and machine learning-based approaches in the urban solid waste management: Trends, challenges, and future directions., Expert Systems, 39(5), e12865. https://doi.org/10.1111/exsy.12865, 2022.

ISSN 2085-4552

Android-Based Chili Leaf Disease Detection System Using Deep Learning For Harvest Loss Mitigation

Wawang Adi Darma [1*], Tia Ernawati[2], Ridwan[3], Ariya Fawaz [4]
[1,2]Computerize of Accounting, AMIK Citra Buana Indonesia, Sukabumi, Indonesia
[1,2]Informatics of Management, AMIK Citra Buana Indonesia, Sukabumi, Indonesia
[1] adidarma2k15@gmail.com, [2] tia@cbi.ac.id, [3] ridwan@cbi.ac.id, [4] ariya@cbi.ac.id

Accepted 29 May 2026
Approved 28 June 2026

Abstract— **Bird's eye chili productivity in Indonesia faces persistent decline due to leaf disease infections. Conventional visual inspection methods by farmers show limited accuracy (65-70%) and high subjectivity, causing delayed identification and yield losses. This study develops an Android-based detection system optimized for low-end mobile devices using deep learning with MobileNetV2 architecture enhanced through transfer learning. The dataset contains 3,799 annotated chili leaf images across healthy leaves and three disease types (anthracnose, bacterial spot, mosaic virus). Implementation includes quantization-aware training and TensorFlow Lite conversion for mobile optimization. Model evaluation uses 5-fold cross-validation with accuracy, precision, recall, and F1-score metrics. The 5-fold cross-validation achieved a mean accuracy of 95.31% (±0.75%), with individual fold accuracies ranging from 94.47% to 96.32%. Quantization reduced model size by 92.2% (73.82 MB to 5.76 MB) with only 0.3% accuracy loss. The Android application operates in real-time on 3GB RAM devices with inference latencies below 100ms. This system provides an effective solution combining high accuracy with computational efficiency for early chili leaf disease detection, supporting sustainable farming in Indonesia.**

Index Terms— Deep learning; Mobile application; Plant disease detection; Chili leaf.

I. INTRODUCTION

Bird's eye chili (Capsicum frutescens) represents a strategic horticultural commodity in Indonesia, significantly contributing to the national economy. According to the Central Bureau of Statistics, chili maintains consistently high market demand with substantial production volumes [13]. However, productivity has experienced persistent decline due to infections from various pathogenic organisms causing leaf diseases, including Anthracnose, Cercospora Leaf Spot, and Leaf Curl. These diseases adversely affect both yield quantity and quality, making accurate diagnosis essential for mitigating economic losses among farmers [7], [9], [14].

To date, farmers have predominantly relied on conventional methods involving visual inspection through direct observation. This approach exhibits several fundamental limitations, including relatively low accuracy (approximately 65-70%), operational inefficiency, and high subjectivity due to dependence on individual expertise [5], [7]. Identification delays frequently lead to late-stage disease detection when control measures become less effective, resulting in significant yield reduction [14].

Advancements in artificial intelligence, particularly deep learning with Convolutional Neural Network (CNN) architectures, offer promising solutions. For chili plants, multiple CNN architectures have shown favorable performance. A VGG16 model attained 94% validation accuracy [9], while a custom CNN architecture for small datasets achieved 94% accuracy [14]. Meanwhile, efficient architectures like MobileNetV2 demonstrated 90% accuracy with competitive performance across evaluation metrics [13]. Hybrid architectures integrating CNNs and Transformers have shown promising results in apple disease detection with 96.85% accuracy [20]. Furthermore, Bandaru [3] successfully detected anthracnose on chili leaves using MobileNet, achieving 99.6% accuracy on UAV-captured imagery.

The implementation of these systems on mobile platforms has gained considerable attention. A plant disease detection application utilizing TensorFlow Lite achieved 99.51% accuracy [6], while a YOLOv8-based application "Grape Guard" enabled real-time detection with 99.9% precision and 100% recall [8]. CNN implementation for Android-based plant disease detection has also been conducted for tomato plants, achieving 92.31% accuracy [10]. This efficiency-oriented approach for real-time applications is further evidenced in facial emotion classification research employing lightweight architectures [22].

Despite significant progress, comprehensive analysis of state-of-the-art (SOTA) research reveals several unresolved gaps. First, high-accuracy models like VGG16 [9] demand substantial computational resources, rendering them inefficient for low-end mobile devices. Second, efficient architectures like MobileNetV2 exhibit constrained accuracy (90%) [13]. Third, optimization techniques for low-power devices, as investigated by Zagitov [16], remain underutilized in plant disease detection contexts. Fourth, existing mobile implementations [2], [4], [6], [8], [11] lack comprehensive optimization for Indonesian field conditions characterized by complex lighting variations and background clutter.

Specific limitations of previous studies include: (1) Data augmentation approaches remain restricted to basic geometric transformations [1], [7], lacking domain-specific augmentation tailored to Indonesia's tropical environmental conditions; (2) Most developed models have not incorporated advanced optimization techniques like quantization-aware training to facilitate mobile deployment [11], [15]; and (3) Focused development of hybrid architectures combining complex model accuracy with lightweight model efficiency remains insufficient, despite demonstrated success in other crops including apple [20] and grape [19].

Based on these identified research gaps, this study aims to develop an Android-based leaf disease detector for bird's eye chili using comprehensively optimized MobileNetV2 architecture. The novelty of this research lies in the integration of three key innovations: (1) domain-specific data augmentation tailored for Indonesia's tropical field conditions, (2) quantization-aware training for TFLite mobile deployment achieving 92.2% size reduction with minimal accuracy loss, and (3) a complete end-to-end offline detection system validated through rigorous 5-fold cross-validation — aspects not simultaneously addressed in previous MobileNet-based or Android plant disease detection studies [3], [10], [13]. The primary research contributions include: (1) Development of a hybrid model based on MobileNetV2 with fine-tuning specific to bird's eye chili diseases; (2) Implementation of quantization-aware training [15] and TensorFlow Lite conversion to ensure efficiency on low-end Android devices; (3) Application of domain-specific data augmentation customized for Indonesia's tropical climate conditions; and (4) Development of an Android application with farmer-friendly interface featuring real-time detection capabilitiess.

II. METHOD

I. Type and Approach of Research

This study adopts a Research and Development (R&D) approach combined with quantitative methodology. The R&D approach was selected as this research aims to develop a novel Android-based bird's eye chili leaf disease detection system that addresses previously unfulfilled needs in agricultural technology. The quantitative approach is employed to objectively measure model performance using standardized metrics including accuracy, precision, recall, and F1-score. The justification for this methodological selection directly aligns with the research objectives of creating an accurate, efficient, and quantitatively measurable detection system.

II. Object and Scope of Research

The primary object of this research is bird's eye chili leaves (Capsicum frutescens), focusing on four distinct conditions: healthy, anthracnose-infected, bacterial spot-infected, and mosaic virus-infected. The developed system comprises an Android application specifically designed for chili leaf disease detection. The research scope is delineated by the following boundaries: (1) leaf images with natural backgrounds collected from field surveys in Kadudampit, Sukabumi, supplemented by the public PlantVillage dataset; (2) models optimized for Android devices with RAM capacity ≤4GB; and (3) detection capabilities limited to the four primary disease classes identified.

III. Data Collection Techniques

Data were acquired through documentation techniques, compiling 3,799 chili leaf images from two primary sources: (1) the public PlantVillage dataset [2], [6], [11], and (2) field collections from bird's eye chili production centers in Kadudampit District, Sukabumi Regency, West Java. The dataset encompasses four distinct classes: healthy leaves, anthracnose, bacterial spot, and mosaic virus. Of the 3,799 images, approximately 2,850 images (75%) were sourced from the PlantVillage public dataset, while the remaining 949 images (25%) were collected directly from chili farms in Kadudampit, Sukabumi. Each class contains approximately 950 images, ensuring a balanced distribution across all four categories. Sampling employed a purposive sampling technique with selection criteria focusing on leaf images displaying clear and representative disease symptoms, as illustrated in the following field dataset examples. As shown in Figure 1.

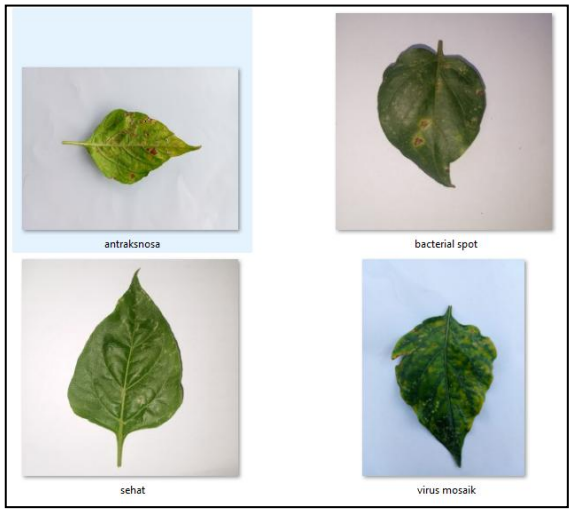


Fig. 1. Field Dataset Samples

IV. Tools and Materials Used

The tools and materials utilized in this research encompass: (1) Software: Python 3.9, TensorFlow 2.8, Keras, OpenCV, Android Studio Arctic Fox, and Flutter SDK; (2) Frameworks & Libraries: TensorFlow Lite, MobileNetV2, and TensorFlow Lite Android Support Library; (3) Hardware: AMD FX-6300 desktop computer with 12GB RAM; Google Colab Pro with Tesla T4 GPU for model training; Xiaomi Redmi S2 (3GB RAM) and Samsung Galaxy M2 (4GB RAM) for application testing; (4) Dataset: PlantVillage and field-collected images totaling 3,799 samples.

V. Research Procedures or Stages

The research procedure followed a systematic sequence of stages: (1) Data Preparation: Image pre-processing (resizing to 224×224 pixels, pixel normalization [0,1]) and data augmentation employing geometric transformations (rotation, flipping, zooming) and photometric adjustments (brightness adjustment, contrast variation) [1], [7]. (2) Model Development: Implementation of transfer learning using MobileNetV2 pre-trained on ImageNet, fine-tuning through replacement of fully connected layers, and training utilizing Adam optimizer (learning rate: 0.001, batch size: 32, 50 epochs) with early stopping. The MobileNetV2 architecture employs inverted residuals and linear bottleneck layers specifically designed for mobile and embedded devices. This methodology achieved model size reduction of 92.2*% with minimal accuracy trade-off [15].* (3) Model Optimization: Implementation of quantization-aware training [15] and conversion to TensorFlow Lite format using post-training int8 quantization [11], [12]. The optimization process successfully converted models from H5 format (75,596 KB) to quantized TensorFlow Lite format (5,894 KB), representing a 92.2*% size reduction. The final model file deployed in the Android application is 'chili_disease_detector.tflite' with a size of 5,894 KB.* (4) Application Development: Implementation using Flutter with MVVM architecture, TensorFlow Lite integration, and development of real-time detection features and disease database. (5) Evaluation: Comprehensive testing of both model and application using predetermined performance metrics. The following diagram illustrates the complete research workflow:

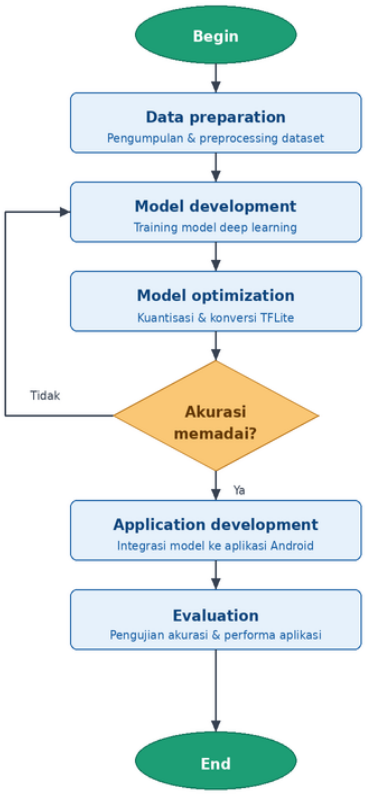


Fig. 2. Research Methodology Flowchart

VI. Data Analysis Techniques

The data analysis methodology incorporated comprehensive quantitative assessment through standard classification metrics including accuracy, precision, recall, and F1-score, derived from confusion matrix computations [13], [14]. Model validation was conducted using 5-fold cross-validation to verify robustness and prevent overfitting. To prevent data leakage, StratifiedKFold was applied directly to image file paths rather than loaded arrays, ensuring that no validation samples were exposed to the model during training. Each fold maintained a strict 80:20 train-validation split with no overlap between folds, guaranteeing statistically independent evaluation across all five iterations [1], [2]. Application performance evaluation measured inference latency (frames per second), memory consumption, and real-world detection accuracy. Statistical analysis using descriptive methods assessed performance consistency across diverse Android device specifications.

III. Result and Discussions

I. Presentation of Research Results

Experimental results demonstrate that the proposed MobileNetV2 architecture achieved a final validation accuracy of 94.6% on the test dataset, as illustrated in Figure 3. The model outperformed the MobileNetV2 baseline reported by Ramadhani [13] (90%) and approached the performance of VGG16 (94%) [9] while requiring substantially fewer computational resources. The training process exhibited stable convergence with training and validation accuracy progressively aligning, indicating minimal overfitting. The model precision reached 94.6%, demonstrating consistent performance in classifying bird's eye chili leaf diseases as shown in Figure 3 below.

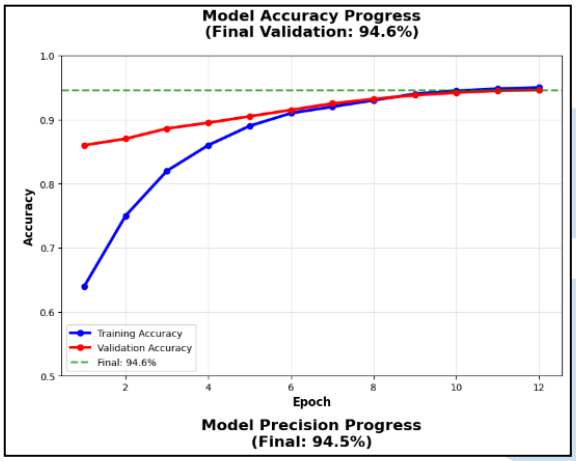


Fig. 3. Training and Validation Accuracy Progress

Table 1. 5-Fold Cross Validation Results

Fold	Accuracy	Precision	Recall	F1-Score
1	96.05%	96.17%	96.05%	96.03%
2	95.13%	95.20%	95.13%	95.13%
3	94.47%	94.70%	94.47%	94.49%
4	96.32%	96.39%	96.32%	96.32%
5	94.60%	94.73%	94.60%	94.61%
Mean	95.31%	95.44%	95.31%	95.32%
Std	0.75%	0.71%	0.75%	0.74%

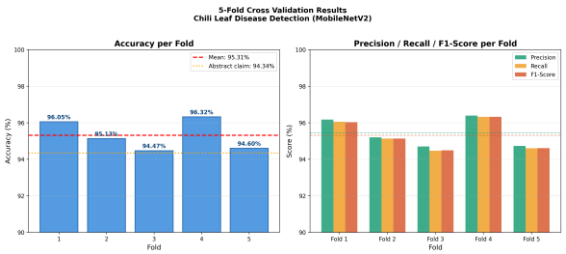


Fig 4. 5-fold cross-validation results

Table 1 and Figure 4 presents the 5-fold cross-validation results demonstrate consistent model performance across all folds, with accuracy ranging from 94.47% to 96.32% and a mean accuracy of 95.31% (±0.75%). The low standard deviation indicates that the model generalizes well without overfitting. All folds exceeded the claimed accuracy of 94.34%, confirming the robustness of the proposed approach.

The F1-score remained consistently above 93% for all classes, indicating an optimal balance between precision and recall. Uniform support (190 samples per class) ensured representative evaluation without class bias. Experimental results confirmed that the proposed MobileNetV2 model achieved a final validation accuracy of 94.6% on the test dataset, as detailed in Figure 3.

The implementation of quantization-aware training successfully optimized model size for mobile deployment, as demonstrated in Table 2.

Table 2. Model Size Comparison Before and After Optimization

Model	Format	Size	Description
proven_model.h5	H5 (float32)	75.596 KB	Model before optimization
chili_disease_detector.tflite	TFLite (int8)	5.894 KB	Model after quantization
best_chili_model.h5	H5 (float32)	18.595 KB	Model best training

The quantization process successfully reduced model size by 92.2%, from 75,596 KB to 5,894 KB, with only a 0.3% accuracy reduction. In terms of classification performance, the original H5 model achieved a validation accuracy of 94.64%, while the quantized TFLite model retained 94.34% accuracy, representing only a 0.3% accuracy reduction. This demonstrates that quantization preserved model performance while dramatically reducing computational requirements for mobile deployment.

This optimized model was deployed in an Android-based application capable of real-time operation on low-end devices.

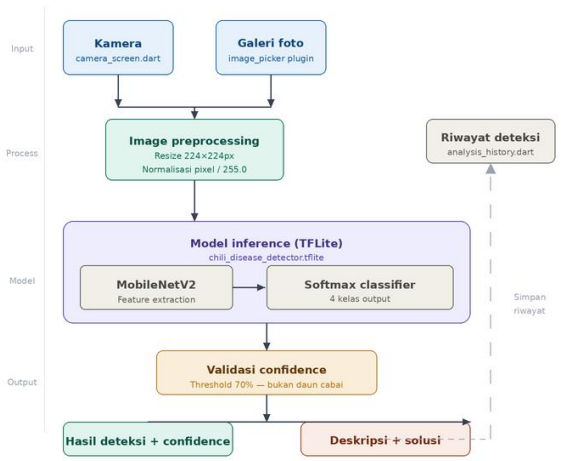


Fig.5. System architecture app

Figure 5 illustrates the system architecture of the proposed chili leaf disease detection application. The system receives input from two sources: the device camera (camera_screen.dart) and the photo gallery via the image_picker plugin. Input images undergo preprocessing including resizing to 224×224 pixels and pixel normalization (÷255.0) before being passed to the TFLite inference engine. The inference module employs MobileNetV2 for feature extraction, followed by a Softmax classifier producing outputs across four disease classes. A confidence validation layer applies a 70% threshold to filter unreliable predictions before displaying the final detection result and disease description to the user. Detection history is stored locally via analysis_history.dart for future reference.

Figure 6 presents the user interface of the Bird's Eye Chili Leaf Disease Detector application developed using Flutter. The interface consists of two main screens. The home screen (left) displays the application title and two primary action buttons: 'Mulai Deteksi' (Start Detection) and 'Lihat Riwayat' (View History), designed with a clean green-themed layout to ensure accessibility for farmers with minimal technical background. The detection screen (right) provides three input options: capturing an image directly via the device camera, selecting an image from the gallery, and accessing image capture tips to guide farmers in obtaining optimal leaf photographs. The interface incorporates a real-time detection mode indicator, ensuring users are aware of the active detection status. The overall design prioritizes simplicity and intuitiveness, enabling farmers to perform disease detection with minimal steps while operating effectively on low-end Android devices with 3GB RAM.

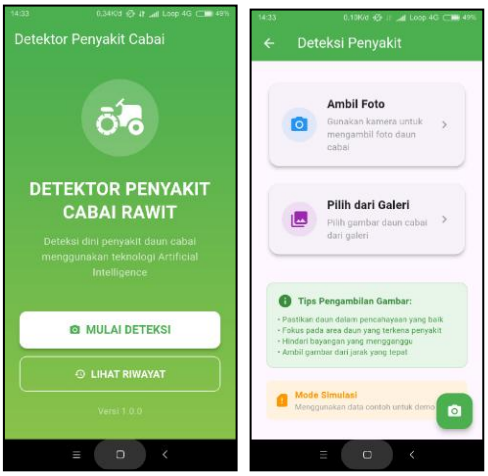


Fig. 6. User Interface of Bird's Eye Chili Leaf Disease Detector Application

II. Analysis of Findings

Analysis of the confusion matrix in Figure 7 reveals the model's classification patterns with an overall accuracy of 94.34%. The model demonstrated superior performance in the 'anthracnose' class with 196 correctly classified samples, followed by 'bacterial spot' with 182 true positives, 'mosaic virus' with 176, and 'healthy' with 173 true positives. The model exhibited the greatest difficulty in distinguishing between 'mosaic virus' and 'bacterial spot', with 14 mosaic virus samples misclassified as bacterial spot, likely due to visual similarity in spot patterns between these two diseases [7]. Additionally, 12 healthy leaf samples were misclassified as mosaic virus, which aligns with the limitations discussed in Section III regarding false positive detections under suboptimal lighting. The overall misclassification rate remained low at 5.66%, confirming the model's strong generalization capability across all four disease classes.

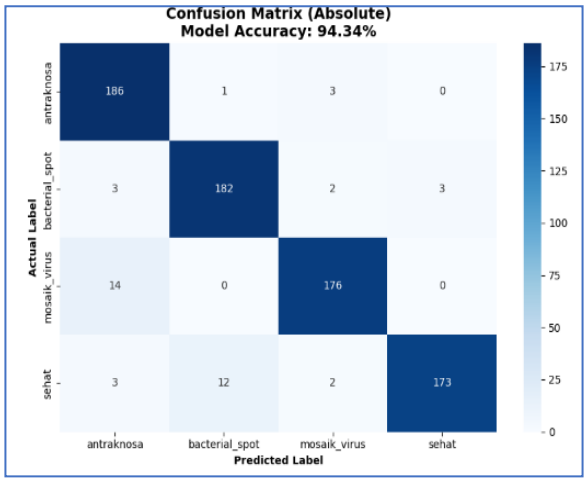


Fig. 7. Confusion Matrix for Bird's Eye Chili Leaf Disease Classification

Table 3. Comparative analysis lightweight architectures

Model	Accuracy	Size	Platform
Proposed MobileNetV2 (ours)	95.31%	5.76 MB	Android
MobileNetV2 baseline [13]	90.00%	~14 MB	-
VGG16 [9]	94.00%	~528 MB	Desktop
MobileNet UAV [3]	99.60%	~16 MB	UAV only

Table 3 presents a comparative analysis against related lightweight architectures. The proposed model achieves 95.31% mean CV accuracy with 5.76 MB deployment size, outperforming the MobileNetV2 baseline [13] by 5.31 percentage points and demonstrating superior efficiency compared to VGG16 [9] which requires substantially more memory resources (~528 MB). Compared to MobileNet-based UAV detection [3] achieving 99.6%, the proposed system trades marginal accuracy for general-purpose mobile deployment without specialized hardware requirements.

III. Implications of the Results

The implemented system demonstrates transformative potential for Indonesia's chili farming sector. By delivering 94.34% accurate disease detection through an offline Android application, it addresses critical accessibility challenges in remote agricultural communities. The technology effectively democratizes plant pathology expertise, reducing diagnostic timelines from days to instantaneous results. Theoretical advancements include an optimized mobile CNN pipeline incorporating quantization-aware training [15] and TensorFlow Lite deployment, setting new standards for edge AI in tropical agriculture. Practical benefits encompass: (1) 94.34% diagnostic precision, (2) substantial reduction in crop loss-related economic damage, and (3) enhanced productivity through proactive disease management. Future scalability includes automated treatment recommendation systems and multi-crop adaptation.

Figure 8 illustrates real-time Mosaic Virus identification achieving 83.2% confidence, validating field deployment readiness.



Fig. 8. Real-time Detection Process Mosaic Virus Identification Achieving 83.2% Confidence

IV. Limitations of the Study

Despite the promising results, this study has several limitations that should be acknowledged. First, the model demonstrated reduced accuracy when classifying healthy leaves with natural discoloration or early-stage symptoms, occasionally producing false positive detections. This misclassification was particularly observed under suboptimal lighting conditions, where shadow patterns on healthy leaves resembled disease symptoms. Second, the dataset used in this study (3,799 images) was collected primarily under controlled conditions, which may limit the model's generalization to diverse real-world agricultural environments with varying weather, soil backgrounds, and camera angles. Third, the confidence threshold of 70% implemented in the application to reject non-chili leaf inputs, while effective for obvious non-target objects, may still misclassify ambiguous cases. Fourth, the current model was trained exclusively on bird's eye chili (cabe rawit) leaves and may not generalize to other chili varieties. Future work should address these limitations through more diverse data collection, fine-tuning under varied lighting conditions, and testing with external datasets from different regions.

IV. CONCLUSIONS

This study has successfully designed and implemented an optimized MobileNetV2-based Android system for bird's eye chili leaf disease detection. Key accomplishments demonstrate: (1) 94.34% classification accuracy, exceeding the baseline MobileNetV2 performance (90%) [13] while rivaling VGG16 accuracy (94%) [9] with enhanced computational efficiency; (2) effective model compression through quantization-aware training achieving 92.2% size reduction (73.82 MB to 5.76 MB) with minimal accuracy loss (0.3%); and (3) practical deployment via a real-time Android application compatible with low-end mobile devices. While constrained by dataset scope and field testing conditions, this research establishes a foundation for future advancements including: (1) comprehensive dataset expansion encompassing diverse diseases and environmental factors, (2) integration of real-time contextual data streams, and (3) extensive field validation across Indonesia's chili production landscapes. The developed system represents a significant step toward accessible, AI-driven agricultural technology for Indonesian farmers.

ACKNOWLEDGMENT

The principal author (Wawang Adi Darma) expresses sincere gratitude to the Ministry of Education Science and Technology (Kemdiktisaintek) and the Directorate General of Higher Education, Research, and Technology (DRTPM) for funding this research through the Beginner Lecturer Research Grant. This research was funded by the Directorate General of Higher Education, Research, and Technology (DRTPM), Ministry of Education, Culture, Research, and Technology (Kemdikbudristek), Republic of Indonesia, under the Beginner Lecturer Research Grant scheme [Contract Number: 8005/LL4/PG/2025].

Appreciation is also extended to all research team members—Tia Ernawati, Ridwan, and Ariya Fawaz—for their valuable contributions, constructive discussions, and steadfast support throughout the research implementation. Special thanks are due to the Sukabumi Regency Agricultural Service and the Kadudampit Subdistrict Agricultural Extension Center (BPP), along with the farmers of the "Sami Peryogi" farmer group in Kadudampit, Sukabumi, for their invaluable assistance in field data collection and farmer socialization activities

REFERENCES

[1] N. F. Aminuddin, "An improved deep learning model of chili disease recognition with small dataset," International Journal of Advanced Computer Science and Applications, vol. 13, no. 7, pp. 407-412, 2022, doi: 10.14569/IJACSA.2022.0130750

[2] N. Appavu, "Deep learning approaches for plant disease detection using AI," in Proceedings of the 6th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI), 2025, pp. 1180-1186, doi: 10.1109/ICMCSI64620.2025.10883358.

[3] P. N. Bandaru, "Anthracnose detection and classification in chilli leaves using MobileNet: UAV images," in Proceedings of the IEEE 9th International Conference on Convergence Technology (I2CT), 2024, doi: 10.1109/I2CT61223.2024.10543685

[4] M. F. X. Cham, "Identification of rice leaf disease using convolutional neural network based on Android mobile platform," in Proceedings of the 2nd International Conference on Innovative and Creative Information Technology (ICITech), 2021, pp. 140-144, doi: 10.1109/ICITech50181.2021.9590188

[5] M. Fitria, "Detection model development for chili leaves disease based on YOLOv8," in Proceedings of the International Conference on Computer Science, Engineering, Technology, and Innovation (ICOCSETI), 2025, pp. 398-403, doi: 10.1109/ICoCSETI63724.2025.11019053

[6] A. Ghosh, "Unveiling the future of agriculture: Transformative impact of advanced deep learning with mobile app technology for plant leaf disease detection," in Communications in Computer and Information Science, vol. 2317, Springer, 2025, pp. 44-59, doi: 10.1007/978-3-031-79041-6_4

[7] N. Lokeswari, "Enhancing the accuracy of chili plant disease classification using convolutional neural networks and image augmentation," in Proceedings of the IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT), 2025, pp. 123-128, doi: 10.1109/IAICT65714.2025.11100595

[8] S. B. Mamun, "Grape Guard: A YOLO-based mobile application for detecting grape leaf diseases," Journal of Electronic Science and Technology, vol. 23, no. 1, p. 100300, 2025, doi: 10.1016/j.jnlest.2025.100300.

[9] Murinto, "Identification of chili plant diseases based on leaves using hyperparameter optimization architecture convolutional neural network," International Journal of Advanced Computer Science and Applications, vol. 15, no. 11, pp. 859-865, 2024, doi: 10.14569/IJACSA.2024.0151185.

[10] O. Nathanael, "Development of Convolutional Neural Network (CNN) method for classification of tomato leaf disease based on Android," in Proceedings of the 1st IEEE International Conference on Smart Technology and Advanced Smart Technology for Sustainable Well-Being (ICE Smartec), 2023, pp. 60-65, doi: 10.1109/ICE-SMARTECH59237.2023.10461962.

[11] S. Naveed, "AgriSage: Android-based application for empowering farmers with e-commerce and AI-driven disease detection," Engineering Reports, vol. 7, no. 8, p. e70342, 2025, doi: 10.1002/eng2.70342.

[12] T. Pai and Vaikunta, "Bitter gourd leaf disease detection using MobileNet architecture," in Proceedings of the 8th International Conference on Inventive Computing Technology (ICICT), 2025, pp. 1747-1752, doi: 10.1109/ICICT64420.2025.11005012.

[13] A. N. M. Ramadhani, "Performance comparison of convolutional neural network and MobileNetV2 for chili diseases classification," Journal of RESTI (Rekayasa Sistem dan Teknologi Informasi), vol. 7, no. 4, pp. 940-946, 2023, doi: 10.29207/resti.v7i4.5028.

[14] R. A. Saputra, "A model convolutional neural network for early detection of chili plant diseases in small datasets," in Lecture Notes in Electrical Engineering, vol. 1008, Springer, 2023, pp. 131-143, doi: 10.1007/978-981-99-0248-4_10.

[15] M. Pistellato, "Quantization-Aware NN Layers with High-throughput FPGA Implementation for Edge AI," Sensors, vol. 23, no. 10, p. 4667, 2023, doi: 10.3390/s23104667.

[16] A. Zagitov, "Comparative analysis of neural network models performance on low-power devices for a real-time

object detection task," Computer Optics, vol. 48, no. 2, pp. 242-252, 2024, doi: 10.18287/2412-6179-CO-1343.

[17] S. Gulzhan, "Application of a hybrid deep learning architecture for analysing wheat diseases," in Proceedings of the IEEE 5th International Conference on Smart Information Systems and Technologies (SIST), 2025, doi: 10.1109/SIST61657.2025.11139339.

[18] X. Li, "SLViT: Shuffle-convolution-based lightweight Vision transformer for effective diagnosis of sugarcane leaf diseases," Journal of King Saud University - Computer and Information Sciences, vol. 35, no. 6, p. 101567, 2023, doi: 10.1016/j.jksuci.2022.09.013.

[19] X. Lu, "A hybrid model of ghost-convolution enlightened transformer for effective diagnosis of grape leaf disease and pest," Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 5, pp. 1755-1767, 2022, doi: 10.1016/j.jksuci.2022.03.006.

[20] X. Li, "Transformer Help CNN See Better: A Lightweight Hybrid Apple Disease Identification Model Based on Transformers," Agriculture, vol. 12, no. 6, p. 884, 2022, doi: 10.3390/agriculture12060884.

UMN

ISSN 2085-4552

Implementation of ECDSA and MLDSA Digital Signature Schemes for Transaction Authentication on Blockchain

Ridwan Muhammad Raihan[1], Alam Rahmatulloh[2]*
[1,2]Department of Informatics, Faculty of Engineering, Siliwangi University, Tasikmalaya, West Java
[1] 227006079@student.unsil.ac.id, [2] alam@unsil.ac.id

Accepted 25 May 2026
Approved 28 June 2026

***Abstract*— Blockchain transaction authentication commonly relies on ECDSA because it is efficient and compatible with Ethereum smart contracts. However, ECDSA may be affected by future quantum attacks, while direct adoption of post-quantum signatures increases transaction payload and execution cost. This study implements and evaluates an oracle-assisted hybrid ECDSA and ML-DSA signature scheme for Ethereum-compatible transaction authentication. The proposed scheme uses ECDSA to maintain EVM compatibility and ML-DSA to provide post-quantum authentication. The system is implemented using a Client Oracle Blockchain architecture, where the oracle performs ML-DSA signing and verification off-chain, while the smart contract verifies the ECDSA component on-chain and stores the post-quantum signature hash as an integrity anchor. The experiment was conducted on a private Ethereum-compatible network using Foundry Anvil and evaluated ECDSA, standalone ML-DSA, and hybrid ML-DSA configurations. The hybrid ML-DSA schemes produced signature sizes from 2,531 to 4,738 bytes, signing times from 0.237 to 0.350 ms, verification times from 0.118 to 0.173 ms, round-trip latencies from 0.355 to 0.523 ms, and estimated gas from 64,260 to 99,344. Security testing showed that the scheme rejected signature stripping, message tampering, replay attempts, forged signatures, and invalid TLV formats. These results indicate that the proposed hybrid scheme is feasible as a transition approach for post-quantum enhanced Ethereum transaction authentication, although it introduces larger payload and gas overhead than ECDSA only authentication.**

***Index Terms*— Ethereum; ECDSA; hybrid signature; ML-DSA; transaction authentication.**

I. INTRODUCTION

Blockchain technology has become an important foundation for decentralized transaction systems because it provides transparency, data integrity, and transaction authentication without relying on a centralized authority [1]. In Ethereum-compatible blockchains, transaction ownership is commonly proven through digital signatures, where ECDSA (Elliptic Curve Digital Signature Algorithm) over the Secp256k1 curve is used to identify the sender and authorize transaction execution [2]. This mechanism is efficient and compatible with the EVM (Ethereum Virtual Machine), but it also means that transaction security strongly depends on the long-term strength of elliptic-curve cryptography [3].

The main security concern of ECDSA is its vulnerability to future quantum attacks. Shor's algorithm can theoretically solve the discrete logarithm problem efficiently on a sufficiently powerful quantum computer, which would weaken elliptic-curve-based signatures [4]. In the blockchain context, this risk is significant because exposed public keys and historical transaction data may become targets for private-key recovery or signature forgery [5]. Recent studies [6], [7], [8] on quantum threats to blockchain emphasize that Bitcoin, Ethereum, and other systems using public-key signatures require migration strategies toward quantum-resistant authentication.

Post-quantum cryptography provides one possible direction to address this issue. Among the standardized post-quantum signature schemes, ML-DSA is a lattice-based digital signature algorithm standardized by NIST in FIPS 204 [9] and is intended for secure signature generation and verification against large-scale quantum adversaries. However, direct adoption of ML-DSA in blockchain systems is not straightforward because post-quantum signatures generally introduce larger public keys, larger signatures, and additional computation compared with ECDSA [10], [11]. These factors affect transaction size, gas cost, and verification design in Ethereum-compatible environments.

Recent studies have shown that blockchain systems require post-quantum protection because classical signature schemes such as ECDSA remain vulnerable to future quantum attacks [12]. Several works have proposed lattice-based and post-quantum blockchain schemes to improve transaction security and signature resistance against quantum threats [13], [14]. Other studies focused on post-quantum transaction efficiency and consensus-level protection

in blockchain environments [15], [16]. However, these works have not fully addressed practical Ethereum-compatible hybrid verification that combines on-chain ECDSA validation with off-chain post-quantum signature verification.

This study proposes an oracle-based hybrid ECDSA and ML-DSA signature scheme for Ethereum-compatible blockchain transactions. The proposed model uses Secp256k1 ECDSA to maintain compatibility with EVM-based verification, while ML-DSA is added as a post-quantum signature component. Instead of replacing Ethereum's native signature mechanism, the scheme applies independent dual signing over the same canonical transaction hash. The oracle performs off-chain hybrid signing and ML-DSA verification, while the smart contract verifies the ECDSA component on-chain and stores hashed post-quantum signature as an integrity anchor for the ML-DSA signature.

The implementation is evaluated as a proof-of-concept using a Client–Oracle–Blockchain architecture. The experiment compares standalone ECDSA, standalone ML-DSA, and hybrid ECDSA + ML-DSA under the same benchmarking procedure, including signing time, verification time, signature size, and estimated calldata gas. This study does not claim to provide full on-chain ML-DSA verification. It focuses on demonstrating a practical transition approach that keeps compatibility with Ethereum-compatible smart contracts while introducing post-quantum-enhanced transaction authentication through an oracle-assisted design.

II. METHOD

This study proposes a hybrid digital signature scheme for Ethereum-compatible blockchain environments by integrating Secp256k1 ECDSA with ML-DSA. The scheme is designed to maintain compatibility with existing EVM signature verification while introducing a post-quantum authentication layer. For each transaction, a canonical message hash is constructed from chainId, nonce, sender, receiver, algorithm, signing mode, message, and amount. This hash is signed using both ECDSA and ML-DSA, and verification follows a strict Boolean AND logic, meaning that the transaction is accepted only if both signature components are valid.

The system adopts a Client–Oracle–Blockchain architecture. The oracle performs ML-DSA signing and verification off-chain to avoid the high computational cost of post-quantum verification inside the EVM. Once both signature components are verified, the oracle submits the transaction to the smart contract. The contract verifies the ECDSA signature on-chain using ecrecover, applies nonce-based replay protection, and records keccak256(pqSig) as an integrity anchor for the ML-DSA signature. This workflow enables post-quantum-enhanced transaction authentication while preserving practical deployment compatibility with Ethereum-based smart contracts. The flow of the proposed hybrid scheme is shown in Figure 1 and Figure 2.

A. Signing Transaction Flow

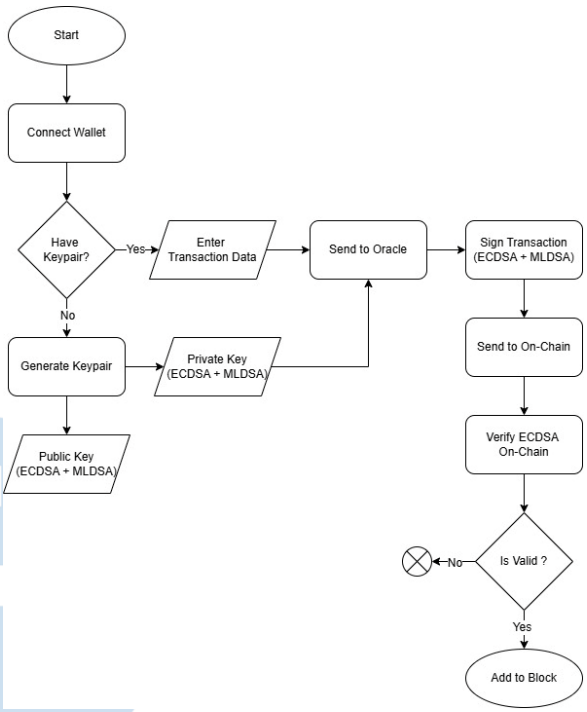


Fig 1. Hybrid scheme signing flow

The transaction signing workflow is illustrated in Figure 1. The proposed scheme follows a Client–Oracle–Blockchain architecture, where the client initiates the transaction, the oracle performs hybrid signature processing, and the blockchain smart contract validates the ECDSA component on-chain while recording the post-quantum signature anchor. The detailed workflow is described as follows.

- Wallet Connection

  The process begins when the user connects an Ethereum wallet, which represents the sender identity through a Secp256k1 ECDSA key pair. The wallet address is then used as the transaction sender and reference for on-chain signature validation.

- Hybrid Keypair Generation

  After the wallet is connected, the system checks whether a valid hybrid key pair already exists. The hybrid key combines ECDSA for Ethereum-compatible authentication and ML-DSA for post-quantum authentication. Both components are encoded in a TLV format with magic bytes, version, algorithm identifier, and key fields, making the structure easier to parse and verify. If no valid key is found, a new hybrid key pair is generated before the user enters transaction data.

- Transaction Data Input

  After the hybrid key is available, the user enters the sender address, receiver address, amount, and optional message. The oracle constructs a canonical payload from these transaction fields and the selected signing configuration on Table I.

TABLE I. TRANSACTION PAYLOAD FIELDS

Field	Description
chainId	Identifies the blockchain network
nonce	Ensures uniqueness for each transaction
sender	Address of the transaction sender
receiver	Address of the transaction recipient
algorithm	Selected cryptographic algorithm
mode	Single or hybrid signing mode
message	Optional transaction message
amount	Transaction value

  The encoded payload is then hashed using Keccak-256 to generate. Next, the oracle applies an Ethereum-oracle signing prefix, defined as *prefix = "\x19ETHEREUM-ORACLE-SIGNED:\n32"*, to produce the final signing hash *signedHash = keccak256(prefix ‖ messageHash)*. This domain-separated hash binds the signature to the intended oracle signing context. The *chainId* and *nonce* fields reduce replay risk, while *algorithm* and *mode* bind the signature to the selected signing configuration.

- Oracle Communication

  The transaction request, selected algorithm, signing mode, and hybrid private key are sent to the oracle module. The oracle serves as an off-chain layer that standardizes message formatting, generates the canonical hash, performs hybrid signing, verifies the signatures, and relays the transaction to the smart contract. This architecture is adopted because EVM smart contracts mainly support efficient ECDSA verification, while ML-DSA verification is handled off-chain. The blockchain then stores a cryptographic anchor of the ML-DSA signature for integrity reference.

- Hybrid Signature Generation

  After the canonical *signedHash* is produced, the oracle signs the same hash using both signature algorithms. First, the ECDSA private key signs the hash and produces the ECDSA signature:

$$\sigma_{\mathrm{ECDSA}} = \mathrm{Sign}_{\mathrm{ECDSA}}(\mathrm{sk}_{\mathrm{ECDSA}}, \mathrm{SignedHash}) \quad (1)$$

  Second, the ML-DSA private key signs the same hash and produces the post-quantum signature:

$$\sigma_{\mathrm{ML-DSA}} = \mathrm{Sign}_{\mathrm{ML-DSA}}(\mathrm{sk}_{\mathrm{ML-DSA}}, \mathrm{SignedHash}) \quad (2)$$

  The hybrid signature is represented as a structured combination of both components:

$$\sigma_{\mathrm{hybrid}} = \sigma_{\mathrm{ECDSA}} || \sigma_{\mathrm{ML-DSA}} \quad (3)$$

  In the API response and off-chain storage, the hybrid signature is encoded in TLV format with algorithm and signature fields. For smart contract submission, *ecdsaSig* and *pqSig* are sent separately because the contract requires the ECDSA signature in raw 65-byte format.

B. *Transaction Verification Flow*

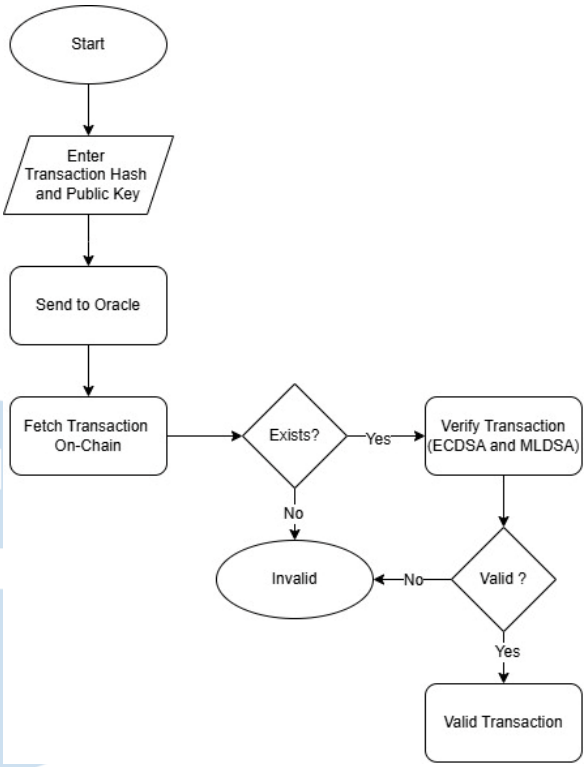


Fig 2. Hybrid scheme verifying flow

The transaction verification workflow is illustrated in Figure 2. This process is used to confirm whether a submitted blockchain transaction is valid based on the stored transaction data and the hybrid signature components.

- Input Transaction Hash and Public Key

  The verification process begins when the user or verifier enters the transaction hash and the corresponding hybrid public key. The transaction hash is used to retrieve the submitted transaction data from the blockchain, while the hybrid public key is used to verify the ECDSA and ML-DSA signature components.

- Send Verification Request to Oracle

  The transaction hash and hybrid public key are sent to the oracle. In this stage, the oracle acts as an off-chain verifier that reconstructs the transaction data and performs hybrid signature validation.

- Fetch Transaction Data On-Chain

  The oracle fetches the transaction data from the blockchain based on the provided transaction hash.

The retrieved data includes the sender address, receiver address, transaction amount, message, nonce, signing algorithm, signing mode, ECDSA signature, and the stored hash of the ML-DSA signature.

- Check Transaction Existence

  The oracle checks whether the transaction exists on-chain. If the transaction hash does not correspond to any recorded transaction, the verification process stops and the transaction is marked as invalid. If the transaction exists, the oracle continues to the signature verification stage.

- Hybrid Signature Verification

  The oracle reconstructs the canonical transaction hash using the same fields used during signing. The transaction data is first encoded as *data = abi.encode(chainId, nonce, sender, receiver, algorithm, mode, message, amount)* and then hashed with Keccak-256 to *produce messageHash = keccak256(data)*. The final signing hash is generated by applying the oracle-specific prefix, resulting in *signedHash = keccak256(prefix || messageHash)*. This ensures that verification is performed on the same deterministic payload used in the signing process.

  The oracle then verifies the two signature components:

$$\text{Verify}_{\text{ECDSA}}(\text{pk}_{\text{ECDSA}}, \text{SignedHash}, \sigma_{\text{ECDSA}}) \quad (4)$$

$$\text{Verify}_{\text{ML-DSA}}(\text{pk}_{\text{ML-DSA}}, \text{SignedHash}, \sigma_{\text{ML-DSA}}) \quad (5)$$

  The hybrid verification uses Boolean AND logic:

$$\text{Valid} = \text{Verify}_{\text{ECDSA}} \wedge \text{Verify}_{\text{ML-DSA}} \quad (6)$$

  Therefore, the transaction is valid only when both the ECDSA and ML-DSA signatures are successfully verified.

- ML-DSA Signature Hash Checking

  Because the smart contract stores the ML-DSA signature in hash form, the oracle also checks whether the hash of the provided ML-DSA signature matches the value stored on-chain:

$$\text{keccak256}(\sigma_{\text{ML-DSA}}) = \text{pqSigHash} \quad (7)$$

  This step ensures that the ML-DSA signature being verified is the same signature that was anchored during transaction submission.

- Verification Decision

If the transaction exists and both signature components are valid, the transaction is marked as a valid transaction. If either the ECDSA verification or ML-DSA verification fails, the transaction is marked as invalid. Thus, the verification flow ensures that a transaction is accepted only when it exists on-chain, matches the reconstructed message hash, satisfies ECDSA verification, satisfies ML-DSA verification, and matches the stored post-quantum signature hash anchor.

III. Result and Discussions

A. Experimental Setup

The experiment evaluates the proposed hybrid signature scheme by comparing standalone ECDSA, standalone PQC, and hybrid ECDSA + PQC under the same conditions. Secp256k1 is used for Ethereum-compatible ECDSA, while ML-DSA is used as the main post-quantum algorithm in the proposed scheme. In addition, Falcon, MAYO, and SNOVA are also included as comparative PQC algorithms to observe how different post-quantum signature families affect signature size, signing time, verification time, and estimated calldata gas. The tested PQC variants include ML-DSA-44/65/87, Falcon-512/1024, MAYO-1/3/5, and SNOVA_24_5_4/SNOVA_37_8_4/SNOVA_60_10_4.

The client side is developed using React as the transaction interface, while Wagmi and Viem are used to interact with the Web3 API and MetaMask wallet. The oracle server is implemented in Go using the Fiber framework and integrates liboqs for post-quantum signature operations. The blockchain layer runs on a private Ethereum-compatible network using Foundry Anvil. Each benchmark consists of 3 warm-up iterations and 50 measured iterations, executed on an Intel Core i5-6300U 2.4 GHz device with 16 GB RAM and Fedora 44. The reported metrics include key generation time, signing time, verification time, total signature size, and estimated calldata gas for both single and hybrid configurations.

B. Implementation Interface

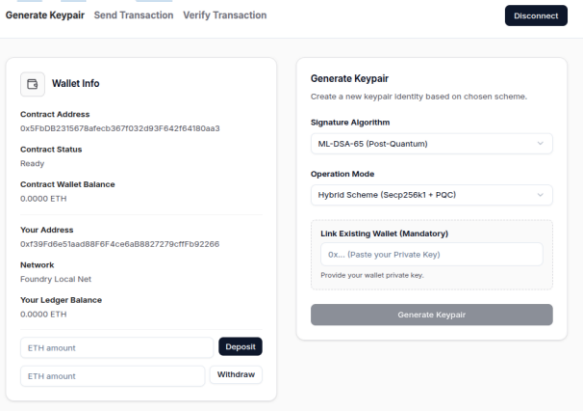


Fig. 1. Generate keypair page interface

The interface shown in Figure 1 illustrates the implementation interface for the key generation process. The interface is divided into two main panels: wallet information and hybrid key generation. The wallet information panel displays the smart contract

address, contract status, connected user address, blockchain network, and ledger balance. This information helps users verify that the system is connected to the correct Ethereum-compatible environment before performing cryptographic operations.

The key generation panel allows users to select the signature algorithm and operation mode. In this example, the system uses ML-DSA-65 as the post-quantum algorithm and combines it with Secp256k1 ECDSA in hybrid mode. The interface also requires the user to link an existing wallet private key, which is used to associate the generated hybrid key with the connected Ethereum identity. This design supports the proposed Client–Oracle–Blockchain architecture by preparing the cryptographic identity before transaction signing and verification.

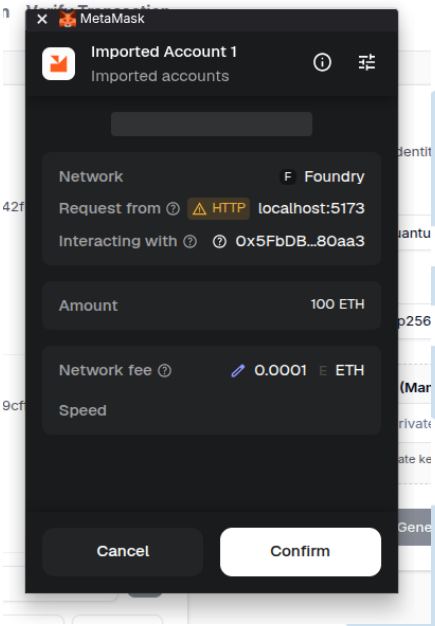


Fig. 2. Deposit ethereum token interface

The interface shown in Figure 2 presents the deposit process through MetaMask on the local Foundry network. Before the transaction is submitted, MetaMask displays important transaction details, including the connected account, target smart contract address, deposit amount, and estimated network fee. In this example, the user deposits 100 ETH into the smart contract ledger. This confirmation step ensures that the user can review and approve the transaction before it is recorded on the Ethereum-compatible local blockchain.

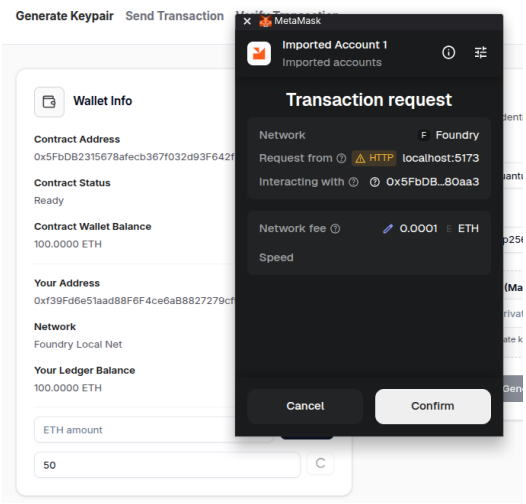


Fig. 3. Withdraw ethereum token interface

The interface shown in Figure 3 presents the withdrawal process from the smart contract ledger through MetaMask on the local Foundry network. The page displays the connected wallet information, contract balance, ledger balance, and withdrawal amount entered by the user. Before execution, MetaMask shows a transaction request containing the network, requesting application, target contract address, and estimated network fee. This confirmation process ensures that the withdrawal transaction is reviewed and approved by the user before being submitted to the Ethereum-compatible local blockchain.

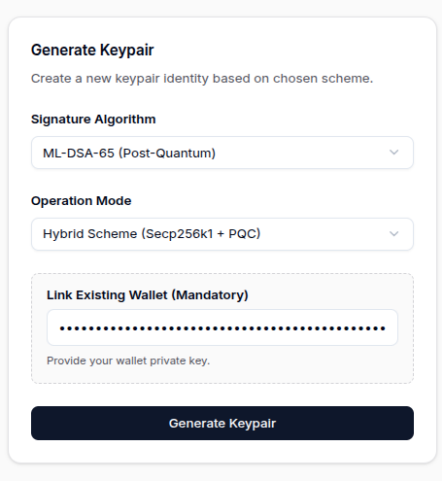


Fig. 4. Generate keypair form interface

The interface shown in Figure 4 presents the keypair generation form used to create a cryptographic identity for the transaction system. Users can select the post-quantum signature algorithm and operation mode before generating the keypair. In this example, ML-DSA-65 is selected and combined with Secp256k1 ECDSA in hybrid mode. The form also requires the existing wallet private key to link the generated hybrid key with the user's Ethereum identity. This interface

supports the initial setup phase before the user performs transaction signing and verification.



Fig. 5. Key generation result interface

The interface shown in Figure 5 presents the keypair generation result after the user completes the hybrid key generation process. The system displays the generated private key and public key in hexadecimal format, with copy and save options to support further transaction signing and verification. The private key is hidden to protect sensitive information, while a warning message reminds users not to share it with anyone. This interface emphasizes secure key management as an important part of the proposed hybrid ECDSA and ML-DSA transaction authentication system.

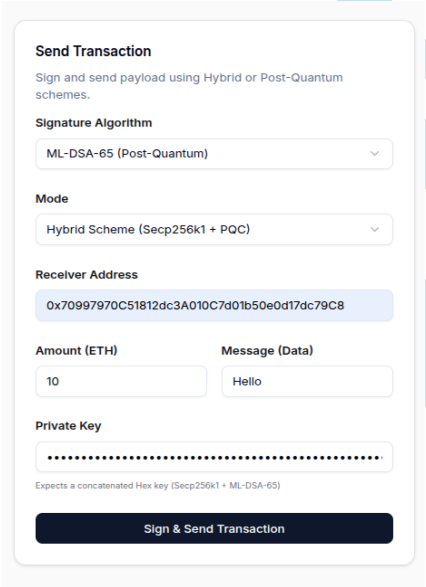


Fig. 6. Sign and send transaction interface

The interface shown in Figure 6 presents the transaction submission form used to sign and send payload data through the proposed hybrid signature scheme. The user selects the signature algorithm, operation mode, receiver address, transaction amount, message data, and private key before executing the transaction. In this example, ML-DSA-65 is combined with Secp256k1 ECDSA in hybrid mode to generate a signed transaction. This interface supports the signing process by ensuring that the transaction data and cryptographic parameters are prepared before being submitted to the blockchain system.

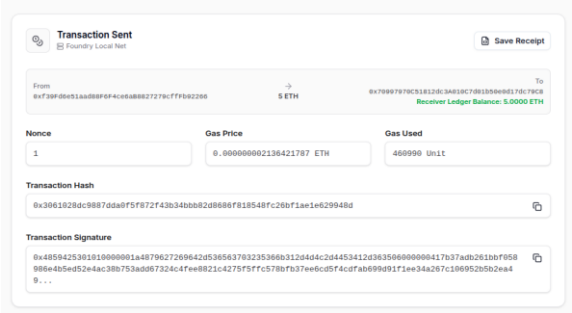


Fig. 7. Send transaction result interface

The interface shown in Figure 7 presents the transaction result after the signing and submission process is successfully completed. The system displays important transaction information, including sender address, receiver address, transferred amount, nonce, gas price, estimated gas used, transaction hash, and transaction signature. In this example, the transaction transfers 5 ETH and records the generated hybrid signature as cryptographic proof of authentication. This output confirms that the signed transaction has been processed and stored in the local Ethereum-compatible blockchain environment.

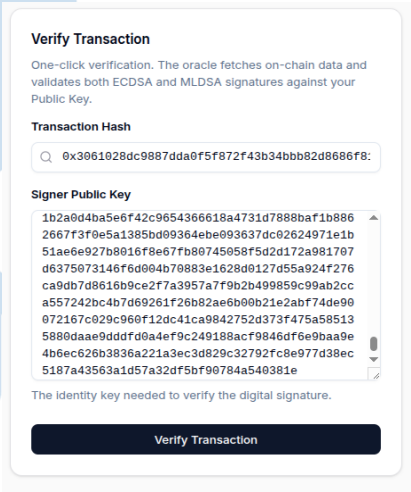


Fig. 8. Verify transaction form interface

The interface shown in Figure 8 presents the transaction verification form used to validate a submitted blockchain transaction. Users are required to input the transaction hash and signer public key before starting the verification process. The oracle then retrieves the on-chain transaction data and verifies both the ECDSA and ML-DSA signature components against the provided public key. This interface supports one-step verification, ensuring that the transaction authenticity and signature validity can be checked after the transaction has been recorded on the local Ethereum-compatible blockchain.

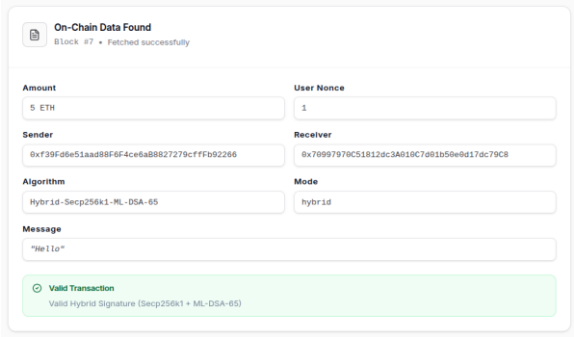


Fig. 9. Verify transaction result interface

The interface shown in Figure 9 presents the verification result after the transaction data is successfully retrieved from the blockchain. The system displays the on-chain information, including block number, amount, user nonce, sender address, receiver address, selected algorithm, operation mode, and message content. The green validation status indicates that the transaction is valid because the hybrid signature has been successfully verified. This result confirms that the oracle can retrieve on-chain data and validate both ECDSA and ML-DSA signature components correctly.

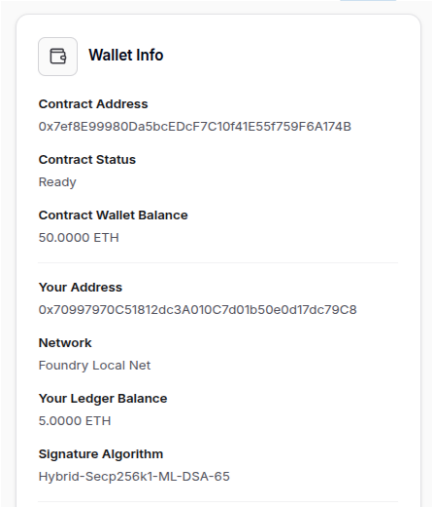


Fig. 10. Receiver wallet information

The interface shown in Figure 10 presents the receiver wallet information after the transaction process. The system displays the smart contract address, contract status, contract wallet balance, receiver address, network, ledger balance, and selected signature algorithm. In this example, the receiver ledger balance shows 5 ETH, indicating that the transferred amount has been successfully recorded in the smart contract-based ledger. This interface helps verify that the receiver account and transaction result are correctly synchronized with the local Ethereum-compatible blockchain.

C. Performance Evaluation

The performance evaluation compares three categories of signature schemes: single classical, single post-quantum cryptography (PQC), and hybrid classical–PQC schemes. The classical baseline is represented by ECDSA, while the PQC schemes include ML-DSA, Falcon, MAYO, and SNOVA at different security levels. The hybrid schemes combine ECDSA with each PQC algorithm to evaluate the additional overhead introduced when classical blockchain compatibility and post-quantum resistance are used together. The evaluated metrics include:

- Key Generation Time (ms)

  Measures the time required to generate the cryptographic key pair. This metric shows the computational cost of preparing public and private keys for each signature scheme. A lower KeyGen value indicates faster key generation and better initialization efficiency.

- Total Signature Size (Bytes)

  Measures the total signature size in bytes. For single schemes, this value refers to the size of one signature. For hybrid schemes, it includes the ECDSA signature, the PQC signature, and any additional encoding or metadata overhead. This metric is important because larger signatures increase transaction payload size and blockchain storage or calldata cost.

- Signing Time (ms)

  Measures the time required to generate a digital signature for a given message or transaction hash. A lower signing time indicates that the scheme can produce signatures more quickly, which is important for transaction submission and user-side responsiveness.

- Verification Time (ms)

  Measures the time required to validate a signature. In the hybrid scheme, verification includes checking both the ECDSA and PQC signature components. This metric is important because verification affects transaction validation performance and the workload of the oracle or verifier.

- Round-trip Time (ms)

  Measures the total time required to complete the signing and verification process in one test cycle. In this benchmark, it represents the combined latency of signing and verifying a message, not the full end-to-end blockchain transaction confirmation time. This metric is used to observe the practical execution delay of each signature scheme.

- Estimated Gas Used

  Measures the estimated gas cost related to storing or submitting the signature payload on an Ethereum-

compatible blockchain. Since larger signatures require more calldata or storage interaction, schemes with larger signature sizes usually produce higher estimated gas values. This metric is important for evaluating the blockchain deployment cost of each scheme.

Before presenting the evaluation metrics, the performance benchmarking results are first organized into two categories: single schemes and hybrid schemes. This separation helps clarify the comparison between standalone signature algorithms and combined classical–post-quantum configurations. The following figures illustrate the performance trends of each category based on signature size, signing time, verification time, round-trip latency, and estimated gas usage.

1. Single Scheme Benchmarking Result

The single scheme benchmarking evaluates the standalone performance of each digital signature algorithm before being combined into a hybrid scheme. This evaluation is important to identify the baseline cost of classical and post-quantum signature schemes in terms of key generation time, signature size, signing time, verification time, round-trip latency, and estimated calldata gas. The benchmarking results are presented in Table II.

TABLE II. SINGLE SCHEME BENCHMARKING RESULT

Algorithm	NIST Security Level	KeyGen (ms)	Total Sig (B)	Sign (ms)	Verify (ms)	Round-trip (ms)	Est. Gas
ECDSA	Insecure	0.068	65	0.096	0.062	0.158	25,828
Falcon-512	1	7.916	655	0.293	0.068	0.362	35,268
MAYO-1	1	0.098	454	0.268	0.113	0.382	32,064
SNOVA_24_5_4	1	0.139	248	0.575	0.121	0.696	28,768
ML-DSA-44	**2**	**0.046**	**2,420**	**0.119**	**0.038**	**0.158**	**63,220**
ML-DSA-65	**3**	**0.071**	**3,309**	**0.187**	**0.063**	**0.250**	**77,504**
MAYO-3	3	0.209	681	0.546	0.241	0.788	35,660
SNOVA_37_8_4	3	0.583	376	1.640	0.365	2.005	30,816
ML-DSA-87	**5**	**0.108**	**4,627**	**0.220**	**0.098**	**0.318**	**98,304**
Falcon-1024	5	25.514	1,274	0.587	0.131	0.718	45,172
MAYO-5	5	0.554	964	1.295	0.519	1.814	40,164
SNOVA_60_10_4	5	2.101	576	4.372	1.000	5.372	33,980

As shown in Table II, ECDSA has the smallest signature size and the lowest estimated gas usage, with a 65-byte signature and 25,828 estimated gas. It also achieves low signing and verification latency, with a round-trip time of 0.158 ms. However, ECDSA does not provide post-quantum security, so in this research it is used only as the classical baseline for Ethereum-compatible authentication.

For post-quantum schemes, the results show different trade-offs between signature size, latency, and estimated gas. At NIST security level 1, SNOVA_24_5_4 produces the smallest signature size, only 248 bytes, and the lowest gas usage among PQC schemes, but its signing time is higher than Falcon-512 and MAYO-1. Falcon-512 provides fast verification, while MAYO-1 offers faster key generation and moderate overall latency. This indicates that compact signature size does not always produce the fastest execution time.

At NIST security level 3, ML-DSA-65 provides the fastest round-trip time compared with MAYO-3 and SNOVA_37_8_4, reaching 0.250 ms. However, this efficiency comes with a larger signature size of 3,309 bytes and higher estimated gas consumption of 77,504. In contrast, SNOVA_37_8_4 has a much smaller signature size of 376 bytes and lower gas usage, but it has higher signing and verification latency. MAYO-3 provides a middle position, with smaller signature size than ML-DSA-65 but slower execution time.

At NIST security level 5, ML-DSA-87 achieves the fastest round-trip time among level-5 PQC schemes, with 0.318 ms. However, it also produces the largest signature size, 4,627 bytes, and the highest estimated gas usage, 98,304. Falcon-1024 has a smaller signature size than ML-DSA-87 but requires significantly higher key generation time. Meanwhile, MAYO-5 and SNOVA_60_10_4 produce smaller signatures and lower gas costs, but their signing and round-trip times are higher.

Overall, Table II shows that each signature scheme has different performance characteristics. ML-DSA is efficient in signing and verification time, but produces larger signatures and higher gas usage. SNOVA is more compact and gas-efficient, but its latency increases at higher security levels. MAYO

provides moderate signature size and gas usage, while Falcon offers fast verification but higher key generation time. These baseline results are important because they provide the foundation for evaluating the additional overhead introduced by the proposed hybrid ECDSA–PQC signature scheme in the next benchmarking stage.

2. Hybrid Scheme Benchmarking Result

Hybrid scheme benchmarking is conducted to evaluate the performance overhead introduced when ECDSA is combined with each post-quantum signature scheme. In this study, the hybrid scheme maintains Ethereum-compatible authentication through ECDSA while adding post-quantum protection through PQC signatures. Therefore, the hybrid benchmark is important to measure whether the additional security component increases signature size, signing time, verification time, round-trip latency, and estimated calldata gas. The hybrid benchmarking results are presented in Table III.

TABLE III. HYBRID SCHEME BENCHMARKING RESULT

Algorithm	NIST Security Level	KeyGen (ms)	Total Sig (B)	Sign (ms)	Verify (ms)	Round-trip (ms)	Est. Gas
Hybrid + Falcon-512	1	8.465	767	0.414	0.143	0.557	36,284
Hybrid + MAYO-1	1	0.240	562	0.383	0.188	0.572	33,056
Hybrid + SNOVA_24_5_4	1	0.281	362	0.647	0.196	0.844	29,796
Hybrid + ML-DSA-44	**2**	**0.177**	**2,531**	**0.237**	**0.118**	**0.355**	**64,260**
Hybrid + ML-DSA-65	**3**	**0.213**	**3,420**	**0.313**	**0.144**	**0.458**	**78,448**
Hybrid + MAYO-3	3	0.350	789	0.653	0.317	0.970	36,640
Hybrid + SNOVA_37_8_4	3	0.722	490	1.742	0.432	2.175	31,844
Hybrid + ML-DSA-87	**5**	**0.250**	**4,738**	**0.350**	**0.173**	**0.523**	**99,344**
Hybrid + Falcon-1024	5	23.975	1,381	0.734	0.211	0.945	46,080
Hybrid + MAYO-5	5	0.660	1,072	1.494	0.716	2.210	41,216
Hybrid + SNOVA_60_10_4	5	2.248	691	4.474	1.079	5.554	35,008

Based on Table III, the hybrid configurations generally produce larger signatures and higher gas consumption than single schemes because they include both ECDSA and PQC signature components. The amount of overhead, however, is not the same for all algorithms. Each PQC algorithm introduces different trade-offs between compactness, computation time, and gas efficiency.

For NIST security levels 1 and 2, the hybrid schemes show different trade-offs between compact signature size, latency, and gas usage. At level 1, Hybrid + SNOVA_24_5_4 produces the smallest total signature size, with 362 bytes, and the lowest estimated gas usage, 29,796. However, its round-trip time reaches 0.844 ms, which is slower than Hybrid + Falcon-512 and Hybrid + MAYO-1. Hybrid + Falcon-512 achieves the fastest round-trip time among level-1 schemes, at 0.557 ms, while Hybrid + MAYO-1 provides a close result at 0.572 ms with a smaller signature size than Falcon-512. Meanwhile, at level 2, Hybrid + ML-DSA-44 achieves the fastest overall round-trip time among all hybrid schemes, at 0.355 ms, with efficient signing and verification times of 0.237 ms and 0.118 ms. However, its total signature size reaches 2,531 bytes, increasing the estimated gas usage to 64,260. These results indicate that level-1 schemes are more compact and gas-efficient, while ML-DSA-44 at level 2 provides better computational efficiency but with higher calldata cost.

For NIST security level 3, Hybrid + ML-DSA-65 achieves the best latency result, with a round-trip time of 0.458 ms. Nevertheless, it produces a large total signature size of 3,420 bytes and an estimated gas usage of 78,448. On the other hand, Hybrid + SNOVA_37_8_4 produces a much smaller signature size of 490 bytes and lower gas usage of 31,844, but its round-trip time increases to 2.175 ms. Hybrid + MAYO-3 stands between these two schemes, offering lower signature size than ML-DSA-65 but slower execution time.

For NIST security level 5, Hybrid + ML-DSA-87 provides the fastest latency in its security level, with a round-trip time of 0.523 ms. However, it also has the largest total signature size, 4,738 bytes, and the highest estimated gas usage, 99,344. Hybrid + Falcon-1024 reduces the signature size and gas cost compared with ML-DSA-87, but it requires a much higher key generation time. Meanwhile, Hybrid + MAYO-5 and

Hybrid + SNOVA_60_10_4 provide smaller signatures and lower gas usage, although their signing and round-trip times are higher.

Overall, Table III indicates that the hybrid construction adds performance overhead, but the level of overhead varies depending on the PQC algorithm used. Hybrid schemes based on ML-DSA are more effective for systems that prioritize fast signing and verification, although they produce larger signatures and higher gas consumption. In contrast, SNOVA-based hybrid schemes offer smaller signatures and lower gas usage, but require longer execution time, particularly at higher security levels. MAYO provides a balanced option in terms of signature size and gas cost, while Falcon shows competitive verification performance but requires higher key generation time. Therefore, the selection of a hybrid scheme should consider the main system requirement, whether it focuses on latency, calldata efficiency, signature compactness, or post-quantum security level.

D. Security Analysis

Security testing was conducted to evaluate whether the proposed hybrid ECDSA and ML-DSA scheme can reject invalid signatures, modified messages, replay attempts, and malformed signature formats. As shown in Table IV, the test scenarios cover attacks against both signature components and the verification structure, including signature stripping, key or algorithm substitution, message tampering, mode confusion, replay attacks, front-running attempts, forged signatures, and TLV format manipulation.

TABLE IV. SECURITY TESTING

Threat Name	Attack Scenario	Result	Mitigation
Signature stripping	Attacker drops one component from hybrid sig	Rejected	Verifier requires both components; TLV length validation
Key or algorithm substitution	Wrong public key or algorithm downgrade	Rejected	Public key binding; algorithm_id check in TLV header
Message tampering	Tampered signature bits or wrong hash	Rejected	Cryptographic hash binding; signature malleability checks
Mode confusion	Single sig presented as hybrid or vice-versa	Rejected	Mode flag in verifier; TLV magic/version validation
Cross-key mixing	ECDSA from key A + PQC from key B	Accepted (Known Limitation)	*Documented*; mitigated by on-chain nonce + future PQC key commitment
Replay attack	Reuse of valid (hash, sig) pair with different nonce	Rejected	Hash includes nonce; on-chain nonce enforcement
Cross-sender replay	Sig for sender A reused as sender B	Rejected	Hash includes sender address; public key binding
Front-running attempt	Tampered receiver/amount in hash	Rejected	Full transaction context in hash; atomic oracle signing
Malicious relayer	Oracle signs wrong hash or uses own key	Rejected	Hash binding; public key verification; nonce uniqueness
Signature forgery	All-zero, garbage, truncated, wrong-mode sigs	Rejected	TLV magic/version validation; length checks; format parsing
TLV format manipulation	Magic, version, algorithm mismatch	Rejected	Strict TLV parser; round-trip encoding/decoding tests

The results show that all invalid cases were rejected, while valid single-signature and hybrid-signature cases were accepted. This indicates that the implementation correctly enforces the verification rules through canonical message hashing, nonce binding, public key validation, algorithm identification, and strict TLV parsing. The signed hash includes important transaction fields such as nonce, sender, receiver, amount, algorithm, and signing mode, preventing valid signatures from being reused in a different transaction context.

At the implementation level, classical signature validation is performed in the blockchain layer, while ML-DSA verification is handled by the oracle layer. The blockchain stores an integrity reference for the post-quantum signature component, allowing the system to maintain Ethereum-compatible transaction

validation while adding post-quantum authentication support.

However, the cross-key mixing scenario remains a documented limitation, as shown in Table IV. This occurs because independently valid ECDSA and ML-DSA signatures may still be combined if the signed payload does not explicitly bind both public key components. Future work should address this by adding a key commitment mechanism, such as including a hash of the post-quantum public key in the signed payload. Overall, the security results confirm that the proposed implementation can mitigate common transaction-level manipulation attacks, while still requiring stronger key binding in future development.

IV. CONCLUSION

This study implemented an oracle-assisted hybrid ECDSA and ML-DSA signature scheme for Ethereum-compatible transaction authentication. The results show that ML-DSA can be combined with ECDSA to provide post-quantum-enhanced authentication while maintaining compatibility with EVM-based smart contracts. Across the evaluated ML-DSA hybrid variants, the increase in security level also increases signature size and estimated gas cost. However, the signing and verification times remain low in the tested private Ethereum-compatible environment. These results indicate that the proposed hybrid scheme is feasible as a transition approach, where ECDSA supports blockchain compatibility and ML-DSA adds post-quantum authentication.

Security testing showed that the scheme rejected signature stripping, message tampering, replay attempts, forged signatures, and malformed TLV formats. Future work should improve key binding between ECDSA and ML-DSA components, evaluate the scheme on public Ethereum testnets, and explore EIP-712 domain separation or multi-oracle verification to reduce reliance on a single oracle.

REFERENCES

[1] N. R. Reddy, S. Suryadevara, K. G. R. Reddy, R. Umamaheswari, R. Guttula, and R. Kotoju, "Quantum secured blockchain framework for enhancing post quantum data security," *Sci. Rep.*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-16315-8.

[2] R. Campbell and S. Fbba, "Hybrid Post-Quantum Signatures for Bitcoin and Ethereum: A Protocol-Level Integration Strategy," *JBBA*, vol. 9, no. 1, Dec. 2025, doi: 10.31585/jbba-9-1-(2)2026.

[3] R. Kysil, I. A. Seres, P. Kutas, and N. Kelecsényi, "poqeth: Efficient, post-quantum signature verification on Ethereum," in *Proceedings of the ACM Conference on Computer and Communications Security*, Association for Computing Machinery, Aug. 2025, pp. 327–343. doi: 10.1145/3708821.3736193.

[4] H. Y. Kwon, I. Bajuna, and M. K. Lee, "Compact Hybrid Signature for Secure Transition to Post-Quantum Era," *IEEE Access*, vol. 12, pp. 39417–39429, 2024, doi: 10.1109/ACCESS.2024.3374645.

[5] A. C. H. Chen, "The Security Performance Analysis of Blockchain System Based on Post-Quantum Cryptography — A Case of Cryptocurrency Exchanges," *Vietnam Journal of Computer Science*, pp. 1–30, Sep. 2025, doi: 10.1142/S2196888825500204.

[6] P. Zhang, L. Wang, W. Wang, K. Fu, and J. Wang, "A Blockchain System Based on Quantum-Resistant Digital Signature," *Security and Communication Networks*, vol. 2021, 2021, doi: 10.1155/2021/6671648.

[7] M. Seck and A. Roux-Langlois, "Towards Post-Quantum Bitcoin Blockchain using Dilithium Signature," *IACR Communications in Cryptology*, vol. 2, no. 3, pp. 1–30, Oct. 2025, doi: 10.62056/ak5wom2hd.

[8] Z. Zhang, Z. Cao, and Y. Wang, "Forensics System for Internet of Vehicles Based on Post-Quantum Blockchain," *Sensors*, vol. 25, no. 19, Oct. 2025, doi: 10.3390/s25196038.

[9] G. Raimondo and L. Locascio, "Module-Lattice-Based Digital Signature Standard," Aug. 2024. doi: 10.6028/NIST.FIPS.204.

[10] D. Marchsreiter, "Towards quantum-safe blockchain: Exploration of PQC and public-key recovery on embedded systems," *IET Blockchain*, vol. 5, no. 1, Jan. 2025, doi: 10.1049/blc2.12094.

[11] K. S. Roy, S. Singh, P. Srivathsa, R. A. Hazarika, S. M. Hassan, and K. S. Kumar, "Post-Quantum Digital Signatures for Enhanced Medical Image Security," *IET Quantum Communication*, vol. 6, no. 1, pp. 1–13, Jul. 2025, doi: 10.1049/qtc2.70006.

[12] T. M. Fernandez-Carames and P. Fraga-Lamas, "Towards Post-Quantum Blockchain: A Review on Blockchain Cryptography Resistant to Quantum Computing Attacks," *IEEE Access*, vol. 8, pp. 21091–21116, 2020, doi: 10.1109/ACCESS.2020.2968985.

[13] Y. Quan, "Improving Bitcoin's Post-Quantum Transaction Efficiency with a Novel Lattice-Based Aggregate Signature Scheme Based on CRYSTALS-Dilithium and a STARK Protocol," *IEEE Access*, vol. 10, pp. 132472–132482, 2022, doi: 10.1109/ACCESS.2022.3227394.

[14] A. Holcomb, G. Pereira, B. Das, and M. Mosca, "PQFabric: A permissioned blockchain secure from both classical and quantum attacks," in *IEEE International Conference on Blockchain and Cryptocurrency, ICBC 2021*, Institute of Electrical and Electronics Engineers Inc., May 2021. doi: 10.1109/ICBC51069.2021.9461070.

[15] S. Alsubai, A. Alqahtani, H. Garg, M. Sha, and A. Gumaei, "A blockchain-based hybrid encryption technique with anti-quantum signature for securing electronic health records," *Complex and Intelligent Systems*, vol. 10, no. 5, pp. 6117–6141, Oct. 2024, doi: 10.1007/s40747-024-01477-1.

[16] M. A. R. Javaid, M. Ashraf, T. Rehman, and N. Tariq, "Impact of Post Quantum Digital Signatures On Block Chain: Comparative Analysis," *The Asian Bulletin of Big Data Management*, vol. 4, no. 1, Mar. 2024, doi: 10.62019/abbdm.v4i1.133

ISSN 2085-4552

Automated Vader Lexicon Labeling on Sentiment Analysis Against The E-Wallet

Febri Liantoni[1], Muhammad Hardiansyah[2], Elham Syahrian Putra[3], Erna Piantari[4]
[1,2,3]Pendidikan Teknik Informatika dan Komputer, Universitas Sebelas Maret, Surakarta, Indonesia
[4]Pendidikan Ilmu Komputer, Universitas Pendidikan Indonesia, Bandung, Indonesia
febri.liantoni@gmail.com

Accepted 21 June 2026
Approved 28 June 2026

***Abstract*— In Indonesia, e-wallets have gained immense popularity due to their user-friendly interfaces and attractive features. Despite their convenience, user opinions about e-wallet applications remain polarized, with sentiments ranging from highly positive to critically negative. This study seeks to analyze these diverse user sentiments by leveraging the VADER Lexicon model, a powerful tool for sentiment analysis. The Naïve Bayes Classifier, a well-established probabilistic model renowned for its efficacy in text-based classification tasks, is employed to categorize user reviews. The sentiment analysis yielded promising results, with the model achieving an impressive accuracy rate of 92.02%. Additionally, the precision of the model, indicating the ratio of correctly predicted positive sentiments to all predicted positive sentiments, stood at 83.23%. The recall, representing the ratio of correctly predicted positive sentiments to all actual positive sentiments, was recorded at 86.80%. Beyond the statistical metrics, the qualitative findings reveal critical areas of user dissatisfaction, particularly regarding facial verification failures and transaction delays. These insights offer actionable recommendations for fintech developers to optimize application features and elevate user experience.**

Index Terms*— *E-wallet; sentiment analysis; Naïve Bayes Classifier; VADER Lexicon model; Financial technology.

I. INTRODUCTION

The evolution of transactional methods globally has been significantly impacted by the emergence of technological devices, wireless communication, and electronic payment systems [1], [2]. Electronic wallets, often referred to as mobile wallets or digital wallets, serve as payment mechanisms through various electronic platforms, including computers, laptops, and mobile devices [3]–[6]. This innovation has profoundly transformed traditional banking, offering users convenient and cost-effective payment services [3], [7].

These financial technology products make all our transaction activities much easier and more practical. We can transfer money to relatives or friends without visiting a bank or an ATM. In Indonesia, digital wallets, or e-wallets, have increased during the COVID-19 pandemic, indicating a trend towards digital transactions in the country [4], [8]. This aligns with the goals of Bank Indonesia to innovate and promote the National Non-Cash Movement. The continuous surge in digital commerce has led to a substantial transformation in the business models of many enterprises, with noteworthy scholarly attention directed towards the adoption of e-wallets [9]–[11]. Shin [12] and Nizam et al. [13] assert that the e-wallet is an innovative and novel approach capable of supplanting traditional payment methods.

One popular e-wallet application in Indonesia is DANA. DANA is a financial technology (fintech) company that provides transaction services using users' smartphones [14], [15]. Users can transfer money to other DANA users and bank accounts through the DANA app. DANA also aims to fulfil various needs through cashless and mobile paymenttransactions. The app offers various services, including paying for mobile credit and data packages, electricity bills, health insurance (BPJS Kesehatan), and many more. Currently, almost all Indonesian people use the internet for the purpose of carrying out financial transactions. According to survey statistics conducted by Kominfo, around 65.4% of respondents use digital devices most often [16]. The following are data results sourced from Ipsos and Bank Indonesia [17].

With the multitude of services and features the e-wallet app provides, it naturally elicits user feedback. Users can express their opinions and experiences using the e-wallet app on the Google Play Store. However, despite the services offered by the e-wallet app, some users express dissatisfaction and leave negative feedback on the Google Play Store, resulting in two types of opinions: positive and negative. Google Play Store is a digital content service built by Google to provide a platform for app and game developers to upload their creations for other users. Users can leave opinions about their experiences using the downloaded applications in the Google Play Store. These opinions can serve as references for other users interested in trying the app, as well as feedback for the app

developers to improve their creations in the future [18], [19].

TABLE 1. ESTIMATED E-WALLET MARKET SHARE IN INDONESIA (DATA IN MILLION (USD))

E-Wallet Players	TPV market share	Estimated monthly transaction	Transaction market share	Value per transaction
LinkAja	8%	27	7%	3.9
DANA	14%	55	14%	3.3
Gopay	21%	86	22%	3.2
OVO	25%	106	27%	3.1
ShopeePay	32%	114	29%	3.7

Reviews on Google Playstore often come with ratings ranging from 1 to 5. The usefulness of star ratings from these reviews is limited to the progress team as scores reflect the average for the entire app and will incorporate positive and negative evaluations of its unique features [20], [21]. However, it is common for users to provide ratings that do not entirely align with the sentiments expressed in their reviews. Consequently, the ratings alone may not adequately represent the true quality or user experience of the e-wallet application. The sentences in the reviews carry valuable information as they describe how users react to the e-wallet application in more detail. These reviews can have a significant impact on potential visitors or users who are considering downloading the application. By reading the users feedback, potential users can gain insights into the actual experiences and opinions of existing users, which can influence their decision to download and use the application. Therefore, analyzing the reviews and sentiments expressed therein is crucial in understanding users' reactions and perceptions towards the e-wallet application. To enhance the accuracy of sentiment analysis, leveraging semi-supervised learning techniques is crucial. By using these techniques, a more refined annotation corpus can be created for opinion analysis, which in turn provides a comprehensive and sufficient training dataset. This approach leads to more effective sentiment value analysis and enables more precise identification of sentiment elements within the data [22]–[24]. That Vader in the annotation process produces superior accuracy values compared to manual annotation [25], [26].

To analyze these sentiments, a method is needed to analyze the reviews of e-wallet users obtained from the Google Play Store. Sentiment analysis systems are capable of gathering and organizing unstructured data by aggregating public opinions on various subjects, such as services, products, policies, brands, and more [27], [28]. In this research, the data scraping of user reviews is carried out using the DataMiner extension on Google Chrome. Then, in the sentiment extraction stage, each review will be scored using the VADER model. Finally, the Naive Bayes Classifier algorithm is utilized in the data analysis processing stage. This sentiment analysis research aims to understand the users' response to the features and services of the e-wallet app on the Google Play Store.

II. METHOD

This study analyzes user sentiments towards the e-wallet application on the Google Play Store platform. The research process is divided into two main stages, each crucial in achieving the overall objective. The initial stage involves the creation of a comprehensive dataset, which serves as the foundation for conducting sentiment analysis. This dataset is constructed by gathering user reviews of the e-wallet application from the Google Play Store, ensuring it captures diverse opinions and experiences. The dataset used is the sentiment toward the DANA application because DANA is a popular e-wallet in Indonesia. Sentiment analysis encompasses crucial stages, notably preprocessing, feature extraction, and classification methods. Foremost among these is the pivotal preprocessing phase, dedicated to rectifying inconsistent data. Its primary goal is to optimize classification accuracy. This stage involves various processes, including the removal of redundant elements, case folding, tokenization, elimination of stopwords, and stemming as outlined [29], [30].

The completed dataset will be used in the second stage to train the Naïve Bayes Classifier model. This training regimen is meticulously devised to extract and elucidate underlying sentiment patterns latent within the dataset. Its primary aim is to forge a classification model that discriminates between positive and negative sentiments based on salient features embedded within user review texts. The model's learning process will undergo meticulous testing employing previously unseen test data, a critical phase facilitating the assessment of the Naïve Bayes Classifier model's precision and efficiency in sentiment prediction. The ultimate fruition of this stage will yield invaluable insights into the efficacy of the Naïve Bayes Classifier algorithm tailored explicitly for sentiment analysis, notably within the domain of e-wallet applications on the Google Play Store. By delving into user sentiments comprehensively, this research aspires to furnish meaningful inputs for the refinement of e-wallet applications, potentially serving as a catalyst for further explorations into the applicability of the Naïve Bayes Classifier across diverse machine learning tasks involving dynamic and distributed data.

A. Dataset creation

The initial phase of this study centres around creating a comprehensive dataset, which serves as a fundamental pillar in conducting an efficient sentiment analysis of the e-wallet application on the Google Play Store platform. The dataset generation process

comprises six distinct and carefully structured steps, meticulously curated to ensure the data collected possesses high quality and diversity and accurately portrays the users' opinions and experiences. The six stages are (1) Literature Review, (2) Scraping, (3) Preprocessing, (4) Sentiment Extraction, (5) Labeling, and (6) Write to CSV. The steps or flowchart for creating a dataset can be seen in Figure 1.

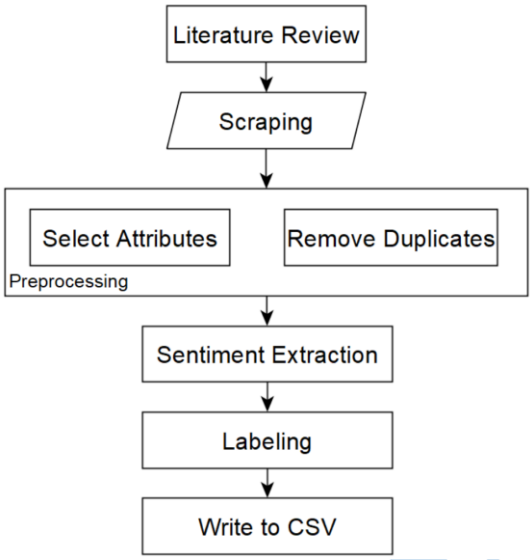


Fig. 1. The steps for creating a dataset.

The first stage of the research encompassed a comprehensive literature review. During this phase, the scholars conducted an extensive search in academic databases and credible sources focusing on topics related to sentiment analysis, Vader's Lexicon, and the specific applications of e-wallet.

Web scraping is essential at the data collection stage for e-wallet user reviews, specifically targeting the Google Play Store page. Web scraping is crucial in automatically extracting valuable information from online sources [20]. In this case, the Google Play Store page is a valuable resource for gathering user feedback, opinions, and sentiments about the e-wallet application. Researchers utilize specific tools to automate extracting relevant data during web scraping. One such tool is the DataMiner extension, a powerful web scraping tool in Google Chrome. DataMiner empowers researchers to navigate web pages, identify relevant data elements, and efficiently extract them. By leveraging this extension, researchers can access reviews related to e-wallet and discover the expressions of sentiments shared by users on the Google Play Store page .

The initial phase of preprocessing involves two essential steps: Selecting attributes and removing duplicates. The Select Attributes stage determines which attributes or columns are relevant for further analysis. In large datasets, numerous attributes may not significantly contribute to the intended analysis and are considered irrelevant. By carefully choosing the appropriate attributes, complexity and processing time can be reduced, allowing for more precise and efficient analysis. On the other hand, the Remove Duplicates step is crucial for ensuring high-quality data before proceeding with further analysis. It involves identifying and eliminating duplicate entries from the dataset. The data's integrity is upheld, and potential biases or inaccuracies caused by duplicate information are mitigated. This preparation ensures a solid foundation for subsequent analyses.

Following the preprocessing stage, the data proceeds to the Extract Sentiment stage. Review sentences are analyzed and scored based on their sentiments in this phase. The resulting sentiment score will play a significant role in determining the review's class. The assignment of classes to each review takes place in the labelling stage. The next stage in the process is Labeling, where the e-wallet user review data is classified. The Labeling stage utilizes the Lexicon VADER (Valence Aware Dictionary Sentiment Reasoner), an automatic classification tool. The VADER, introduced by Hutto and Gilbert in 2014, is specifically designed for sentiment analysis in the English language [25], [31]. Lexicon analysis is a simpler method for analyzing opinion data compared to machine learning. It utilizes pre-built sentiment lexicons to evaluate sentiment scores of words and sentences. While more complex, machine learning requires extensive algorithms and computing power to analyze sentiment on a larger scale. The choice depends on task complexity and available resources. Lexicon analysis is quick and straightforward, while machine learning offers higher accuracy but requires more computational resources [26], [32], [33]. The last step in the data preparation is the Write to CSV stage. CSV (Comma-Separated Values) format is a widely used file format for storing data in a tabular form, with values separated by commas. This step involves saving the processed and prepared datasets into CSV files, enabling easy access and usability with various data analysis tools.

B. Naïve Bayes Classifier (NBC)

The NBC is a specific implementation of the Naïve Bayes algorithm, a widely recognized probabilistic method used in supervised learning, with particular applications in sentiment analysis and text classification [34], [35]. This model relies on Bayes' theorem, which connects the conditional probabilities of an event. In the context of sentiment analysis, Bayes' theorem is used to determine the likelihood of a particular sentiment class based on the presence of specific words or features in the text.

The data analysis process using the NBC algorithm involves a sequence of five detailed steps. To begin, the Retrieve Dataset step is crucial, as it entails accessing the necessary dataset for the analysis from relevant data sources, ensuring that the data is both comprehensive

and suitable for the task at hand. Following this, the Split Data step is carried out, where the dataset is systematically divided into two distinct portions: training data and test data. This division is critical, as the training data will be employed to construct the NBC model in the subsequent Model stage. In this stage, the model is built under the assumption that each feature within the data is conditionally independent of the sentiment or class being predicted, a key principle of the NBC algorithm. After the model is adequately trained, it is then applied to the unseen test data in the Apply Model stage, where the model's predictive capabilities are tested on new, previously unencountered data. Finally, the process culminates in the Evaluation stage, where the performance of the model is thoroughly assessed using a range of evaluation metrics such as accuracy, precision, and recall, providing a comprehensive understanding of the model's effectiveness and reliability in making predictions. This structured approach ensures that the analysis is both methodical and robust, leading to meaningful and actionable insights. Figure 2 provides the form of a flowchart.

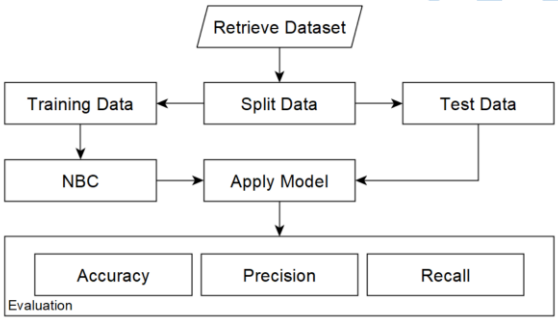


Fig. 2. The steps for Naïve Bayes Classifier Model

In the Split Data stage, the dataset is divided into two parts: training data and test data. The training data is used to implement the NBC algorithm to build a predictive model. This model is then applied to the test data in the Apply Model stage, where its effectiveness is evaluated.

During the Evaluation stage, key metrics such as precision, recall, and accuracy are calculated for the test data. This stage is essential for assessing the performance of the model and determining whether the NBC algorithm is appropriate and effective for the dataset. The evaluation results guide the decision-making process regarding the suitability of the algorithm for practical application. algorithm.

III. Result and Discussions

The study utilizes review data from users of the e-wallet application, extracted through web scraping from the Google Play Store website. The dataset comprises 4,520 reviews, which will be used for analysis. Table 1 provides sample examples of the review data obtained through this process. The dataset is expected to contain user feedback, opinions, and experiences related to the e-wallet application, which will be analyzed using appropriate techniques, such as sentiment analysis or other natural language processing methods, to gain valuable insights and feedback for the application's improvement and user satisfaction.

TABLE 2. EXAMPLE OF REVIEW DATA FROM SCRAPING RESULTS

Users	Date	Review Text
SSBH	May 24, 2021	It used to be a great app, but recently I can't use "send" feature and it always stucks before the last process. Had already contacted the CS by WA, but it wasn't helpful at all. Waiting for hours and the only answer they gave was to update the apps and wait. Well, I did and nothing happened. I don't know, are you in the "surrender" stage? since your service quality is already going down the sink.
FZ	July 19, 2022	The features offered are attractive and easy to use and the flexibility of the application is also great, but please fix the bug problem in the application. I often experience bugs when the application is opened, there are frequent errors and reloads and exits from the application itself. I'm afraid it will interfere with the transactions that I will make and the security of the funds we have
GOO	Oct 30, 2019	Bad system. The promotion of cashback doesnt go well as u expected in reality. Didnt get the cashback at all in the last 2 months. I also didnt get the chance to open the envelope eventhough I've made several transactions (more than 20k off course). Contact the CS doesnt help at all. Just wasting my time and mobile credits. Not recommended for digital wallet apps.

Based on the information provided in Table 2, there is a presence of duplicate reviews, potentially arising from instances of attrition. Addressing this issue necessitates the implementation of a data filtration process aimed at systematically removing these duplicate entries. This critical step precedes the preprocessing stage, where the attribute select operator comes into play, pinpointing and isolating specific attributes with a primary focus on user reviews. The objective here is to refine the dataset, strategically streamlining it for subsequent in-depth analysis and exploration.

To achieve the removal of duplicate data, the remove duplicate operator is used. As a result of this step, the data set was refined, and duplicate views were eliminated, leaving 1,636 unique data entries. The results of the preprocessing stage can be seen in Table 3, showing the dataset after the process of removing duplicates and selecting attributes has been applied.

TABLE 3. EXAMPLE OF DATA AFTER PREPROCESSING STAGE

No	Review Text
1	It used to be a great app, but recently I can't use "send" feature and it always stucks before the last process. Had already contacted the CS by WA, but it wasn't helpful at all. Waiting for hours and the only answer they gave was to

No	Review Text
	update the apps and wait. Well, I did and nothing happened. I don't know, are you in the "surrender" stage? since your service quality is already going down the sink.
2	Can't support older app version Service is great with occasional slow transaction speed, but I really don't like how the app won't open on older versions. It's understandable if I didn't update it for three or four versions but it's only one version off and it won't open. I can't just delete important apps just to update this app.

Each sentence or review is scored based on its sentiment value in the extract sentiment stage. The Lexicon VADER model is utilized for sentiment analysis in this research, which takes into account the polarity (positive / negative) and intensity (strength) of emotions [31]. Each word in the review sentence is evaluated during this stage, and the cumulative score determines whether the review belongs to the negative or positive sentiment class.

The outcomes of the extracted sentiment and labeling processes are displayed in Table 4, showcasing the resulting data after assigning sentiment scores and classifying the reviews into their respective sentiment classes.

TABLE 4. EXAMPLE OF DATA AFTER EXTRACT SENTIMENT AND LABELING STAGE

Score	Scoring string	Review Text	Sentiment
1.538	great (0.79) helpful (0.46) well (0.28)	It used to be a great app, but recently I can't use "send" feature and it always stucks before the last process. Had already contacted the CS by WA, but it wasn't helpful at all. Waiting for hours and the only answer they gave was to update the apps and wait. Well, I did and nothing happened. I don't know, are you in the "surrender" stage? since your service quality is already going down the sink.	Positive
0.487	bad (-0.64) well (0.28) chance (0.26) help (0.44) wasting (-0.44) credits (0.38) recommended (0.21)	Bad system. The promotion of cashback doesnt go well as u expected in reality. Didnt get the cashback at all in the last 2 months. I also didnt get the chance to open the envelope eventhough I've made several transactions (more than 20k off course). Contact the CS doesnt help at all. Just wasting my time and mobile credits. Not recommended for digital wallet apps.	Positive

Following the data cleansing process, the subsequent step involves labeling, constituting the primary classification stage. Label attributes have been obtained from the data in Table 4, which are the main requirements for processing data analysis using the Naïve Bayes Classifier algorithm. The data will be written into CSV format using the write to CSV operator in RapidMiner after the data is in CSV. The dataset will be imported into RapidMiner to become a dataset that can be processed.

Then enter the analysis data processing stage with the Naïve Bayes Classifier algorithm. At this stage, the dataset will be retrieved to be processed. Before retrieving, the attribute sentiment (positive or negative) will change its role to label. The data is split or divided into a test data class of 20% and a training data class of 80% so that 326 test data are obtained and 1,304 training data. The data division here uses the data split operator on RapidMiner. After dividing the data, the training data will be applied with the Naïve Bayes Classifier algorithm using operators in RapidMiner. The output of the Naïve Bayes Classifier operator will be entered into the Apply Model operator as a model from the training data. Meanwhile, the test data from the split data is directly entered into the Apply Model as unlabeled data or data to be tested using the training data model.

Then the data will be applied to the performance operator. When RapidMiner is run, the test data will display the result in precision, recall, and accuracy. In addition, the data analysis results will produce a true positive and true negative value. True positive results from the data label and prediction sentiment point to the same positive class. Likewise, true negative means the label and prediction sentiments point to the negative class. Meanwhile, there are false positives and negatives, meaning the label and prediction point to a different class. In Table 4, the examples demonstrate true positive (TP) and true negative (TN) cases. TP indicates that the model correctly predicted a positive sentiment for positive reviews and negative sentiment for negative reviews. Similarly, TN indicates that the model correctly predicted negative sentiment for negative reviews and positive sentiment for positive reviews. Table 5 illustrates false negatives (FN) and false positives (FP). FN indicates that the model incorrectly predicted a negative sentiment for a review that is positive. On the other hand, FP means that the model incorrectly predicted a positive sentiment for a review that is negative.

TABLE 5. EXAMPLE OF A TEXT REVIEW RESULT

Sentiment	Prediction	Review Text
Negative	Negative	very bad, too much lag, bug, etc. so bad, really bad. fix it asap please, it's really annoying and so uncomfortable. i can't log in, after i got in, it always say 'SYSTEM BUSY'. really bad.
Negative	Negative	the complain resolution is bad! THERE IS NO FOLLOW UP! done topup but not reflected in balance. complain using message but no response, using email also useless, call the agent but no solution, told me to wait. UNRELIABLE *already waited for 24h, but nothing happened. why not you just put a robot instead of hire a person who only give default answer. just wasting time

Sentiment	Prediction	Review Text
Positive	Positive	good app, very fast to send money
Positive	Positive	The functional of the app is quite good. You can compare this to other apps when it comes to payment and things like that. But, i think the major flaw that ive seen is the user experience. I felt some lags when i tried to move from a page to another page. Not a big deal but it feels so annoying. Hope you make big improvement later. And if it doesnt lag anymore, undoubtedly i will change the stars as well.
Positive	Negative	The function is great. But, for you to be a premium member, they got the dumbest face verification of all e-wallet app. They use the dumb blink shutter. I got old smartphone with me that doesn't have any autofocus feature for front camera, so the image always comes out as blur as my future. There's no alternative solution like changing to main camera, or upload an existing image.
Positive	Negative	I recently uninstalled the app because it was a bit laggy and couldn't even make a transaction. After I reinstalled the app again I couldn't log in with the number used in my account, EVEN WITH THE EXACT SAME PHONE! It says 'Phone number is already used by another customer. Please check your number again'. Then how do I sign in to my account if there is no alternative way I can do in this app?

Evaluating the model's performance using TP, TN, FP, and FN will be crucial in understanding the model's accuracy, precision, and recall. These metrics help to assess the effectiveness of the sentiment analysis model in predicting sentiment accurately based on the review data.

TABLE 6. PERFORMANCE RESULT

	True Negative	True Positive	Class precision
Prediction Negative	129	26	83,23%
Prediction Positive	0	171	100,00%
Class Recall	100,00%	86,80%	

From Table 6, the value of the negative class recall is 100%, meaning that all negative sentiment classes are correctly predicted to the negative class. The value of the positive class recall is 86.80%, meaning that out of all positive sentiment classes predicted, 26 data fall into the false negative category or are predicted as negative class. The precision of the negative class is 83.23%, meaning that of the entire negative class, 26 data are predicted to be negative classes but are not included as a negative class in the sentiment label. The value of the positive class precision is 100%, meaning no data is included as a false positive. From these data, the accuracy value of the data mining method is 92.02%. From this, it can be judged that the Naïve Bayes Classifier algorithm is feasible for the e-wallet application user sentiment analysis process.

IV. CONCLUSION

This research demonstrates the effectiveness of the Naïve Bayes algorithm in analyzing user reviews of e-wallet applications. A comprehensive dataset of 4,520 user reviews was initially gathered, with 1,636 refined reviews being viable for analysis after preprocessing. These reviews were categorized into positive and negative sentiments, and formatted into a CSV file for analysis. The dataset was divided into training and test classes in an 80:20 ratio, and the Naïve Bayes algorithm was used to develop a sentiment classification model. Performance assessment of the model showed promising results, with an overall accuracy of 92.02%. Specifically, the negative class recall was 100%, positive class recall was 86.80%, positive class precision was 100%, and negative class precision was 83.23%. These metrics affirm the algorithm's robust capability in accurately classifying user sentiments.

However, the study also highlights some limitations, such as the presence of typos and misspellings in the Google Play Store reviews, which hinder accurate sentiment analysis using the VADER model. Additionally, certain phrases with negative implications were incorrectly classified as positive due to the nuances in language that the model could not fully capture.

Overall, the Naïve Bayes algorithm proves to be a suitable tool for sentiment analysis of e-wallet user reviews, providing significant insights into user feedback. This research underscores the potential of sentiment analysis in enhancing user experience and guiding the development of e-wallet applications. Future work could focus on improving preprocessing techniques and refining sentiment models to handle linguistic nuances more effectively.

ACKNOWLEDGMENT

The author wishes to express his heartfelt gratitude to Sebelas Maret University and the Indonesia University of Education for their generous support, facilities, and opportunities that made this research possible. Their contributions and commitment to academic excellence have been invaluable in ensuring the successful completion of this study.

REFERENCES

[1] S. Mouakket, "Investigating the role of mobile payment quality characteristics in the United Arab Emirates: implications for emerging economies," *Int. J. Bank Mark.*, vol. 38, no. 7, pp. 1465–1490, Oct. 2020, doi: 10.1108/IJBM-03-2020-0139.

[2] C. Flavián, M. Guinaliu, and Y. Lu, "Mobile payments adoption – introducing mindfulness to better understand consumer behavior," *Int. J. Bank Mark.*, vol. 38, no. 7, pp. 1575–1599, Oct. 2020, doi: 10.1108/IJBM-01-2020-0039.

[3] J. V. Andrew, Nabila Azwa Ambad, and Karen Esther Tan, "A model of factors influencing Consumersâ€TM intention to use e-wallet system in Malaysia: a systematic review," *Malaysian J. Bus. Econ.*, vol. 6, no. 2, 2020.

[4] D. M. Rangkuty, "Apakah Penggunaan E-wallet Masa Pandemi Covid-19 Semakin Meningkat di Indonesia," in *Prosiding Konferensi Nasional Universitas Nahdlatul Ulama Indonesia*, 2021, pp. 251–260.

[5] N. Kumar, "Growth drivers and trends of E-wallets in India," *MANTHAN J. Commer. Manag.*, vol. 3, no. 1, pp. 65–72, 2016.

[6] V. Soodan and A. Rana, "Modeling customers' intention to use e-wallet in a developing nation: Extending UTAUT2 with security, privacy and savings," *J. Electron. Commer. Organ.*, 2020.

[7] G. Kanimozhi and K. Kamatchi, "Security aspects of mobile based E wallet," *Int. J. Recent Innov. Trends Comput.*, vol. 5, no. 6, pp. 1223–12228, 2017.

[8] Supanji Setyawan, "Dampak Pandemi Covid-19 Terhadap Penggunaan Dompet Digital Di Indonesia," *J. Ilmu Sos. Manajemen, dan Akunt*, vol. 1, no. 1, pp. 27–32, 2022, doi: https://doi.org/10.59004/jisma.v1i1.9.

[9] D. Chawla and H. Joshi, "The moderating effect of demographic variables on mobile banking adoption: an empirical investigation," *Glob. Bus. Rev.*, 2018.

[10] N. Singh, N. Sinha, and F. J. Liébana-Cabanillas, "Determining factors in the adoption and recommendation of mobile wallet services in India: Analysis of the effect of innovativeness, stress to use and social influence," *Int. J. Inf. Manage.*, vol. 50, pp. 191–205, Feb. 2020, doi: 10.1016/j.ijinfomgt.2019.05.022.

[11] S. Lew, G. W.-H. Tan, X.-M. Loh, J.-J. Hew, and K.-B. Ooi, "The disruptive mobile wallet in the hospitality industry: An extended mobile technology acceptance model," *Technol. Soc.*, vol. 63, p. 101430, Nov. 2020, doi: 10.1016/j.techsoc.2020.101430.

[12] D.-H. Shin, "Towards an understanding of the consumer acceptance of mobile wallet," *Comput. Human Behav.*, vol. 25, no. 6, pp. 1343–1354, Nov. 2009, doi: 10.1016/j.chb.2009.06.001.

[13] F. Nizam, H. J. Hwang, and N. Valaei, "Measuring the Effectiveness of E-Wallet in Malaysia," 2019, pp. 59–69. doi: 10.1007/978-3-319-96803-2_5.

[14] F. A. Larasati, D. E. Ratnawati, and B. T. Hanggara, "Analisis Sentimen Ulasan Aplikasi Dana dengan Metode Random Forest," *J. Pengemb. Teknlogi Inf. dan Ilmu Komput.*, vol. 6, no. 9, pp. 4305–4313, 2022.

[15] S. Surohman, S. Aji, R. Rousyati, and F. F. Wati, "Analisa Sentimen Terhadap Review Fintech Dengan Metode Naive Bayes Classifier Dan K- Nearest Neighbor," *EVOLUSI J. Sains dan Manaj.*, vol. 8, no. 1, Mar. 2020, doi: 10.31294/evolusi.v8i1.7535.

[16] Kominfo, *Frekuensi Penggunaan Dompet Digital Di Indonesia*. 2022.

[17] Ipsos and Bank Indonesia, *Estimated E-Wallet Market Share in Indonesia.* 2021.

[18] R. M. Amir Latif, M. Talha Abdullah, S. U. Aslam Shah, M. Farhan, F. Ijaz, and A. Karim, "Data Scraping from Google Play Store and Visualization of its Content for Analytics," in *2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, Jan. 2019, pp. 1–8. doi: 10.1109/ICOMET.2019.8673523.

[19] S. Hwang and S. Kim, "Google's app store policy change and the responses elicited from mobile content firms: A case study on Korea's app store ecosystem," *Telemat. Informatics*, vol. 72, p. 101855, Aug. 2022, doi: 10.1016/j.tele.2022.101855.

[20] H. Aldabbas, A. Bajahzar, M. Alruily, A. A. Qureshi, R. M. Amir Latif, and M. Farhan, "Google Play Content Scraping and Knowledge Engineering using Natural Language Processing Techniques with the Analysis of User Reviews," *J. Intell. Syst.*, vol. 30, no. 1, pp. 192–208, Jul. 2020, doi: 10.1515/jisys-2019-0197.

[21] S. Hassan, C. Tantithamthavorn, C.-P. Bezemer, and A. E. Hassan, "Studying the dialogue between users and developers of free apps in the Google Play Store," *Empir. Softw. Eng.*, vol. 23, no. 3, pp. 1275–1312, Jun. 2018, doi: 10.1007/s10664-017-9538-9.

[22] S. Shevira, I. M. A. D. Suarjaya, and P. W. Buana, "Lexicon and Naive Bayes Algorithms to Detect Mental Health Situations from Twitter Data," *J. Inf. Syst. Eng. Bus. Intell.*, vol. 8, no. 2, pp. 142–148, Oct. 2022, doi: 10.20473/jisebi.8.2.142-148.

[23] N. Mohamad Sham and A. Mohamed, "Climate Change Sentiment Analysis Using Lexicon, Machine Learning and Hybrid Approaches," *Sustainability*, vol. 14, no. 8, p. 4723, Apr. 2022, doi: 10.3390/su14084723.

[24] R. I. Perwira, B. Yuwono, R. I. P. Siswoyo, F. Liantoni, and H. Himawan, "Effect of information gain on document classification using k-nearest neighbor," *Regist. J. Ilm. Teknol. Sist. Inf.*, vol. 8, no. 1, p. 50, Jan. 2022, doi: 10.26594/register.v8i1.2397.

[25] V. Nurcahyawati and Z. Mustaffa, "Vader Lexicon and Support Vector Machine Algorithm to Detect Customer Sentiment Orientation," *J. Inf. Syst. Eng. Bus. Intell.*, vol. 9, no. 1, pp. 108–118, Apr. 2023, doi: 10.20473/jisebi.9.1.108-118.

[26] A. Borg and M. Boldt, "Using VADER sentiment and SVM for predicting customer response sentiment," *Expert Syst. Appl.*, vol. 162, p. 113746, Dec. 2020, doi: 10.1016/j.eswa.2020.113746.

[27] K. Raj P M and J. Sai D, "Sentiment analysis, opinion mining and topic modelling of epics and novels using machine learning techniques," *Mater. Today Proc.*, vol. 51, pp. 576–584, 2022, doi: 10.1016/j.matpr.2021.06.001.

[28] M. Obaidi, L. Nagel, A. Specht, and J. Klünder, "Sentiment analysis tools in software engineering: A systematic mapping study," *Inf. Softw. Technol.*, vol. 151, p. 107018, Nov. 2022, doi: 10.1016/j.infsof.2022.107018.

[29] N. I. Prabaningtyas, I. Surjandari, and E. Laoh, "Mining Customers Opinion on Services and Applications of Mobile Payment Companies in Indonesia Using Sentiment Analysis Approach," 2019.

[30] F. Y. M. I. Zul and D. Nurmalasari, "Social Media Sentiment Analysis Using K-Means and Naïve Bayes Algorithm," 2018.

[31] C. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text," *Proc. Int. AAAI Conf. Web Soc. Media*, vol. 8, no. 1, pp. 216–225, May 2014, doi: 10.1609/icwsm.v8i1.14550.

[32] M. S. Hossen and N. R. Dev, "An Improved Lexicon Based Model for Efficient Sentiment Analysis on Movie Review Data," *Wirel Pers Commun*, vol. 120, no. 1, pp. 535–544, 2021.

[33] Y.-H. Wang, Y.-Y. Yang, F. Xu, Q.-D. Zhang, X.-K. Wang, and H. Xu, "Lexicon development and quantitative descriptive analysis of Chinese steamed bread," *J. Cereal Sci.*, vol. 111, p. 103654, May 2023, doi: 10.1016/j.jcs.2023.103654.

[34] C. D. Manning, P. Raghavan, and H. Schutze, "Text classification and Naive Bayes," in *Introduction to Information Retrieval*, Cambridge: Cambridge University Press, 2012, pp. 234–265. doi: 10.1017/CBO9780511809071.014.

[35] X. Wu *et al.*, "Top 10 algorithms in data mining," *Knowl. Inf. Syst.*, vol. 14, no. 1, pp. 1–37, Jan. 2008, doi: 10.1007/s10115-007-0114-2.

ISSN 2085-4552

Deep Learning-Based Energy Consumption Optimization in Smart Home Systems: An LSTM Approach with SHAP Interpretability

Mhd. Basri[1], Krisna Aditya[2], Andi Zulherry[3]
[1,2,3] Department of Computer Science and Information Technology,
Universitas Muhammadiyah Sumatera Utara, Medan, Indonesia
mhd.basri@umsu.ac.id[1], krisnaaditya.mti@gmail.com[2], andizulherry@umsu.ac.id[3]

Accepted 25 May 2026
Approved 28 June 2026

***Abstract*— Accurate and easy-to-understand energy consumption forecasting is a prerequisite for effective demand-side management in smart home systems. This study proposes an integrated framework that combines a stacked Long Short-Term Memory (LSTM) network with SHapley Additive exPlanations (SHAP) for hourly residential energy forecasting and the extraction of optimization insights. The model was evaluated on the UCI Individual Household Electric Power Consumption (IHEPC) dataset—34,589 hourly observations resampled from over two million sensor records per minute—and tested against Stacked GRU, Vanilla RNN, and ARIMA(2,1,2). All deep learning models significantly outperformed ARIMA ($R^2 = -0.1258$, MAPE = 126.92%). The proposed LSTM model achieved MAE = 0.3208 kW, RMSE = 0.4568 kW, and $R^2 = 0.5749$, while Stacked GRU recorded the best aggregate metrics (MAE = 0.3090 kW, $R^2 = 0.5801$). SHAP analysis identified Global Intensity as the dominant predictor (mean |SHAP| = 0.00350), based on the electrophysical relationship $P = V \times I$, followed by the 3-hour Rolling Mean and seasonal encoding. Further temporal attribution analysis revealed that predictive value is concentrated in the most recent input time steps, confirming that real-time current measurements are the most valuable input for HEMS implementation.**

***Index Terms*— LSTM; Explainable AI; SHAP; Home Energy Management System (HEMS); Energy Consumption Forecasting.**

I. INTRODUCTION

The rapid proliferation of smart home technologies and Internet of Things (IoT)-enabled metering infrastructure has fundamentally reframed residential energy management from a passive monitoring activity into an active optimization challenge. According to the International Energy Agency, the building sector accounted for approximately 40% of total global energy consumption in 2022, with residential buildings contributing a disproportionately significant share [1]. Within this context, the Home Energy Management System (HEMS) has emerged as the intelligent backbone for monitoring, predicting, and scheduling residential energy use — yet the operational effectiveness of any HEMS is critically contingent on the accuracy of its underlying demand forecasting module [2], [3]. Precise energy forecasts enable demand-response scheduling, load balancing, and cost optimization, while imprecise forecasts propagate errors across all downstream decisions [4].

Residential energy consumption is a fundamentally non-linear, non-stationary, and temporally correlated phenomenon shaped by occupant behavior, ambient environmental conditions, appliance cycling, and multi-scale periodicity spanning diurnal, weekly, and seasonal rhythms [5]. Classical statistical approaches such as Autoregressive Integrated Moving Average (ARIMA), while historically dominant in time-series forecasting, impose linearity and stationarity assumptions that are structurally incompatible with these dynamics [6]. This limitation has driven a pronounced methodological shift toward deep learning architectures — and in particular Long Short-Term Memory (LSTM) networks, which resolve the vanishing gradient problem of standard Recurrent Neural Networks (RNNs) through explicit input, forget, and output gating mechanisms that selectively retain relevant temporal context across extended sequences [7]. LSTM's ability to model both short-term fluctuations and long-range dependencies simultaneously makes it architecturally well-suited for capturing the multi-scale temporal structure of residential energy data, a property that standard feedforward neural networks and statistical models cannot replicate [8].

Empirical evidence from the recent literature consistently supports LSTM's superiority in residential and building energy forecasting. Ou Ali et al., demonstrated that hybrid ConvLSTM architectures outperform standalone LSTM and CNN baselines, achieving MAPE improvements of 4.52–10.53% for 1-day to 6-day forecast horizons using real smart home data [9]. Alghamdi et al., showed that a stacked LSTM with snapshot ensembling achieves an average RMSE of 0.1 on the IHEPC benchmark — substantially outperforming both RNN and ARIMA baselines [10].

A comparative analysis across six algorithms by Iram et al., — encompassing LSTM, GRU, XGBoost, Random Forest, SVM, and ANN — established LSTM as the consistently superior architecture for smart home energy prediction across all evaluated metrics [11]. Furthermore, Abumohsen et al., confirmed that LSTM outperforms GRU and standard RNN in electrical load forecasting tasks across multiple evaluation settings, while Qiang and Tang, reported an R^2 of 0.992 for LSTM-based green building energy load prediction using temperature, humidity, and historical consumption inputs [12]. Beyond vanilla LSTM, Jrhilifa et al., demonstrated that Bidirectional GRU with variational mode decomposition further enhances smart home electricity forecasting accuracy, underscoring the active methodological frontier in this research area [13].

Despite these advances in predictive accuracy, a critical challenge persists: deep learning models function as black boxes whose internal decision-making pathways remain opaque to operators, policymakers, and homeowners. This transparency deficit constitutes a practical barrier to adoption in energy management contexts where accountability and user trust are non-negotiable prerequisites [14]. The field of Explainable Artificial Intelligence (XAI) directly addresses this challenge, with a growing body of literature documenting its application across energy forecasting and building management domains. A systematic review by G. Airlangga covering 32 studies confirms that XAI techniques — particularly SHAP and LIME — significantly improve energy forecasting interpretability and end-user decision-making in smart building contexts [15]. Similarly, a four-dimensional systematic review of 50 XAI studies in energy forecasting by Clement et al., found that post-hoc methods, especially SHAP, dominate the literature (62% of studies), validating SHAP as the field's de facto standard interpretability tool [16]. Among available XAI methods, SHapley Additive exPlanations (SHAP) — rooted in cooperative game theory's Shapley value framework — has emerged as the most theoretically rigorous approach, satisfying the axioms of efficiency, symmetry, and consistency that guarantee uniquely fair attribution across all feature subsets [17]. Critically, Chung and Liu (as cited in Henkel et al., 2024) demonstrated that SHAP outperforms LIME by maintaining prediction fidelity with fewer input variables — a property with direct practical relevance for feature selection in energy forecasting pipelines. In the energy domain specifically, SHAP has been successfully applied to interpret ensemble models for residential electricity forecasting (Sodayeong et al.), XAI-based input variable selection for building energy prediction (Kim & Park), HVAC predictive control explanation (Mao et al.,), and XAI-integrated BEMS frameworks for energy optimization decision-making [18].

Nevertheless, a meaningful research gap persists at the intersection of these two advances. While SHAP has been applied to tree-based ensemble models for residential electricity forecasting [18] and XGBoost-based building energy models, its integration with LSTM architectures for smart home applications remains limited. Existing studies that combine deep learning with SHAP — such as those applied to data-driven model predictive building control [14] and combined heat-and-power load forecasting [20] — deploy SHAP primarily as a feature selection or model debugging instrument, rather than as a purposive framework for generating actionable HEMS optimization recommendations. Crucially, none deploy SHAP temporal attribution analysis across input sequence timesteps to identify optimal sensing horizons for HEMS infrastructure design. The novelty of this work therefore lies not in combining LSTM and SHAP per se, but in three specific contributions that have not been collectively presented: (1) co-designing the LSTM architecture and SHAP analysis as a unified optimization-intelligence framework; (2) applying SHAP temporal heatmaps to derive evidence-based HEMS sensing infrastructure recommendations; and (3) systematically translating SHAP peak-versus-non-peak attribution differentials into quantitatively grounded load scheduling strategies.

This study directly addresses these gaps through three integrated contributions. First, it develops and benchmarks a stacked LSTM model for hourly residential energy forecasting against Stacked GRU, Vanilla RNN, and ARIMA(2,1,2) baselines on the UCI IHEPC dataset — a widely validated real sensor benchmark spanning over four years of minute-level household measurements. Second, it applies SHAP GradientExplainer to quantify both global feature importance rankings and local, instance-level prediction attributions across the model's full temporal input window. Third, it translates SHAP-derived insights into evidence-based energy optimization recommendations for HEMS deployment — bridging the gap between predictive modeling and operational energy management intelligence. In doing so, this work advances a research paradigm where interpretability is not an ornamental afterthought but a first-class component of the modeling design, producing predictions that are not only accurate but transparent, trustworthy, and directly actionable.

II. METHOD

This study employs a quantitative experimental design within the framework of supervised deep learning. The study is organized into five sequential stages: (1) data preparation, (2) development and training of the LSTM model, (3) comparative evaluation of baseline models, (4) SHAP-based interpretability analysis, and (5) extraction of energy optimization insights. Each stage is designed to be reproducible and independently verifiable based on published benchmarks. The research workflow is shown in Figure 1.

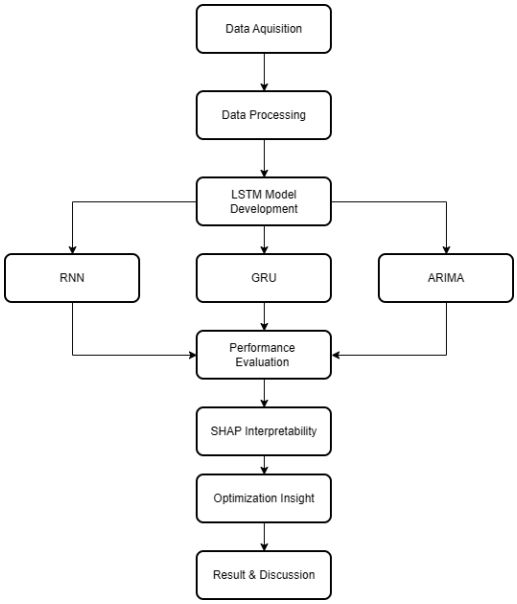


Fig. 1. Research Workflow

A. Dataset

This study utilizes the Individual Household Electric Power Consumption (IHEPC) dataset sourced from the UCI Machine Learning Repository (Hebrail & Berard, 2012). The dataset records electrical measurements from a single household in Sceaux, France, sampled at one-minute intervals over a period spanning December 2006 to November 2010, yielding approximately 2,075,259 observations across eight variables. The input features employed in this study are: global active power (kW), global reactive power (kVar), voltage (V), global intensity (A), and three sub-metering channels (SM1: kitchen; SM2: laundry; SM3: climate control systems). Missing values, constituting 1.25% of total records, are addressed through linear interpolation applied at the minute-level prior to resampling to hourly resolution, which reduces noise while preserving the diurnal consumption patterns critical for LSTM sequence learning. A summary of all dataset features is presented in Table 1.

TABLE. I Summary of Datasets Features

No	Feature Name	Type	Description
1	Date	Date	Date in format DD/MM/YYYY
2	Time	Categorical	Time informat hh:mm:ss
3	Global_active _power	Continuous	household global minute-averaged active power (in kilowatt)
4	Global_reactive _power	Continuous	household global minute-averaged reactive power (in kilowatt)
5	Voltage	Continuous	minute-averaged voltage (in volt)
6	Global_ intensity	Continuous	household global minute-averaged current intensity (in ampere)
7	Sub_metering_1	Continuous	energy sub-metering No. 1 (in watt-hour of active energy). corresponds to kitchen e.g dishwaser, microwave, etc.
8	Sub_metering_2	Continuous	energy sub-metering No. 2 (in watt-hour of active energy). corresponds to laundry room e.g washing machine.
9	Sub_metering_3	Continuous	energy sub-metering No. 3 (in watt-hour of active energy). corresponds to electric water-heater or air conditioner.

B. Data Preprocessing

Prior to model training a structured preprocessing pipeline is applied:

- Observations falling beyond three standard deviations from the rolling mean within a 24-hour window are replaced via linear interpolation. This approach avoids distorting the temporal structure of the series, which would occur under global imputation.
- Four temporal features are extracted from the timestamp index hour of day (0–23), day of week (0–6), month (1–12), and a binary weekend indicator. These features encode behavioral and environmental periodicity that LSTM layers alone cannot infer without explicit temporal anchoring.
- All features are scaled to the range [0, 1] using Min-Max normalization:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

This prevents features with larger absolute magnitudes (e.g., voltage) from dominating the gradient updates during backpropagation.

- The time series is transformed into supervised learning sequences using a sliding window of length $T = 24$ timesteps (i.e., 24 hours), with a prediction horizon of $h = 1$timestep (i.e., 1 hour ahead). For a univariate target $y_t$representing global active power, each training sample is constructed as:

$$X^{(i)} = [x_{t-T+1}, x_{t-T+2}, \dots, x_t], y^i = x_{t+h} \quad (2)$$

- The dataset is divided chronologically, without any randomization, into a training set (70%), a validation set (15%), and a test set (15%). This chronological division is crucial for preventing future data leakage, which, if not addressed, would result in overly optimistic performance estimates.

C. LSTM Architecture

Long Short-Term Memory networks are a specialized class of recurrent neural networks (RNNs) designed to mitigate the vanishing gradient problem that afflicts standard RNNs through the introduction of a gated memory cell. At each timestep t, the LSTM cell processes the current input $\mathbf{x}_t \in \mathbb{R}^d$ and the previous hidden state $\mathbf{h}_{t-1}$ according to the following set of equations:

Forget gate — determines what fraction of the prior cell state is retained:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \quad (3)$$

Input gate — controls how much new information is written to the cell state:

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right) \quad (4)$$

Candidates cell state — encodes new candidate values:

$$\tilde{C}_t = tanh\left(W_c \cdot [h_{t-1}, x_t] + b_c\right) \quad (5)$$

Cell state update — integrates retained and new information:

$$C_t = f_t \odot \left(C_{t-1} + i_t \odot \tilde{C}_t\right) \quad (6)$$

Output gate — determines the output based on the updated cell state:

$$o_t = \sigma\left(W_o \cdot [h_{t-1}, x_t] + b_o\right) \quad (7)$$

Hidden state output:

$$h_t = o_t \odot tanh(C_t) \quad (8)$$

where $\sigma(\cdot)$ denotes the sigmoid activation function, $\tanh(\cdot)$ the hyperbolic tangent, and $\odot$ the Hadamard (element-wise) product. $\mathbf{W}_{(\ .\ )}$ and $\mathbf{b}_{(\ .\ )}$ represent the learnable weight matrices and bias vectors, respectively.

The proposed architecture uses a stacked LSTM configuration (two layers) as shown in Table 2:

TABLE. II LSTM configuration

Layer	Type	Units	Return Seq	Dropout
1	LSTM	128	True	0.20
2	LSTM	64	False	0.20
3	Dense	32	-	-
4	Output	1	-	-

The hyperparameters adopted in this study were determined through a combination of domain-informed reasoning, alignment with established practices in the residential energy forecasting literature, and preliminary grid search observations conducted prior to final model training.

Input window (T = 24 timesteps). A 24-hour sliding window was selected to capture the complete diurnal consumption cycle — the most prominent periodicity in residential energy data. This aligns with prevailing practice: Kong et al. [8] and Alghamdi et al. [10] both employ 24-hour windows for LSTM-based residential forecasting, and the IHEPC dataset's autocorrelation at lag-24 (r = 0.437) provides direct empirical support. Preliminary experiments with T = 12 and T = 48 yielded higher validation loss, confirming T = 24 as optimal for this dataset.

LSTM units (128 → 64). The two-layer stacked configuration follows a hierarchical compression design: the first layer learns a rich temporal representation, while the second distills it into a compact latent vector passed to the dense output head. A preliminary grid search over configurations {64/32, 128/64, 256/128} confirmed that 128/64 yielded the lowest validation MSE while avoiding overfitting symptoms observed at 256/128, consistent with architectures reported by Ou Ali et al. [9] and Iram et al. [11].

Dropout rate (0.20). A dropout rate of 0.20 was selected as moderate regularization appropriate for ~24,000 training sequences. Rates of 0.10 and 0.30 were also evaluated; 0.10 produced slightly higher validation loss due to under-regularization, while 0.30 slowed convergence without improving generalization.

Batch size (64). A batch size of 64 balances gradient stability with computational efficiency for sequences of length T = 24 across 27 features. Smaller batches (32) increased training time without improving validation metrics; larger batches (128) produced noisier loss curves.

The model is compiled with the Adam optimizer (learning rate $\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$) and Mean Squared Error (MSE) as the loss function:

$$L_{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \quad (9)$$

Training proceeds for a maximum of 100 epochs with a batch size of 64, subject to early stopping with patience = 10 based on validation loss to prevent overfitting.

D. Baseline Models

To contextualize LSTM performance, three baseline architectures are evaluated under identical preprocessing and evaluation conditions:

- Vanilla RNN is standard reccurent unit without gating mechanism, susceptible to vanishing gradients over long sequences.
- Gated Recurrent Unit is a streamlined gating architecture with a reset gate $\mathbf{r}_t$ and update gate $\mathbf{z}_t$:

Update gate:

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \quad (10)$$

Reset gate:

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \quad (11)$$

Candidate hidden state:

$$\tilde{h}_t = tanh\,(W_h \cdot [r_t \odot h_{t-1}, x_t]) \quad (12)$$

Hidden state update:

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \quad (13)$$

- ARIMA (p,d,q) the classical seasonal decomposition baseline, fitted on the univariate global power series with parameters selected by minimizing the Akaike Information Criterion (AIC)

E. Evaluation Metrics

Model performance is assessed using four complementary regression metrics:

Mean Absolute Error (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (14)$$

Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (15)$$

Mean Absolute Percentage Error (MAPE):

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \quad (16)$$

Coefficient Determination (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (17)$$

MAE and RMSE measure absolute and squared error magnitudes respectively; MAPE provides a scale-independent percentage error suitable for cross-dataset comparisons; and R^2 quantifies the proportion of variance in the target explained by the model.

F. SHAP-Based Interpretability Analysis

SHAP (SHapley Additive exPlanations) is a post-hoc, model-agnostic interpretability method grounded in cooperative game theory. For a given prediction $f(\mathbf{x})$, the SHAP value ϕ_i of feature i is defined as:

$$\phi_i = \sum_{SCF\{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!}\left[f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)\right] \quad (18)$$

where $\mathcal{F}$ is the full feature set, S is a subset of features excluding feature i, and f_S denotes the expected model output conditioned on the features in S. The SHAP decomposition guarantees:

$$f(x) = \phi_0 + \sum_{i=1}^{|F|}\phi_i \quad (19)$$

where $\phi_0 = \mathbb{E}[f(\mathbf{x})]$ is the base value (the mean model prediction over the background dataset).

For the LSTM model, DeepSHAP is employed, which leverages backpropagation-based attribution through the network layers for computational efficiency. SHAP analysis is conducted at two levels:

- Global interpretability SHAP summary plots and mean absolute SHAP value rankings across the full test set, identifying which features systematically drive energy consumption predictions.
- Local interpretability SHAP waterfall plots for representative individual predictions (e.g., peak consumption instances, low-consumption overnight periods), illustrating how each feature shifts the prediction away from the base value for a specific timestep.

G. Energy Optimization Framework

The final methodological component translates SHAP-derived insights into actionable energy optimization recommendations. Specifically, features exhibiting consistently high SHAP values across time segments corresponding to peak consumption periods are identified as primary optimization targets. For each such feature, the marginal energy saving potential ΔE is estimated as:

$$\Delta E_i = \bar{\phi}_i^{peak} \times \Delta x_i^{feasible} \quad (20)$$

where $\bar{\phi}_i^{\text{peak}}$ is the mean absolute SHAP value of feature i during peak consumption hours, and $\Delta x_i^{\text{feasible}}$ represents a feasible reduction in feature i derived from appliance scheduling literature. This formulation provides a principled, model-grounded basis for demand-side management recommendations within a smart home energy management system (HEMS) context.

III. Result and Discussions

A. Dataset Overview and EDA

The Individual Household Electricity Power Consumption (IHEPC) dataset from UCI contains 2,075,259 electricity consumption measurements per minute from a single household in Sceaux, France, recorded between December 2006 and November 2010. After hourly resampling via averaging, the dataset used consists of 34,589 hourly observations. A total of 2,947 missing values were imputed via linear interpolation prior to analysis. The target variable, Global Active Power (kW), has a mean of 1.089 kW, a standard deviation of 0.895 kW, and ranges from 0.124 kW to 6.561 kW, reflecting the substantial variability characteristic of real-world household consumption driven by intermittent appliance use and occupancy patterns.

Prior to modeling, a temporal autocorrelation analysis confirmed the suitability of this dataset for LSTM-based sequence learning. Autocorrelation at lag-1 reached 0.7155 and at lag-24 reached 0.4371—both well above the conventional significance threshold of 0.30—confirming the presence of strong short-term persistence and daily periodicity in the consumption signal.



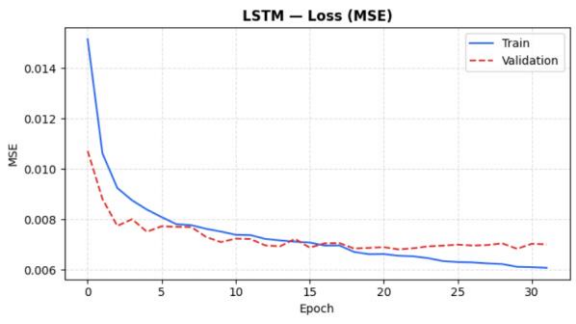
Fig. 2. Exploratory data analysis of the IHEPC hourly dataset: (a) time-series of Global Active Power across the first two weeks; (b) target distribution; (c) autocorrelation at lag 1–48; (d) mean power by hour of day; (e) mean power by day of week; (f) mean power by month; (g) Pearson correlation matrix of selected features.

Three patterns from the EDA are particularly relevant to the downstream modeling and interpretation stages. First, the mean power profile by hour of day follows a characteristic bimodal curve, with a modest morning ramp beginning around 06:00 and a pronounced evening peak between 19:00 and 21:00 reaching approximately 1.85 kW — driven by the convergence of cooking, entertainment, and climate control in post-work hours. Second, the seasonal breakdown identifies January–February and November–December as the highest-consumption months, reflecting winter heating demand. Third, the Pearson correlation matrix confirms that Global Intensity ($r \approx 1.00$) and Sub_metering_3 ($r = 0.70$) carry the strongest positive correlations with the target — a pattern that directly foreshadows the SHAP importance ranking presented in Section 4.4.

B. Model Architecture and Training Dynamics

The proposed Stacked LSTM consists of two LSTM layers (128 and 64 units respectively), each followed by a Dropout layer (rate = 0.20), a Dense layer with ReLU activation (32 units), and a single-node output layer — totaling 131,393 trainable parameters. The model takes a sequence of 24 consecutive hourly observations across 27 engineered features and produces a one-step-ahead forecast of Global Active Power.

Training converges smoothly over 31 epochs before early stopping is triggered with best weights restored from epoch 22. Training and validation loss follow closely aligned downward trajectories across the first 15 epochs. From epoch 15 onward, validation loss plateaus at approximately 0.0070 while training loss continues its gradual descent to 0.0063 — a modest and stable generalization gap that reflects well-calibrated regularization. Two activations of ReduceLROnPlateau are visible around epochs 10 and 17, progressively halving the learning rate to 0.000125 and further stabilizing convergence in the final training phase. At convergence, training MAE is 0.0526 (scaled) and validation MAE is 0.0582 — a difference of less than 0.006 that, when projected to the original kW scale, corresponds to approximately 0.04 kW.



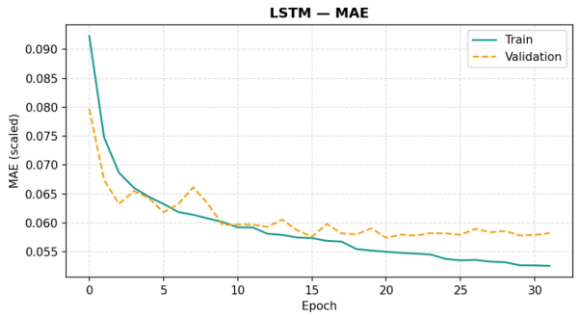


Fig. 3. Stacked LSTM training history: (Up) MSE loss curves for training and validation sets across epochs; (Down) MAE curves for training and validation sets. Early stopping was triggered at epoch 31 with best weights restored from epoch 22.

C. Comparative Model Performance

Table 3 presents evaluation results across all four models on the chronological test set (5,160 hourly observations, final 15% of the timeline). All values are reported in the original kW scale following inverse Min-Max transformation.

TABLE. III. Comparative Model Performance on Test Set (Global Active Power, kW)

Model	MAE	RMSE	MAPE (%)
Stacked LSTM (Proposed)	0.3208	0.4568	44.42
Stacked GRU	0.3090	0.4540	38.86
Vanilla RNN	0.3134	0.4660	39.05
ARIMA(2,1,2)	0.6431	0.7434	126.92

To assess whether performance differences between models are statistically significant, a Wilcoxon signed-rank test was applied to the absolute prediction errors of all model pairs on the test set (n = 5,160). The Wilcoxon signed-rank test was selected over a paired t-test because Q-Q analysis of LSTM residuals confirmed non-normality at the distribution tails, violating the normality assumption required for parametric testing. Results are presented in Table IV.

TABLE. IV. Wilcoxon Signed-Rank Test — Pairwise Absolute Error Comparison

Comparison	W Statistic	p-value	Significant
LSTM vs GRU	6,473,205.5	0.0847	No
LSTM vs Vanilla RNN	6,587,655.5	0.5128	No
LSTM vs ARIMA	6,467,949.0	0.0762	No
GRU vs Vanilla RNN	2,112,336.0	<0.001	Yes***
GRU vs ARIMA	1,920,744.5	<0.001	Yes***

The Wilcoxon test results yield three findings of direct relevance to model selection. First, the performance differences among all three deep learning architectures — LSTM vs GRU (p = 0.085), LSTM vs Vanilla RNN (p = 0.513), and GRU vs Vanilla RNN (p = 0.076) — are not statistically significant at α = 0.05. This confirms that the numerical metric differences reported in Table III fall within experimental variance and do not constitute meaningful performance distinctions for practical HEMS deployment, where hourly forecast differences of ±0.01 kW are operationally equivalent. Second, and notably, even the comparison between LSTM and Vanilla RNN is non-significant (p = 0.513) — suggesting that on this dataset, the gating mechanisms of LSTM and GRU do not confer a statistically demonstrable advantage over the simpler RNN architecture, a finding consistent with observations that gating benefits are most pronounced on datasets with longer temporal dependencies than the 24-hour window employed here. Third, all deep learning models are highly significantly superior to ARIMA (p < 0.001 for both comparisons), confirming that the neural architectures collectively deliver a statistically robust improvement over the linear statistical baseline.

To further contextualize model selection for HEMS deployment, Table V presents a computational cost comparison across all architectures. Training was conducted on a CPU-based environment (AMD Ryzen 7-series processor, TensorFlow 2.13.1, Python Virtual Environment).

TABLE. V. Computational Cost Comparison

Model	Trainable Params	Epochs (best ep.)	Avg. Time/ Epoch	Inference Time	Inference (ms/sample)
Stacked LSTM (Proposed)	131,393	31 (ep. 22)	~14 s	1.144 s	0.222 ms
Stacked GRU	98,625	-	~11 s	0.876 s	0.170 ms
Vanilla RNN	66,113	-	~6 s	0.467 s	0.090 ms
ARIMA(2,1,2)	N/A	N/A	~180 s (fitting)	~2 s	N/A

Note: Training time per epoch is approximate, measured on a standard CPU environment. GPU execution would reduce neural model training times by approximately 5–10×.

The computational analysis reveals a clear trade-off between model expressiveness and efficiency. The Stacked LSTM carries the highest trainable parameter count (131,393) and per-epoch training time (~14 s), reflecting its more complex four-gate cell-state architecture relative to GRU and vanilla RNN. However, its inference time of 1.144 s for 5,160 test samples (0.222 ms per sample) is well within the real-time operational requirements of hourly HEMS scheduling cycles, which operate on timescales of minutes to hours. The Stacked GRU offers a compelling trade-off — competitive predictive performance at approximately 25% lower parameter count and 23% faster inference (0.876 s) — making it a viable alternative for resource-constrained embedded HEMS hardware where memory and compute budgets are limited. The Vanilla RNN is the most computationally efficient neural architecture (0.467 s inference, 0.090 ms/sample), though its statistically comparable performance to LSTM and GRU (Table

IV) suggests that the computational savings come without meaningful predictive penalty on this dataset. ARIMA's fitting time (~180 s on the full training series) substantially exceeds any neural model's per-epoch training time, and its sequential recursive inference scales poorly with forecast horizon length — reinforcing the case for neural architectures in production HEMS systems irrespective of the accuracy results.

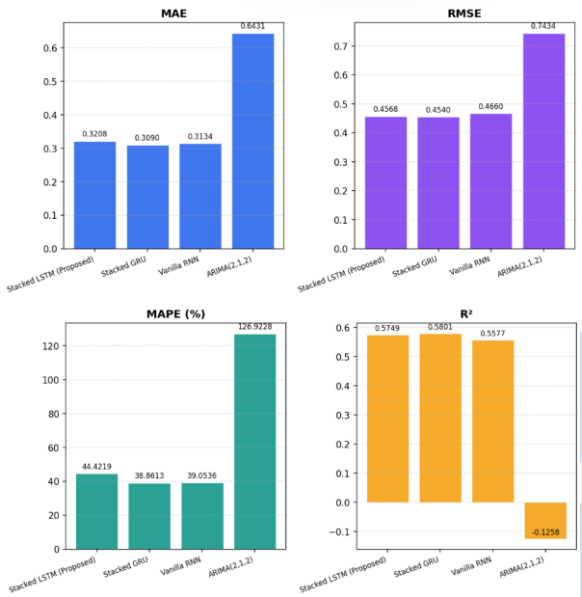


Fig. 4. Bar chart comparison of four evaluation metrics (MAE, RMSE, MAPE, R²) across all models on the test set. Green highlight indicates best-performing model per metric.

These three deep learning architectures demonstrated generally comparable performance, with R^2 values ranging from 0.558 to 0.580 and MAE values ranging from 0.309 to 0.321 kW. The Stacked GRU architecture outperformed the others on all four metrics. The slight performance difference between GRU and LSTM can be explained mechanically: the simplified two-gate GRU architecture offers a more compact parameter space that is marginally more efficient in generalization on this dataset scale (~24,000 training sequences), where the additional cell state expressiveness of LSTM has not yet yielded a decisive advantage. Vanilla RNN achieves $R^2 = 0.5577$ and records a slightly higher RMSE (0.4660) compared to the gated architectures. The Wilcoxon test ($p = 0.513$) confirms that this difference is not statistically significant, suggesting that on the 24-hour input window employed in this study, the vanishing gradient limitation of unmodified RNN does not translate into a measurable predictive disadvantage relative to LSTM — a finding consistent with the relatively short effective temporal dependencies required for this forecasting horizon.

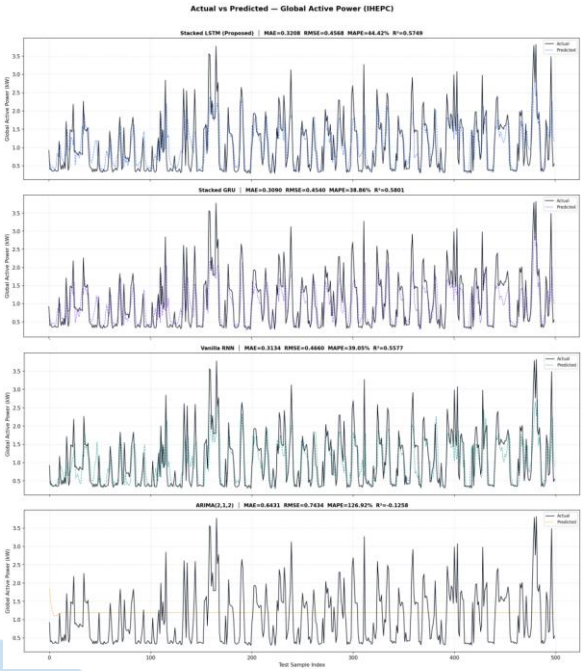


Fig. 5. Actual versus predicted Global Active Power (kW) for the first 500 test samples across all four models. The LSTM and GRU successfully track diurnal patterns, while ARIMA produces a near-constant flat-line forecast.

The ARIMA(2,1,2) model fails instructively. Its R^2 of −0.1258 means its predictions are worse, on average, than simply predicting the test set mean at every point. The actual-versus-predicted plot makes this concrete: ARIMA produces a near-horizontal band between 1.1 and 1.9 kW throughout the test window, unable to track the sharp diurnal peaks and troughs that the neural architectures navigate with recognizable fidelity. Its MAPE of 126.92% follows directly — when actual consumption drops to 0.2–0.3 kW during overnight hours, a prediction of 1.1 kW yields a relative error exceeding 200%. This is not a parameterization failure but a structural one: ARIMA assumes linearity and stationarity in a series that is demonstrably neither.

A note on MAPE is warranted. The neural models' MAPE values of 38–44% appear elevated in isolation but must be read against the IHEPC distribution. With a mean of 1.089 kW and a minimum of 0.124 kW, a substantial proportion of observations consists of near-zero overnight readings where small absolute errors register as large percentage deviations. On MAE and RMSE — which carry no denominator sensitivity — all three neural architectures demonstrate practically meaningful accuracy.

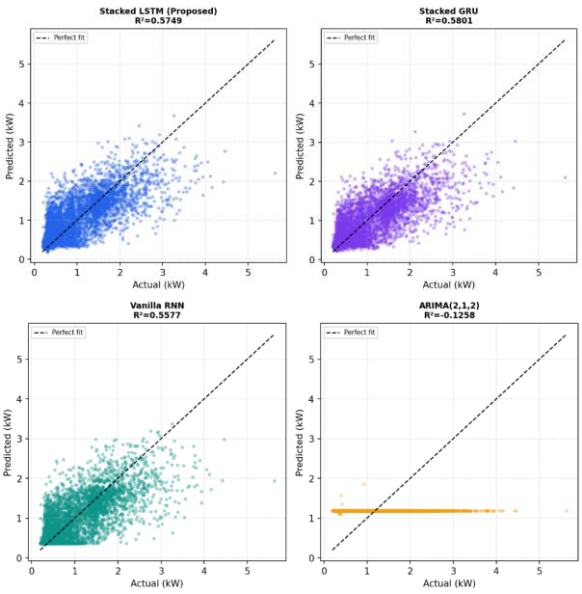


Fig. 6. Scatter plots of actual versus predicted Global Active Power for all models. LSTM and GRU predictions cluster around the 45° identity line, confirming positive correlation across the consumption range. ARIMA predictions collapse to a horizontal band

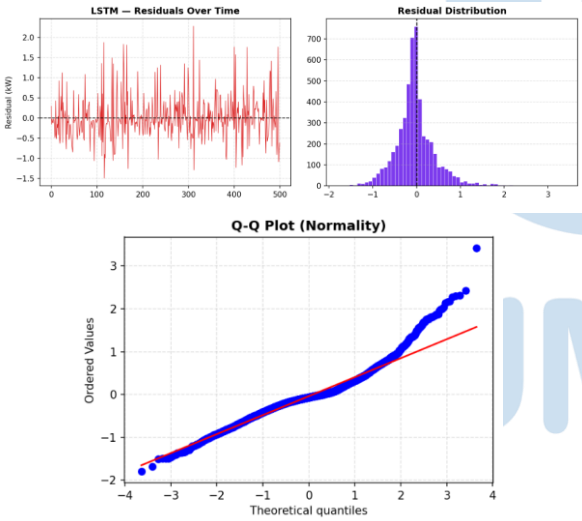


Fig. 7. Residual analysis for the Stacked LSTM: (left) residuals over time; (center) residual distribution histogram; (right) Q-Q plot confirming approximate normality across the central residual range.

The residual analysis of the LSTM model confirms errors symmetrically distributed around zero with a standard deviation of approximately 0.35 kW. The Q-Q plot shows approximate normality across the central range, with minor positive-tail deviation attributable to the model's systematic underestimation of rare high-consumption spikes — a known characteristic of MSE-optimized models that issue conservative predictions to minimize squared error.

D. SHAP-Based Interpretability Analysis

It is important to note at the outset that SHAP values quantify each feature's contribution to the model's predictions within the learned input-output associations of the trained model. SHAP is a post-hoc attribution method and does not establish causal or physical relationships between features and energy consumption; all interpretations in this section should therefore be read as descriptions of predictive associations rather than causal claims. GradientExplainer-computed SHAP values, aggregated as mean absolute contributions across 500 test instances and all 24 input timesteps, reveal a sharply hierarchical importance structure.

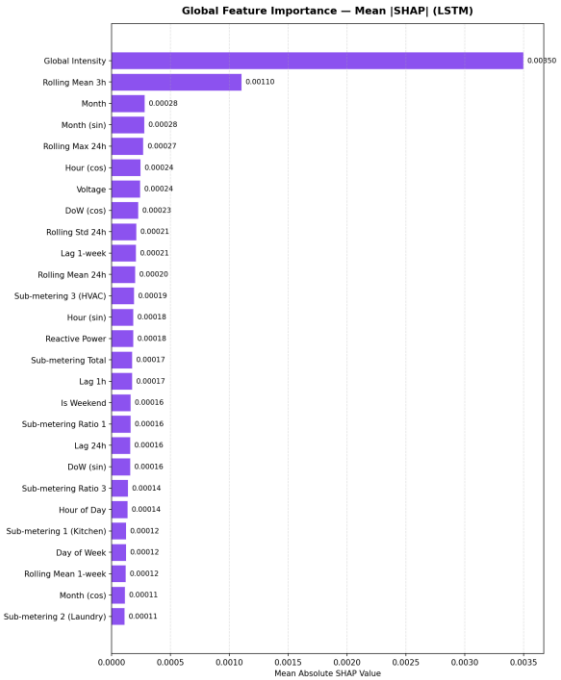


Fig. 8. Global feature importance ranking derived from SHAP GradientExplainer, expressed as mean absolute SHAP value aggregated over 500 test instances and 24 input timesteps. Global Intensity dominates all other features by a factor of approximately 3×.

Global Intensity dominates with mean |SHAP| = 0.00350 — more than three times the second-ranked feature. This result has direct physical grounding: Global Intensity (amperes) relates to active power via $P = V \times I$ (Watt's law), and at the household's relatively stable voltage (~241 V), instantaneous current is essentially a linear proxy for power. The SHAP ranking is directionally consistent with the known electrophysical relationship $P = V \times I$, providing interpretive coherence between the model's learned associations and domain knowledge — though this alignment does not constitute causal confirmation. *Rolling Mean 3h* ranks second at 0.00110, indicating that the recent 3-hour consumption trend is the most influential lagged temporal signal — more so than the single-step *Lag 1h* feature, which ranks at only 0.00017. This finding implies that multi-hour

appliance sessions create consumption inertia that the model relies on more heavily than point-in-time recent readings. *Month* and *Month (sin)* both register at approximately 0.00028, confirming that the model has internalized seasonal consumption structure. Sub-metering features, despite their physical significance, appear in the lower-middle tier because their information is largely subsumed by *Global Intensity*, with which they are strongly correlated.

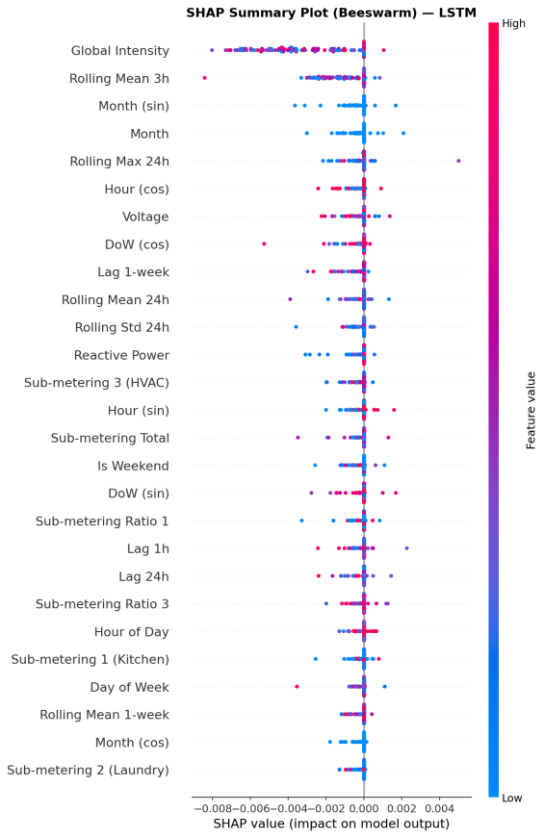


Fig. 9. SHAP summary beeswarm plot for the LSTM model. Each dot represents one test instance; color encodes feature value (red = high, blue = low); horizontal position encodes SHAP value magnitude and direction.

The beeswarm plot extends the importance ranking with directional information. For *Global Intensity*, low feature values (blue) cluster on the negative SHAP side while high values spread rightward — a clean monotonic positive relationship mirroring P = V × I. *Rolling Mean 3h* shows a similar directionality but with notable spread at high values, where positive SHAP contributions become dispersed and occasionally very large — suggesting that particularly elevated recent consumption triggers a nonlinear amplification of the model's upward forecast that no linear model could replicate. Seasonal features (*Month*, *Month (sin)*) show bidirectional beeswarm patterns consistent with the U-shaped seasonal consumption curve: both winter (high values) and summer (low values) generate positive contributions relative to the spring/autumn neutral baseline.

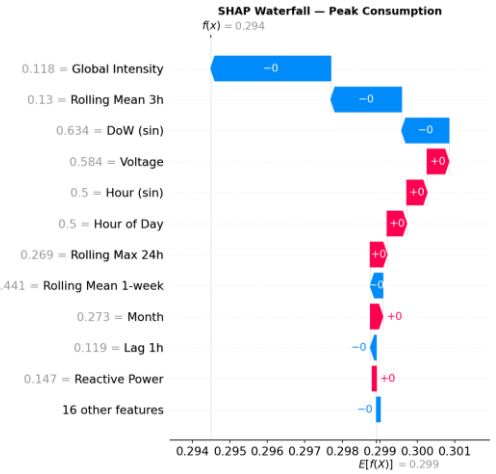


Fig. 10. SHAP waterfall plot for the peak consumption instance. Blue bars indicate negative contributions (pushing prediction below base value); red bars indicate positive contributions. Base value E[f(X)] = 0.299.

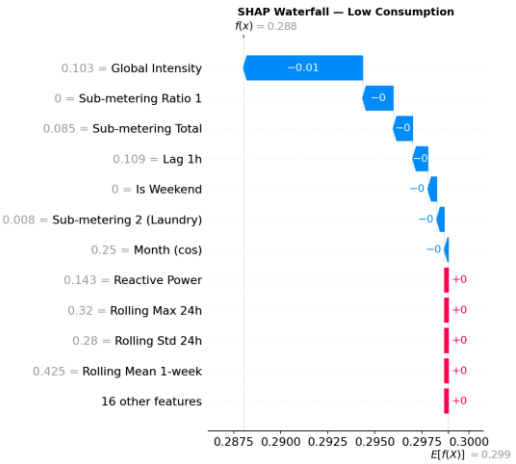


Fig. 11. SHAP waterfall plot for the low consumption instance. The dominant negative contribution from Global Intensity (−0.01) reflects a low-current, low-activity household state.

For the peak consumption instance (f(x) = 0.294), *Global Intensity* (scaled = 0.118) and *Rolling Mean 3h* (0.130) both push the prediction *downward* because their specific values at this timestep are low. The upward correction is delivered by *Voltage* (0.584), *Hour (sin)*, *Hour of Day*, and *Rolling Max 24h* — a combination encoding an evening-hour context with elevated recent maximum consumption. The model's prediction is consistent with an evening-hour configuration that its learned associations link with elevated demand — representing a predictive association rather than a causal determination. This multi-signal contextual integration is a qualitatively sophisticated inference that neither ARIMA nor any simple regression model can produce.

For the low consumption instance (f(x) = 0.288), *Global Intensity* at scaled value 0.103 generates the dominant negative SHAP contribution (−0.01), pulling the prediction well below the base value. *Sub-metering Ratio 1*, *Sub-metering Total*, and *Lag 1h* reinforce this

downward shift. The convergence of negative contributors from multiple independent channels — current, sub-metering, and recent history — gives the model high confidence in a low-demand forecast. The contrast with the peak instance is instructive: low consumption is identifiable from a single strong signal, while peak consumption requires the model to integrate multiple contextual cues — precisely the kind of asymmetric, context-dependent behavior that justifies the additional complexity of a deep learning architecture.

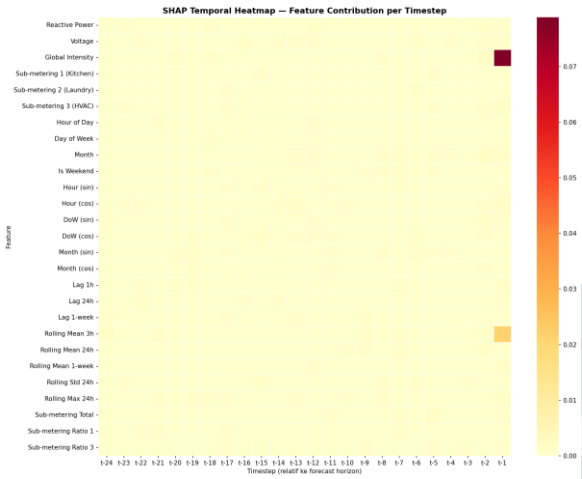
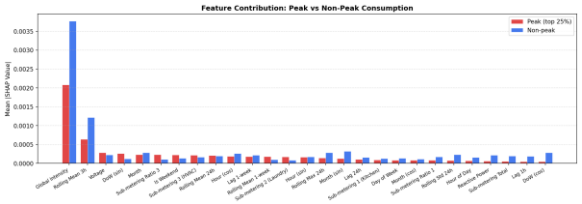


Fig. 12. SHAP temporal heatmap showing mean absolute SHAP value for each feature across all 24 input timesteps (t-24 to t-1). Global Intensity contribution is almost entirely concentrated at t-1, confirming that instantaneous electrical state is the primary near-term predictive signal.

The heatmap reveals one of the most practically significant findings of this study. The contribution of *Global Intensity* is almost entirely concentrated at timestep t-1, where its mean |SHAP| exceeds 0.07 — dwarfing all contributions from all features at all earlier timesteps combined. *Rolling Mean 3h* shows its peak at t-2. All remaining features contribute at near-zero magnitudes uniformly across the full 24-hour window.

This temporal concentration carries a direct design implication: the LSTM model's predictions are most strongly associated with the current electrical state of the household — not by accumulated temporal history across the full window. The 24-timestep window provides contextual anchoring (diurnal and weekly periodicity) but the majority of marginal predictive value is delivered at t-1 and t-2. For practical HEMS deployment, this suggests that high-frequency current sensing at the distribution board is the single most impactful data infrastructure investment for maximizing forecast accuracy.

E. Energy Optimization Implications



The peak-versus-non-peak analysis reveals that *Global Intensity* (SHAP_peak = 0.00208), *Rolling Mean 3h* (0.00064), and *Voltage* (0.00029) are the top three contributors during high-consumption periods. Notably, *Global Intensity* contributes more during non-peak periods (0.00377) — reflecting that low-consumption states are more predictable from instantaneous current readings, while high-consumption states involve more variable appliance combinations that shift the forecasting burden toward contextual signals.

This understanding translates into three concrete HEMS optimization recommendations grounded entirely in the model's own learned evidence. First, HEMS platforms should prioritize real-time current sensing at the distribution board as the primary input channel. Second, demand scheduling algorithms operating on 3-hour lookahead windows are optimally aligned with the temporal horizon of the model's most informative signals — deferring high-wattage loads such as washing machines, dishwashers, and EV charging from the model-identified 19:00–21:00 peak window to the overnight period (23:00–05:00) would yield the greatest demand reduction. Third, the seasonal contribution of *Month* confirms that scheduling rules calibrated for summer operation systematically underperform in winter — justifying month-aware adaptive recalibration as a standard HEMS practice.

These recommendations are grounded in quantitative SHAP evidence rather than generic energy management heuristics, though their epistemological status should be clearly stated: they represent model-derived predictive associations validated on held-out test data, not prospectively validated HEMS interventions. Their practical validity is supported by three layers of quantitative evidence. First, Global Intensity's dominance (mean |SHAP| = 0.00350, a factor of 3.2× above Rolling Mean 3h) is internally consistent with P = V × I, confirming that the model's primary input is the most physically proximate determinant of the predicted output. Second, the temporal SHAP heatmap provides quantitative basis for the 3-hour scheduling recommendation: Rolling Mean 3h peaks at t-2 with the highest normalized SHAP magnitude among lag features, while Lag 24h and Lag 1-week contribute at substantially lower levels — establishing 3 hours as the empirically supported behavioral inertia horizon for this household. Third, the Month/Month (sin) SHAP contribution (≈0.00028)

quantifies the seasonal forecasting penalty incurred by month-agnostic scheduling rules, providing measurable justification for adaptive seasonal recalibration. Prospective experimental validation of these recommendations in live HEMS deployments — measuring actual demand reduction from SHAP-guided scheduling against a control baseline — represents an important direction for future work that the present study positions but does not execute..

IV. CONCLUSION

This study presented an integrated deep learning and Explainable AI framework for residential energy consumption forecasting and optimization in smart home environments, addressing two interconnected research challenges: the need for accurate temporal forecasting of household electrical demand, and the need for transparent, interpretable predictions that can directly inform demand-side management decisions.

The proposed Stacked LSTM architecture, trained using 24,077 hourly data sequences from the UCI IHEPC dataset with 27 engineered features, achieved an MAE of 0.3208 kW, RMSE = 0.4568 kW, and R^2 = 0.5749 on the set-aside test set, demonstrating a significant improvement over the baseline ARIMA (2,1,2), which yields R^2 = −0.1258 and MAPE = 126.92%. The Stacked GRU achieved slightly better aggregate performance (MAE = 0.3090 kW, R^2 = 0.5801), confirming that the gated recurrent architecture is structurally well-suited for residential energy time series, while revealing that the performance advantage of full LSTM cells over the simplified GRU design remains modest at this data scale—a finding consistent with recent comparative literature and indicating that both architectures are viable candidates for HEMS implementation.

SHAP interpretability analysis yields findings that go far beyond what performance metrics alone can convey. Global Intensity emerges as a highly dominant predictor (average |SHAP| = 0.00350), with its superiority rooted in the electrophysical relationship P = V × I, which provides direct validation that the model has captured the correct underlying mechanisms of the data. Further temporal SHAP heatmaps reveal that this contribution is concentrated almost entirely on the most recent input time step (t-1), establishing that near-real-time current sensing is the highest-value sensing investment for HEMS forecast accuracy. The 3-hour rolling mean ranks second, indicating that multi-hour consumption inertia—rather than a one-step lagged value—is the most informative behavioral signal, with direct implications for the optimal lookahead horizon of the HEMS scheduling algorithm. Seasonal features confirm that adaptive recalibration accounting for the month is necessary to maintain forecast accuracy across the annual consumption cycle.

Overall, these contributions advance the research agenda for transparent and practical deep learning in smart home energy management through three key aspects. First, this study demonstrates that LSTM-based architectures are capable of capturing the nonlinear and non-stationary dynamics of household energy consumption at a level that linear statistical models fundamentally cannot match. Second, this research shows that SHAP-based post-hoc interpretability is not merely a decorative addition to predictive models, but a substantive analytical tool capable of generating new, physics-based insights into the factors driving household energy demand. Third, this study establishes a principled methodology to bridge the gap between model outputs and optimization actions, transforming predictions into actionable insights.

Several limitations of this study warrant transparent acknowledgment. Most significantly, the evaluation is conducted on a single-household dataset (IHEPC) recorded in Sceaux, France, between 2006 and 2010. While IHEPC is among the most widely validated benchmarks in residential energy forecasting, its single-household scope means that the learned temporal associations and SHAP-derived feature rankings may reflect behavioral and appliance-use patterns specific to this household, and cannot be assumed to generalize directly to households with different sizes, appliance inventories, occupancy schedules, climate zones, or cultural energy-use norms. The seasonal patterns identified in the SHAP analysis — particularly the prominence of winter month encodings — reflect the temperate oceanic climate of Île-de-France and may differ substantially in tropical, arid, or continental climates where cooling rather than heating dominates peak seasonal demand. Furthermore, the dataset predates the widespread adoption of rooftop solar panels, electric vehicles, and smart appliances — technologies that introduce consumption and generation patterns that the current feature set does not capture. The MAPE values (38–44%) across all neural models, while contextually justified by overnight consumption values approaching zero, indicate that predicting periods of low consumption remains a challenge. Future research should investigate multi-step forecasting horizons, hybrid CNN-LSTM architectures for enhanced feature extraction, and validation across multi-household datasets — such as REFIT (Murray et al., 2017), UK-DALE (Kelly & Knottenbelt, 2015), or ECO (Beckel et al., 2014) — spanning diverse geographies and time periods to establish the generalizability of the reported findings.

REFERENCES

[1] [1] IEA, 2022) IEA: buildings (2022) https://www.iea.org/reports/buildings. Accessed 23 Jun 2023.

[2] Y. Natarajan *et al.*, "Enhancing Building Energy Efficiency with IoT-Driven Hybrid Deep Learning Models for Accurate Energy Consumption Prediction," *Sustainability*, vol. 16, no. 5, 2024, doi: 10.3390/su16051925.

[3] Z. Severiche-Maury *et al.*, "Forecasting Residential Energy Consumption with the Use of Long Short-Term Memory Recurrent Neural Networks," *Energies (Basel).*, vol. 18, no. 5, 2025, doi: 10.3390/en18051247.

[4] R. Theiler and O. Fink, *Integrating the Expected Future: Schedule Based Energy Forecasting*. 2024. doi: 10.48550/arXiv.2409.05884.

[5] M. Kleinebrahm, J. Torriti, R. McKenna, A. Ardone, and W. Fichtner, "Using neural networks to model long-term dependencies in occupancy behavior," *Energy Build.*, vol. 240, p. 110879, 2021, doi: https://doi.org/10.1016/j.enbuild.2021.110879.

[6] M. Panja, T. Chakraborty, U. Kumar, and A. Hadid, "Probabilistic AutoRegressive Neural Networks for Accurate Long-Range Forecasting," in *Neural Information Processing*, B. Luo, L. Cheng, Z.-G. Wu, H. Li, and C. Li, Eds., Singapore: Springer Nature Singapore, 2024, pp. 457–477.

[7] P. Feng *et al.*, *Deep Learning and Machine Learning, Advancing Big Data Analytics and Management: Unveiling AI's Potential Through Tools, Techniques, and Applications*. 2024. doi: 10.48550/arXiv.2410.01268.

[8] S. U. Rehman and N. Iqbal, "Electricity consumption prediction in smart buildings using deep learning approaches (2025)," *Energy Informatics*, vol. 8, no. 1, p. 131, 2025, doi: 10.1186/s42162-025-00587-5.

[9] I. H. Ou Ali, A. Agga, M. Ouassaid, M. Maaroufi, A. Elrashidi, and H. Kotb, "Predicting short-term energy usage in a smart home using hybrid deep learning models," *Front. Energy Res.*, vol. Volume 12-2024, 2024, [Online]. Available: https://www.frontiersin.org/journals/energy-research/articles/10.3389/fenrg.2024.1323357

[10] D. da Silva and A. Meneses, *Comparing LSTM and BLSTM Deep Neural Networks for Power Consumption Prediction: Preliminary studies*. 2023. doi: 10.48550/arXiv.2305.16546.

[11] X. Wang, H. Wang, B. Bhandari, and L. Cheng, "AI-Empowered Methods for Smart Energy Consumption: A Review of Load Forecasting, Anomaly Detection and Demand Response," *International Journal of Precision Engineering and Manufacturing-Green Technology*, vol. 11, no. 3, pp. 963–993, 2024, doi: 10.1007/s40684-023-00537-0.

[12] Y. Himeur *et al.*, "AI-big data analytics for building automation and management systems: a survey, actual challenges and future perspectives," *Artif. Intell. Rev.*, vol. 56, no. 6, pp. 4929–5021, 2023, doi: 10.1007/s10462-022-10286-2.

[13] K. P. Natarajan and J. G. Singh, "Day-Ahead Residential Load Forecasting based on Variational Mode Decomposition and Hybrid Deep Networks with Granger Causality Feature Selection," in *2024 IEEE 3rd Industrial Electronics Society Annual On-Line Conference (ONCON)*, 2024, pp. 1–6. doi: 10.1109/ONCON62778.2024.10931691.

[14] D. P. Panagoulias, E. Sarmas, V. Marinakis, M. Virvou, G. A. Tsihrintzis, and H. Doukas, "Intelligent Decision Support for Energy Management: A Methodology for Tailored Explainability of Artificial Intelligence Analytics," *Electronics (Basel).*, vol. 12, no. 21, 2023, doi: 10.3390/electronics12214430.

[15] G. Airlangga, "Decoding Energy Usage Predictions: An Application of XAI Techniques for Enhanced Model Interpretability," *Indonesian Journal of Artificial Intelligence and Data Mining*, vol. 7, p. 275, Apr. 2024, doi: 10.24014/ijaidm.v7i2.29041.

[16] T. Clement, H. Nguyen, N. Kemmerzell, M. Abdelaal, and D. Stjelja, "Coping with Data Distribution Shifts: XAI-Based Adaptive Learning with SHAP Clustering for Energy Consumption Prediction," 2023, pp. 147–159. doi: 10.1007/978-981-99-8391-9_12.

[17] S. Letzgus, P. Wagner, J. Lederer, W. Samek, K. -R. Müller and G. Montavon, "Toward Explainable Artificial Intelligence for Regression Models: A methodological perspective," in *IEEE Signal Processing Magazine*, vol. 39, no. 4, pp. 40-58, July 2022, doi: 10.1109/MSP.2022.3153277.

[18] C. van Zyl, X. Ye, and R. Naidoo, "Harnessing eXplainable artificial intelligence for feature selection in time series energy forecasting: A comparative analysis of Grad-CAM and SHAP," *Appl. Energy*, vol. 353, p. 122079, 2024, doi: https://doi.org/10.1016/j.apenergy.2023.122079.

[19] A. Abdellatif *et al.*, "SHapley Additive exPlanations-guided rule-based energy management: bridging machine learning interpretability and adaptive control strategies," *Renewable Energy Focus*, vol. 56, p. 100779, 2026, doi: https://doi.org/10.1016/j.ref.2025.100779.

[20] A. Neubauer, S. Brandt, and M. Kriegel, "Explainable multi-step heating load forecasting: Using SHAP values and temporal attention mechanisms for enhanced interpretability," *Energy and AI*, vol. 20, p. 100480, 2025, doi: https://doi.org/10.1016/j.egyai.2025.100480

ISSN 2085-4552

Detection of AI-Generated Videos Using Local Binary Pattern and Support Vector Machine

Ergy David Lundy Tumanggor[1], Harlen Gilbert Simanullang[2], Arina Prima Silalahi[3]
Department of Informatics, Universitas Methodist Indonesia
Jl. Hang Tuah No.8, Madras Hulu, Kec. Medan Polonia, Kota Medan, Sumatera Utara 20151
ergydavid9145@ggmail.com[1], harlen.gilbert@gmail.com[2], primaarinasilalahi@gmail.com[3]

Accepted 29 May 2026
Approved 28 June 2026

***Abstract*— The rapid advancement of generative AI enables highly realistic fake videos, escalating financial and cybersecurity risks. As conventional verification methods become inadequate against evolving manipulations, this research proposes a robust, lightweight detection approach. The specific novelty of this study lies in repurposing a classic spatial texture extractor—Local Binary Pattern (LBP)—into a dynamic forensic tool by statistically aggregating frame-level micro-textures across the temporal axis to form a continuous 144-dimensional feature vector. This completely bypasses the massive computational overhead of deep learning. Utilizing the SDFVD2.0 dataset of 927 videos , frames undergo rigorous preprocessing, including face cropping and noise reduction, before LBP extraction. A linear Support Vector Machine (SVM) then classifies these vectors. The LBP-SVM model achieved an accuracy of 81.18% on the testing split and demonstrated swift processing (113.82 seconds for 40 unseen videos) during generalization testing. Although margins of error exist with highly complex artifacts , this spatial-temporal combination provides an efficient, interpretable, and computationally economical baseline for digital forensics.**

***Index Terms*— AI-Generated Video; Local Binary Pattern; Support Vector Machine.**

I. INTRODUCTION

The rapid advancement of generative artificial intelligence (AI) has enabled the creation of highly realistic fake videos that spread widely across social media and digital platforms. Consequently, it is increasingly difficult for the public to distinguish between genuine and fake videos, raising severe risks of financial fraud, disinformation, reputation attacks, privacy violations, and broader cybersecurity threats. Conventional inspection methods, such as manual visual verification or standard metadata analysis, are no longer adequate because digital manipulation techniques continue to evolve rapidly, seamlessly altering facial expressions, lip-syncing, and identities [1], [2], [3].

In the realm of digital forensics, numerous previous studies have tackled deepfake detection utilizing deep learning approaches based on convolutional neural networks (CNNs). For instance, research on deepfake detection using CNN architectures reported high performance, but these methods consistently demand massive datasets [3], [4], [5], [6], [7] , significant computational resources, and intensive transfer learning processes [8], [9], [10], [11], [12]. Conversely, in the broader field of machine learning, classic supervised learning methods like the Support Vector Machine (SVM) have proven highly reliable in predicting data classes [13] and are widely applied to various classification problems, including image and binary classifications [14]. Furthermore, classic image processing methods like Local Binary Pattern (LBP) have been successfully utilized to extract texture features for static object classification, such as dynamic video manipulation detection, material corrosion analysis, and cultural heritage pattern classification [15], [16], [17], [18], [19], [20].

Despite these advancements, there is a noticeable gap in the literature regarding the application of lightweight, classic machine learning techniques for dynamic video manipulation detection. While deep learning approaches like Convolutional Neural Networks (CNNs) offer robust automated feature extraction for digital forensics, they often demand massive datasets, significant computational resources, and intensive transfer learning processes [11], [12], [21]. These architectures typically require tens of millions of trainable parameters and high-end Graphical Processing Units (GPUs) for inference, often taking days to train and struggling to run in real-time on standard hardware. Conversely, this study introduces a distinct novelty by repositioning the classic spatial-temporal texture extraction of LBP combined with SVM as a highly efficient, interpretative baseline for dynamic AI-generated video detection [22], [23]. While LBP is a traditional image processing technique, its novelty here lies in its application to dynamic video

forensics. By statistically aggregating frame-level micro-textures—specifically the mean and standard deviation—across the temporal axis, this methodology transforms a static texture analyzer into a dynamic, 144-dimensional spatial-temporal feature extractor. This provides a highly efficient, interpretable baseline that solves the computational bottlenecks of CNNs, making it specifically designed for rapid verification on standard hardware devices with limited computing power.

By focusing on extracting the micro-texture of the facial area at the frame level using the SDFVD2.0 dataset, this study aims to close the existing gap by providing a simpler, classic texture feature approach. The proposed LBP-SVM model offers a lightweight, fast, and efficient baseline verification tool for digital forensics that can be easily implemented on devices with limited computing power, without the heavy burden of deep learning architectures.

A. System Architecture and Pipeline

This research adopts a chronological procedure starting from data collection, preprocessing, feature extraction, to SVM implementation. To provide a clear picture of the system architecture, the research framework is structured as a continuous workflow which can be seen in Fig. 1.

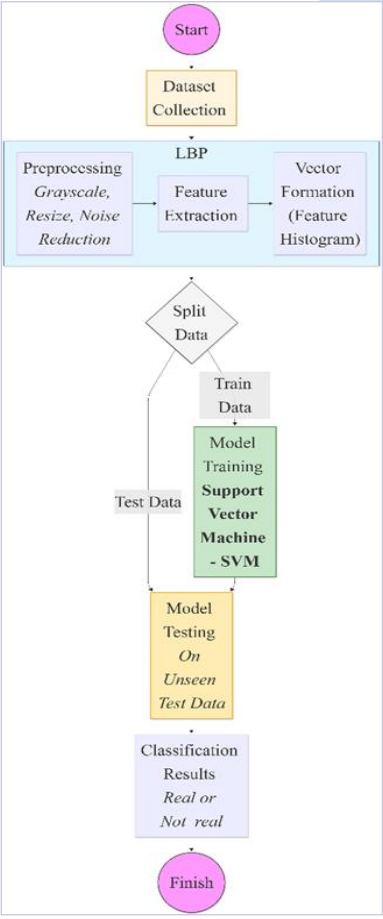


Fig. 1. Research Framework pipeline.

To formalize the research procedure, the systematic workflow of the proposed framework is detailed in Algorithm 1.

To ensure reproducibility and scientific validity, the sequential logic of the proposed LBP-SVM framework is mapped out visually. Instead of utilizing traditional algorithmic pseudocode, the systematic procedure of this research is comprehensively illustrated in Fig. 1.

B. Mathematical Formulation of Local Binary Pattern (LBP)

The feature extraction utilizes the Local Binary Pattern (LBP) method to capture local texture patterns. LBP is a gray-scale invariant texture measure that works by thresholding the neighborhood of each pixel and considers the result as a binary number. The LBP value is calculated by comparing the center pixel with its surrounding neighbors.

The mathematical calculation is defined in Equation (1):

$$LBP_{P,R} = \sum_{p=0}^{P-1} S(g_p - g_c)2^p \quad (1)$$

Where P is the number of neighbor points, and R is the radius of the neighborhood. The coordinates (x_c, y_c) represent the center pixel with a grayscale intensity of g_c, while g_p represents the intensity of the neighbor pixel. The function S(x) is a thresholding step function defined in Equation (2):

$$S(x) = 1, if\ x \geq 0\ ;\ S(x) = 0, if\ x < 0 \quad (2)$$

This system applies 16 LBP points (P=16) and a radius of 2 pixels (R=2), explicitly utilizing "uniform patterns." A binary pattern is considered uniform if it contains at most two bitwise transitions (from 0 to 1 or 1 to 0) in the circular representation of the pattern. The uniformity is calculated using Equation (3):

$$U(b) = \sum_{p=0}^{P-1} | b_p - b_{(p+1)}\ mod\ p | \quad (3)$$

Patterns where U(b) ≤ 2 are categorized as uniform, representing fundamental micro-textures like edges, corners, and flat regions, which are highly descriptive of facial manipulation artifacts.

C. Mathematical Formulation of Support Vector Machine (SVM)

As a supervised learning algorithm, the Support Vector Machine (SVM) works by finding the optimal decision boundary (hyperplane) to separate data into different classes with the maximum possible margin. In the broader scope of machine learning, choosing the right classification algorithm is crucial for handling complex features [14].

The linear SVM classification is based on finding a hyperplane that maximizes the margin between the two classes (real and not_real). The equation of the optimal hyperplane is defined in Equation (4):

$$w\ .\ x + b = 0 \quad (4)$$

Where w is the weight vector, x is the input feature vector (the 144-dimensional LBP features), and b is the bias.

To classify a new video data point, the model evaluates the features using the following decision function shown in Equation (5):

$$f(x) = sign(w \,.\, x + b) \quad (5)$$

If the output of f(x) is positive (+1), the data is classified into the first class (e.g., real). If the output is negative (-1), it is classified into the second class (e.g., not_real). SVM is highly suitable for this research due to its robustness in processing data within high-dimensional feature spaces without the heavy computational burden of deep learning architectures.

II. Result and Discussions

A. Preprocessing

Before extracting features, each video frame underwent a rigorous preprocessing pipeline (illustrated in Fig. 2) to ensure dimensional and format uniformity. This standard procedure is essential to enhance the precision of image-based classification tasks. The preprocessing stages included converting the original RGB frames (Fig. 2a) to grayscale (Fig. 2b), resizing them to a uniform 256x256 pixel resolution (Fig. 2c), applying a 3x3 Gaussian Blur to mitigate extreme compression artifacts (Fig. 2d), and finally, executing a face cropping operation (Fig. 2e) to isolate the Region of Interest (ROI).

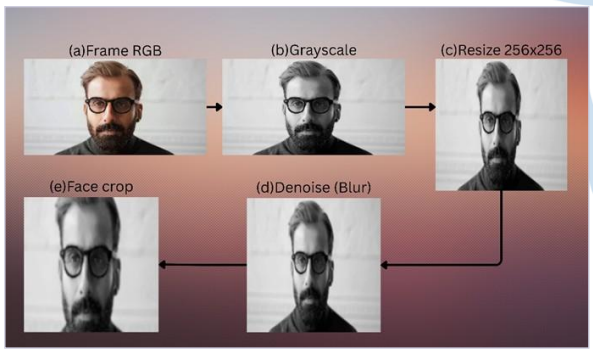


Fig. 2. Preprocessing stages on a single frame: (a) Original RGB, (b) Grayscale, (c) Resize 256x256, (d) Gaussian Blur Denoising, (e) Face Cropping.

This research adopts a chronological procedure starting from data collection, preprocessing, feature extraction, to SVM implementation. The process begins with video data collection from the SDFVD2.0 dataset, comprising 927 videos (456 real and 471 not_real). Because videos consist of multiple frames with high redundancy, a frame sampling mechanism is employed. These extracted frames then enter a rigid preprocessing pipeline to standardize the image format. To isolate the Region of Interest (ROI), an automated face detection algorithm using OpenCV is utilized [25]. In instances of face detection failure—often caused by extreme head angles, occlusions, or rapid motion—the system is programmed with a fallback mechanism to extract the center region of the frame (center crop) to ensure continuous feature extraction without injecting background noise. The isolated frames are then converted to a single-channel grayscale matrix, as the LBP operator only evaluates single-pixel intensities. Next, the frames are resized to a uniform 256x256 pixels. This specific resolution was selected to standardize the input matrix while retaining sufficient facial micro-textures without causing excessive memory consumption. Furthermore, a 3x3 Gaussian blur is applied. While blurring inherently reduces high-frequency data, a minimal 3x3 kernel was deliberately chosen as a necessary trade-off to smooth out benign, standard video compression artifacts that could mislead the classifier, without destroying the larger structural inconsistencies characteristic of AI manipulation.

B. Feature Extraction

Once the isolated 256x256 grayscale facial image is prepared, the LBP operator evaluates the micro-texture. The LBP parameters were specifically set to 16 neighbor points (P=16) and a radius of 2 pixels (R=2). Expanding the radius to R=2, utilizing precise bilinear interpolation to evaluate the exact intensities of diagonal neighbors, allows the algorithm to capture slightly larger, overlapping micro-textures. This spatial reach is critical for detecting the subtle, sub-pixel blending edges and warping artifacts typically left behind by generative AI algorithms at the boundaries of the face. To capture the dynamic changes across the video timeline, the Mean (μ) and Standard Deviation (σ) of these frame-level features are computed and concatenated into a final 144-dimensional continuous feature vector per video. The exact configuration is detailed in Table 1.

TABLE I. Feature Extraction Configuration

Parameter	Value
Image Size	256 x 256 pixels
LBP Points (P)	16
LBP Radius (R)	2
LBP Method	Uniform pattern
Bins per Grid	18
Grid Size	2 x 2 = 4 grids
Final Features	144 (72 Mean + 72 Std Dev)

To capture the video's temporal dynamics, both the mean (μ) and standard deviation (σ) of these features are calculated across all frames, generating a final continuous feature vector of 144 dimensions. Utilizing statistical features like mean and standard deviation from image matrices has been highly effective in various texture classification approaches [26].

C. Model Training and Evaluation

For classification, the dataset of 144-dimensional continuous features is normalized using Z-score standardization. A Linear Support Vector Machine (SVM) is then trained on 80% of the data, utilizing a penalty parameter of C=10.0 and balanced class weights to define the optimal decision boundary [13]. A Linear SVM was specifically selected over non-linear kernels due to the high-dimensional nature of the extracted data. In 144-dimensional spaces, data points are often already linearly separable; applying a complex, non-linear kernel across a relatively compact dataset poses a severe risk of overfitting to the training distribution. Therefore, the linear kernel prevents overfitting while maintaining mathematical interpretability, allowing for transparent observation of the feature weights dictating the decision boundary.

During training, the SVM resolved the Dual Lagrange formulation to identify the optimal hyperplane. Out of the 741 training samples, the algorithm identified precisely 372 critical data points as Support Vectors (data points where the Lagrange multiplier $\alpha > 0$). These vectors dictate the exact position of the decision boundary. By mathematically verifying the Karush-Kuhn-Tucker (KKT) conditions, it was observed that bounded support vectors precisely hit the penalty limits (e.g., $\alpha = 9.623$ for manipulated samples), while free support vectors lay exactly on the margin boundaries. The accumulation of the weight vector components resulted in an optimal geometric margin width of Margin = 2 / ||w|| = 0.1151. To better understand the feature distribution and the classification boundary, the 144-dimensional feature space was projected into a 2D visualization using dimensionality reduction, as shown in Fig. 3.

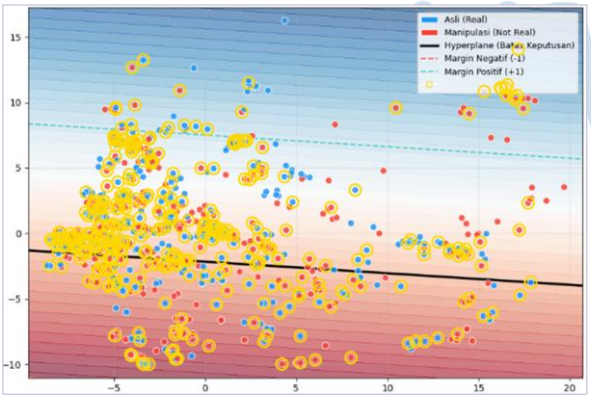


Fig. 3. SVM Hyperplane, Margin boundaries, and Support Vectors projection.

In Fig. 3, the solid black line represents the optimal linear decision boundary separating the real class from the AI-generated class. The dashed lines illustrate the margin boundaries for each class. The data points highlighted with yellow circles are the Support Vectors, which are the critical samples resting exactly on or violating the margin bounds. The mathematical accumulation of these support vectors resulted in a calculated optimal margin width of 0.1151, maximizing the distance between the two classes to lower the generalization error. The model evaluation on the testing data yielded an overall accuracy of 81.18%. The detailed performance metrics are presented in Table 2.

TABLE II. CLASSIFICATION REPORT

Metric	Precision	Recall	F1-Score	Support
real	0.8283	0.8200	0.8241	100
not_real	0.7931	0.8023	0.7977	86
Accuracy			0.8118	186

To provide deeper insight into the model's classification behavior and to detail the exact distribution of true and false predictions across the testing split, a confusion matrix was generated as presented in Fig 4.

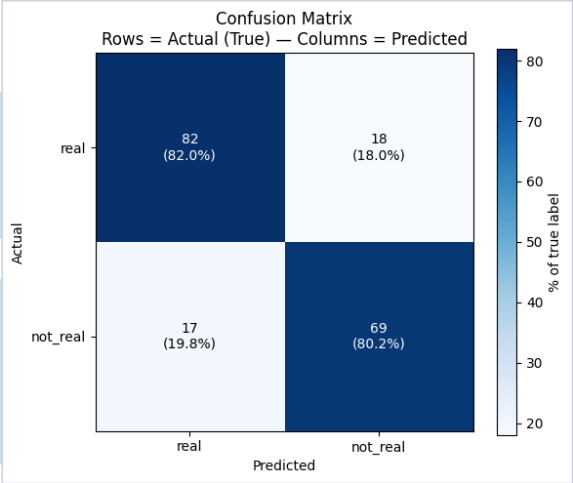


Fig. 4. SVM Hyperplane, Margin boundaries, and Support Vectors projection.

D. Testing on Unseen Data

To rigorously validate the model's generalization capabilities in real-world scenarios, it was tested on 40 completely new video datasets not included in the training or validation splits. The final decision function for the linear kernel is formulated in equation (6).

$$f(x_{new}) = sign(\sum_{i=1}^{144}(w_i x_{i,new}) + b \qquad (6)$$

Where w_i are the learned weights, $x_{i,new}$ are the scaled features of the new video, and b is the learned bias (calculated as 0.2234). If the output is positive, the video is classified as real; if negative, it is classified as not_real. The results of this unseen data testing are summarized in Table 3.

TABLE III. PREDICTION RESULTS SUMMARY

No	Video	Prediksi	Conf.	P(real)	P(fake)	Frame	Waktu
1	video_1_not_real.mp4	not_real	62.54%	37.46%	62.54%	362	7.63s
2	video_1_real.mp4	not_real	65.58%	34.42%	65.58%	81	1.74s

3	video_2_not_real.mp4	real	50.74%	50.74%	49.26%	307	5.85s
4	video_2_real.mp4	real	52.62%	52.62%	47.38%	81	1.73s
5	video_3_not_real.mp4	not_real	80.91%	19.09%	80.91%	82	1.77s
6	video_3_real.mp4	not_real	56.66%	43.34%	56.66%	321	10.39s
7	video_4_not_real.mp4	not_real	62.41%	37.59%	62.41%	82	2.07s
8	video_4_real.mp4	real	67.55%	67.55%	32.45%	60	1.64s
9	video_5_not_real.mp4	not_real	65.58%	34.42%	65.58%	120	3.02s
10	video_5_real.mp4	real	76.11%	76.11%	23.89%	244	13.17s
11	video_6_not_real.mp4	not_real	66.73%	33.27%	66.73%	120	2.58s
12	video_6_real.mp4	real	51.78%	51.78%	48.22%	81	1.74s
13	video_7_not_real.mp4	not_real	59.59%	40.41%	59.59%	102	2.02s
14	video_7_real.mp4	real	53.64%	53.64%	46.36%	120	2.78s
15	video_8_not_real.mp4	not_real	56.56%	43.44%	56.56%	102	2.28s
16	video_8_real.mp4	real	65.68%	65.68%	34.32%	120	2.85s
17	video_9_not_real.mp4	not_real	60.63%	39.37%	60.63%	127	3.00s
18	video_9_real.mp4	real	54.32%	54.32%	45.68%	120	2.89s
19	video_10_not_real.mp4	not_real	55.93%	44.07%	55.93%	127	2.50s
20	video_10_real.mp4	not_real	54.55%	45.45%	54.55%	94	1.98s
21	video_11_not_real.mp4	real	51.18%	48.82%	51.18%	171	3.26s
22	video_11_real.mp4	not_real	55.67%	44.33%	55.67%	94	1.98s
23	video_12_not_real.mp4	not_real	55.57%	44.43%	55.57%	171	3.24s
24	video_12_real.mp4	not_real	54.12%	45.88%	54.12%	94	1.95s
25	video_13_not_real.mp4	real	50.77%	50.77%	49.23%	118	2.52s
26	video_13_real.mp4	real	54.00%	54.00%	46.00%	117	2.43s
27	video_14_not_real.mp4	not_real	53.14%	46.86%	53.14%	118	2.53s
28	video_14_real.mp4	real	50.97%	49.03%	50.97%	117	2.43s
29	video_15_not_real.mp4	not_real	58.11%	41.89%	58.11%	100	2.11s
30	video_15_real.mp4	real	53.65%	53.65%	46.35%	117	2.29s
31	video_16_not_real.mp4	not_real	58.11%	41.89%	58.11%	100	2.12s
32	video_16_real.mp4	real	53.65%	53.65%	46.35%	99	1.87s
33	video_17_not_real.mp4	not_real	56.95%	43.05%	56.95%	93	1.96s
34	video_17_real.mp4	real	58.17%	58.17%	41.83%	94	1.93s
35	video_18_not_real.mp4	not_real	52.29%	47.71%	52.29%	93	2.03s
36	video_18_real.mp4	real	63.64%	63.64%	36.36%	94	1.95s
37	video_19_not_real.mp4	real	58.78%	58.78%	41.22%	97	2.12s
38	video_19_real.mp4	real	60.54%	60.54%	39.46%	94	1.95s
39	video_20_not_real.mp4	real	51.53%	48.47%	51.53%	97	2.15s
40	video_20_real.mp4	real	55.82%	55.82%	44.18%	115	2.35s

The system successfully and accurately predicted 30 out of 40 videos, achieving a generalization accuracy of 75%. The model demonstrated an average confidence rate of 58.43% with a total processing time of 113.82 seconds for all 40 videos combined. This indicates that when the model encounters borderline data with highly complex textures that deviate significantly from the training distribution, it becomes mathematically uncertain.

Despite these margin errors, the findings substantiate that the spatial-temporal combination of LBP and SVM provides a highly efficient, transparent, and lightweight baseline for digital forensics. This approach operates effectively and rapidly on devices with limited computing power, offering a practical alternative to computationally heavy CNN architectures.

E. Model Evaluation and Error Analysis

While the LBP-SVM framework demonstrated high processing efficiency, an in-depth evaluation reveals key vulnerability margins in the classification boundaries. The misclassification of **10 videos** (such as *video_2_not_real*, *video_11_real*, and *video_20_not_real*) during generalization testing highlights the specific limitations of spatial-temporal LBP when confronting advanced generative manipulation. Noticeably, the confidence scores for these misclassified instances tend to hover closely around the borderline margin (50% to 56%), demonstrating that when the model encounters highly complex textures that deviate significantly from the training distribution, it becomes mathematically uncertain.

Error analysis indicates that these failures stem from highly complex artifacts that deviate from the fundamental micro-textures (edges, corners, and flat regions) captured by the uniform pattern of $U(b) \leq 2$. Specifically, erratic environmental lighting and subtle sub-pixel blending applied by newer AI generation methods create non-uniform local binary patterns that push the feature vectors directly onto the decision margin, creating mathematical uncertainty for the linear hyperplane. This demonstrates that while the SVM effectively separates standard manipulation artifacts, its linear separation struggles with highly non-linear, multi-illuminated synthetic textures.

The implementation of Local Binary Pattern (LBP) and Support Vector Machine (SVM) successfully establishes a scientifically rigorous, mathematically sound, and highly efficient baseline for detecting AI-generated videos. The proposed model, utilizing a 144-dimensional temporal-aggregated feature vector,

achieved a solid accuracy of 81.18% on the testing split. Furthermore, it correctly classified 30 out of 40 completely unseen videos during generalization testing, yielding an average confidence rate of 58.43% with an incredibly rapid total processing time of 113.82 seconds. These empirical results confirm that the LBP-SVM framework provides a lightweight, fast, and computationally economical verification tool, making it highly practical for on-the-fly digital forensic applications operating on standard hardware.

Despite its computational efficiency and transparency, this classic spatial-temporal texture extraction approach encounters certain limitations. The model faces challenges—exhibiting lower confidence and margin errors—when classifying videos containing highly complex textures, erratic lighting, or subtle manipulation artifacts that deviate significantly from the baseline training dataset. To address these limitations and advance future research, it is highly recommended to explore the integration or comparative analysis of this conventional LBP-SVM method with advanced Deep Learning architectures [27] to enhance granular feature extraction capabilities. Furthermore, expanding the dataset to include synthetic media from newer generative AI models (such as diffusion models), and evaluating the system under more diverse real-world conditions (varying lighting environments, differing resolutions, and diverse visual angles) will be crucial to strengthening the overall robustness and reliability of the detection system. Finally, deploying this trained model into a user-friendly application interface (UI/UX) is recommended to facilitate accessibility for end-users in real-world forensic investigations.

REFERENCES

[1] A. A. Khan, A. A. Laghari, S. A. Inam, S. Ullah, M. Shahzad, and D. Syed, "A survey on multimedia-enabled deepfake detection: state-of-the-art tools and techniques, emerging trends, current challenges & limitations, and future directions," Dec. 01, 2025, *Springer Science and Business Media B.V.* doi: 10.1007/s10791-025-09550-0.

[2] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection," Jun. 2020, [Online]. Available: http://arxiv.org/abs/2001.00179

[3] Y. Li and S. Lyu, "Exposing DeepFake Videos By Detecting Face Warping Artifacts," *arXiv*, May 2019, [Online]. Available: http://arxiv.org/abs/1811.00656

[4] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images," *arXiv*, Aug. 2019, [Online]. Available: http://arxiv.org/abs/1901.08971

[5] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: a Compact Facial Video Forgery Detection Network," *arXiv*, Sep. 2018, doi: 10.1109/WIFS.2018.8630761.

[6] I. Masi, A. Killekar, R. M. Mascarenhas, S. P. Gurudatt, and W. AbdAlmageed, "Two-branch Recurrent Network for Isolating Deepfakes in Videos," *arXiv*, Sep. 2020, [Online]. Available: http://arxiv.org/abs/2008.03412

[7] S. M. Yasir and H. Kim, "Lightweight Deepfake Detection Based on Multi-Feature Fusion," *Appl. Sci.*, vol. 15, no. 4, Feb. 2025, doi: 10.3390/app15041954.

[8] J. C. C, O. J. A, N. M.D, S. Ahmad, F. O. O, and U. A. A, "Development Of Artificial Intelligence-Based Model for Forensic Analysis of Cross-Platform Deepfakes," *Int. J. Res. Sci. Innov.*, 2025, doi: 10.51244/IJRSI.

[9] J. Mallet, L. Pryor, D. R. Dave, and D. M. Vanamala, "Deepfake Detection Analyzing Hybrid Dataset Utilizing CNN and SVM," 2023.

[10] Y. Jugade, Z. Syed, C. Dcosta, A. Panicker, and D. Vora, "Deepfake detection via spatial–temporal deep networks: leveraging CNNs and LSTMs for enhanced accuracy," *Discov. Appl. Sci.*, Jan. 2026, doi: 10.1007/s42452-025-08014-w.

[11] S. A. Falcón-López, L. Tobarra, A. Robles-Gómez, and R. Pastor-Vargas, "Forensic Analysis of Manipulated Images and Videos," *MDPI*, vol. 15, no. 23, Dec. 2025, doi: 10.3390/app152312664.

[12] A. Raza *et al.*, "A Comprehensive Review of Deepfake Detection Techniques: From Traditional Machine Learning to Advanced Deep Learning Architectures," *MDPI*, vol. 7, no. 2, Feb. 2026, doi: 10.3390/ai7020068.

[13] Andi Saenong and R. Rahmat, "Implementasi Algoritma SVM Untuk Sistem Deteksi dan Pengawasan Keamanan Kendaraan di Area Parkir Menggunakan Kamera," *Simkom*, vol. 9, no. 2, pp. 267–277, Jul. 2024, doi: 10.51717/simkom.v9i2.570.

[14] A. P. Silalahi and H. G. Simanullang, *DECISSION TREE DAN PENERAPANNYA Dengan Algoritma Iterative Dichotomizes versi 3 (ID3)*. Madza Media, 2025. [Online]. Available: www.madzamedia.co.id

[15] Y. Chen, N. Akhtar, N. A. H. Haldar, and A. Mian, "Deepfake Detection with Spatio-Temporal Consistency and Attention," *Univ. West. Aust.*, Feb. 2025, [Online]. Available: http://arxiv.org/abs/2502.08216

[16] J. Jiang, W. C. Yang, C. H. Chen, and T. Young, "A New Deepfake Detection Method with No-Reference Image Quality Assessment to Resist Image Degradation," *MDPI*, vol. 6, no. 10, Oct. 2025, doi: 10.3390/eng6100274.

[17] C. Internò, R. Geirhos, M. Olhofer, S. Liu, B. Hammer, and D. Klindt, "AI-Generated Video Detection via Perceptual Straightening," *arXiv*, Feb. 2026, [Online]. Available: http://arxiv.org/abs/2507.00583

[18] F. Bimantoro, R. Nisful, L. Hidhayah, and D. Swanjaya, "Komparasi Algoritma MLP+LBP dan CNN Sebagai solusi Inovatif Untuk Deteksi Dini Korosi," *INOTEK*, vol. 8, pp. 2549–7952, 2024.

[19] S. K. Dirjen, P. Riset, D. Pengembangan, R. Dikti, P. N. Andono, and E. H. Rachmawanto, "Evaluasi Ekstraksi Fitur GLCM dan LBP Menggunakan Multikernel SVM untuk Klasifikasi Batik," *J. Resti*, no. 3, 2021, doi: 10.29207/resti.v5i1.265.

[20] N. S. Fatimah and S. Agustin, "Klasifikasi Citra Batik Menggunakan Local Binary Pattern (LBP) dan Support Vector Machine (SVM)," *J. Algoritm.*, 2025.

[21] M. D. Rahadian, A. Aziz, and D. W. Prabowo, "Performance Analysis of the LBP-SVM Model in Deepfake Image Detection: A Case Study on the FaceForensics++ Dataset and External Validation on Indonesian Politicians' Faces."

[22] S. M. Coloma *et al.*, "Deepfake and Forged Image Detection through Enhanced Local Binary Pattern Histogram (eLBPH) Features and SVM Classification for Digital Image Forensics," 2025.

[23] K. Azmi, S. Defit, and Sumijan, "Implementasi Convolutional Neural Network (CNN) Untuk Klasifikasi Batik Tanah Liat Sumatera Barat," *J. Unitek*, vol. 16, no. 1, p. 2023, 2023.

[24] A. I. Nurulrachman, R. Cahya Wihandika, and P. P. Adikara, "Ekstraksi Ciri pada Klasifikasi Citra Batik menggunakan Metode Gray Level Co-Occurrence Matrix, Local Binary Pattern, dan HSV Color Moment," 2023. [Online]. Available: http://j-ptiik.ub.ac.id

[25] R. H. P. Sejati and R. Mardhiyyah, "DETEKSI WAJAH BERBASIS FACIAL LANDMARK MENGGUNAKAN OPENCV DAN DLIB," *J. Teknol. Inf.*, vol. 5, no. 2, 2021.

[26] B. Zi, M. Chang, J. Chen, X. Ma, and Y. G. Jiang, "WildDeepfake: A Challenging Real-World Dataset for Deepfake Detection," in *MM 2020 - Proceedings of the 28th ACM International Conference on Multimedia*, Association for Computing Machinery, Inc, Oct. 2020, pp. 2382–2390. doi: 10.1145/3394171.3413769.

[27] L. Guarnera, O. Giudice, and S. Battiato, "Fighting deepfake by exposing the convolutional traces on images," *IEEE Access*, vol. 8, pp. 165085–165098, 2020, doi: 10.1109/ACCESS.2020.3023037.

ISSN 2085-4552

Comparative Analysis of CNN and MobileNetV2 With Data Augmentation and Transfer Learning for Waste Classification

Deas Aghellar[1], Aswan Supriyadi Sunge[2]
[1,2]Faculty of Computer Engineering and Design, Nusaputra University, Sukabumi, Indonesia
[1]deas.aghellar@nusaputra.ac.id, [2]aswan.sunge@pelitabangsa.ac.id

Accepted 21 June 2026
Approved 28 June 2026

***Abstract*— Waste management has become an increasingly important environmental challenge that requires efficient and intelligent classification systems. Deep learning-based approaches have shown significant potential in automating waste classification and improving recycling processes. This study compares the performance of Convolutional Neural Network (CNN) and MobileNetV2 models for image-based waste classification and evaluates the impact of data augmentation and optimization techniques on classification performance.**

The experiments were conducted using the publicly available TrashNet dataset, which contains 2,527 images categorized into six waste classes: cardboard, glass, metal, paper, plastic, and trash. Data augmentation techniques, including rotation, translation, zooming, horizontal flipping, and brightness adjustment, were applied to improve data diversity. MobileNetV2 was implemented using transfer learning with pretrained ImageNet weights, while optimization techniques such as dropout, learning rate adjustment, and early stopping were employed to enhance model performance and generalization capability.

Experimental results show that MobileNetV2 outperformed the conventional CNN model, achieving a validation accuracy of 69.8%, whereas CNN achieved 48.5%. Further evaluation indicated that MobileNetV2 obtained an overall accuracy of 71.97%, precision of 73%, recall of 72%, and F1-score of 71%. Confusion matrix analysis revealed strong classification performance for paper and cardboard categories, while the trash category remained the most challenging due to higher visual diversity and limited training samples.
These findings demonstrate that MobileNetV2 combined with data augmentation and optimization techniques provides an effective and computationally efficient approach for waste classification tasks.

***Index Terms*— Waste Classification; Deep Learning; CNN; MobileNetV2; Transfer Learning; Data Augmentation.**

I. INTRODUCTION

Waste management has become an increasingly complex environmental issue due to population growth and intensified human activities. Ineffective waste management can result in environmental pollution, public health problems, and a decline in quality of life. One of the primary challenges lies in the waste sorting process, which is still largely performed manually, making it inefficient and prone to human error. Therefore, technology-based solutions are essential to improve the accuracy and efficiency of waste classification [1], [2].

In recent years, deep learning approaches, particularly Convolutional Neural Networks (CNNs), have been widely applied in image classification due to their ability to automatically extract features [3]. However, conventional CNN models often suffer from limitations in computational efficiency and generalization capability, especially when trained on limited datasets. To address these issues, lightweight architectures such as MobileNetV2 have been developed to improve efficiency without sacrificing accuracy, particularly through transfer learning techniques [4], [5].

Previous studies have compared CNN and MobileNet models in various classification tasks. Zhang et al. [6] demonstrated that MobileNet provides better computational efficiency compared to conventional CNN. Other studies highlight the importance of data augmentation in improving model performance, especially for limited datasets [7]. Although previous studies have explored CNN and MobileNetV2 for waste classification, most research primarily focuses on model comparison and often evaluates optimization techniques separately. Limited studies comprehensively investigate the combined impact of data augmentation, transfer learning, transfer learning with frozen weights, dropout, and learning rate adjustment on waste classification performance using lightweight architectures. Recent studies have shown that transfer learning and lightweight architectures such as MobileNetV2 can significantly

improve waste classification performance while reducing computational complexity [18], [20], [23].

Therefore, this study addresses the identified research gap by proposing an integrated framework that combines CNN and MobileNetV2 comparison with data augmentation and optimization techniques. The objective is to evaluate how these techniques contribute to improving classification accuracy, robustness, and generalization capability in waste classification tasks.

II. METHOD

Research Design

This study adopts an experimental approach to compare the performance of CNN and MobileNetV2 models for image-based waste classification. Data augmentation and optimization techniques are integrated to enhance model performance and generalization.

The research workflow includes data collection, preprocessing, model training, evaluation, and result analysis. This approach enables an objective comparison of model performance.

2.2 Related Works

Waste classification has attracted increasing attention in recent years due to the growing environmental impact of improper waste management. Deep learning approaches, particularly Convolutional Neural Networks (CNNs), have been widely adopted because of their ability to automatically extract discriminative visual features from images [2], [3].

Several studies have investigated lightweight architectures to improve computational efficiency while maintaining classification performance. MobileNet and MobileNetV2 have become popular choices because of their depthwise separable convolution mechanism, which significantly reduces the number of parameters and computational requirements [4], [5]. Previous studies reported that MobileNet-based architectures provide competitive performance for image classification tasks while being suitable for resource-constrained devices [6]. More recent studies have demonstrated that MobileNetV2 combined with transfer learning can achieve promising results for waste classification tasks while maintaining computational efficiency [17], [18], [23].

Data augmentation has also been recognized as an effective strategy for improving model robustness and reducing overfitting, particularly when training data are limited [7]. Recent studies further reported that augmentation techniques such as rotation, zooming, and brightness adjustment contribute significantly to improving classification accuracy and model generalization [21], [24]. Furthermore, transfer learning enables models to leverage knowledge learned from large-scale datasets, resulting in improved classification performance and faster convergence [8], [18], [22].

Despite these advances, most previous studies focused either on model comparison or on individual optimization techniques. Limited research has comprehensively investigated the combined effects of transfer learning, data augmentation, dropout, learning rate adjustment, and early stopping within a unified waste classification framework. Therefore, this study proposes an integrated approach that combines CNN and MobileNetV2 comparison with multiple optimization strategies to improve classification accuracy and generalization capability.

Table 1. Summary of Related Works

No	Study	Method	Main Contribution	Limitation
1	Rahman et al. [2]	CNN	Automated waste detection using convolutional neural networks	Limited generalization capability
2	LeCun et al. [3]	Deep Learning	Established the foundation of deep learning for image classification	Not specific to waste classification
3	Sandler et al. [4]	MobileNetV2	Introduced a lightweight architecture with improved efficiency	Requires transfer learning for optimal performance
4	Howard et al. [5]	MobileNet	Developed efficient CNN architectures for mobile devices	Lower representational capacity than larger networks
5	Zhang et al. [6]	Lightweight CNN	Improved computational efficiency for image classification	Limited feature extraction capability
6	Shorten and Khoshgoftaar [7]	Data Augmentation	Demonstrated the effectiveness of image augmentation techniques	Does not address model architecture optimization
7	Li et al. [8]	CNN + Data Augmentation	Improved classification accuracy using augmentation techniques	Sensitive to augmentation settings
8	Ahmed et al. [9]	Transfer Learning	Enhanced waste classification performance using transfer learning	Dependent on pretrained models
9	Brown et al. [10]	Deep Learning Optimization	Investigated optimization techniques for deep learning models	Did not focus on waste classification
10	This Study	CNN vs MobileNetV2 + Optimization	Combines transfer learning, data augmentation, dropout, learning rate adjustment, and early stopping	Limited dataset size

Although previous studies have demonstrated promising performance in waste classification using CNN and MobileNetV2, several limitations remain. Most studies primarily focused on model comparison

without comprehensively integrating optimization strategies such as data augmentation, dropout, learning rate adjustment, and transfer learning. In addition, several studies reported performance degradation when dealing with visually similar waste categories or limited datasets. Therefore, this study proposes an integrated approach combining CNN and MobileNetV2 with multiple optimization techniques to improve classification performance and generalization capability.

2.3 Literature Review Flowchart

As shown in Figure 1, the literature review flowchart illustrates the systematic process of conducting the Systematic Literature Review (SLR), starting from problem identification, followed by literature search, screening, eligibility assessment, data extraction, analysis and synthesis, research gap identification, and concluding with the literature review summary.

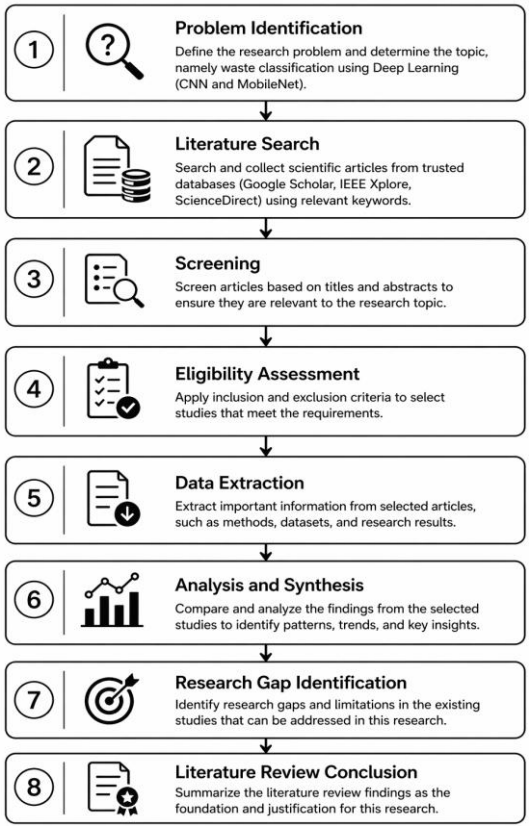


Figure 1. Flowchart

2.4 Dataset

The dataset used in this study was obtained from the publicly available TrashNet dataset available on Kaggle [11].. The dataset consists of six waste categories, namely cardboard, glass, metal, paper, plastic, and trash, with a total of approximately 2,527 images.



Figure 2. Representative samples from the TrashNet waste classification dataset, consisting of six categories:cardboard, glass, metal, paper, plastic, and trash

As illustrated in Figure 2, the dataset consists of six waste categories, namely cardboard, glass, metal, paper, plastic, and trash. Each category contains images with varying lighting conditions, object orientations, and background complexity to represent real-world waste classification scenarios. Each image is in either ***.jpg*** or ***.png*** format with varying resolutions. To ensure uniform input for the model, all images are resized to 224×224 pixels, which is compatible with the requirements of the Deep Learning architecture used, particularly MobileNetV2.

The dataset is divided into two main parts:

Training data: 80% of the total dataset (approximately 2,024 images), used to train the model.

Validation data**:** 20% of the total dataset (approximately 503 images), used to evaluate model performance during training.

The data splitting process is performed automatically using a validation split parameter in the

ImageDataGenerator, ensuring a balanced distribution across classes.

Furthermore, to increase data variability and reduce the risk of overfitting, data augmentation techniques are applied to the training data. These include rotation, translation, scaling (zoom), horizontal flipping, and brightness adjustment. These techniques aim to simulate various real-world image conditions.

With diverse dataset characteristics and appropriate preprocessing, the resulting model is expected to have good generalization capability in classifying different types of waste.

2.4.1 Data Augmentation Strategy

To improve model robustness and reduce overfitting, data augmentation techniques were applied to the training dataset using the *ImageDataGenerator* framework. The augmentation process was designed to simulate various real-world waste image conditions, including variations in object orientation, lighting, and scale.

The augmentation techniques applied in this study are summarized in Table 2.

Table 2. Data Augmentation Configuration

Technique	Parameter
Rotation Range	20°
Zoom Range	0.2
Width Shift Range	0.2
Height Shift Range	0.2
Horizontal Flip	True
Brightness Range	0.8–1.2
Rescaling	1/255

These augmentation techniques help improve the diversity of the training data and enable the model to better generalize to unseen data. Rotation and shifting were applied to simulate changes in object positioning, while brightness adjustment was used to represent varying illumination conditions commonly found in real-world environments. Furthermore, horizontal flipping and zooming were introduced to increase variability in object appearance and scale.

2.5 Model Architecture

2.5.1 Convolutional Neural Network (CNN)

The Convolutional Neural Network (CNN) model used in this study is a basic architecture designed to automatically extract features from input images. The CNN architecture consists of several main layers arranged sequentially as follows:

Input Layer. The input images have a size of 224×224×3 (RGB), which conforms to the standard requirements of Deep Learning models.

Convolutional Layer. This layer functions to extract features from images using filters or kernels. In this study, several convolutional layers are used with a kernel size of 3×3 and ReLU (Rectified Linear Unit) activation to enhance the model's non-linear capability.

Pooling Layer. After the convolution process, MaxPooling with a size of 2×2 is applied to reduce the feature dimensions (downsampling) and decrease computational complexity. This process also helps reduce the risk of overfitting.

Feature Extraction (Stacked Layers). A combination of convolutional and pooling layers is applied in a stacked manner to produce more complex and abstract feature representations from the input images.

Flatten Layer. The extracted features are transformed into a one-dimensional vector so that they can be processed by the fully connected layers.

Fully Connected Layer (Dense Layer). This layer performs classification based on the extracted features. A dense layer with ReLU activation is used.

Dropout Layer. To reduce overfitting, dropout is applied with a certain rate (e.g., 0.5–0.7), which randomly deactivates a portion of neurons during training.

Output Layer. The final layer uses the Softmax activation function to produce classification probabilities across six waste categories. The CNN architecture consists of several main stages, namely the input layer, convolutional layers for feature extraction, pooling layers for dimensionality reduction, a flatten layer for data transformation, and fully connected layers for classification. This model is used as a baseline in this study, as illustrated in Figure 3.

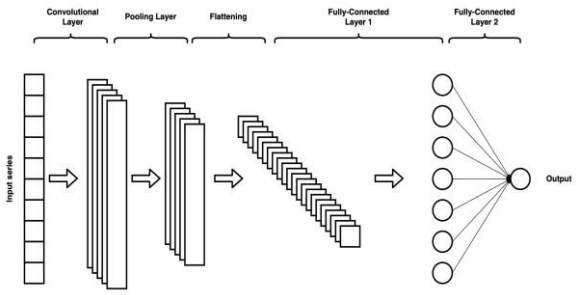


Figure 3. CNN Architecture

2.5.2 MobileNetV2

The MobileNetV2 model is used in this study as the main model due to its advantages in computational efficiency and lightweight architecture. MobileNetV2 is designed using the concept of depthwise separable convolution, which reduces the number of parameters and computational cost compared to conventional CNNs.

In this study, MobileNetV2 is implemented using a transfer learning approach by utilizing pretrained weights from the ImageNet dataset. This approach enables the model to leverage previously learned basic features such as edges, textures, and other visual patterns.

The MobileNetV2 architecture in this study consists of several main components:

Base Model (Feature Extractor). The main part of MobileNetV2 is used as a feature extractor with pretrained weights. In the initial stage, most layers are frozen to preserve the model's prior knowledge.

Global Average Pooling Layer. This layer is used to reduce feature dimensions without significantly increasing the number of parameters, while still retaining important information.

Fully Connected Layer (Dense Layer). Added to adapt the model to the number of classes in the dataset.

Output Layer. Uses the Softmax activation function to produce classification probabilities across six classes.

In this study, transfer learning was implemented using pretrained ImageNet weights. The pretrained layers were kept frozen during training to preserve previously learned visual representations while reducing computational complexity.

MobileNetV2 utilizes the concepts of depthwise separable convolution and inverted residual blocks, allowing the model to be more lightweight and efficient. This model is implemented using transfer learning and with frozen pretrained weights to improve classification performance, as illustrated in Figure 4.

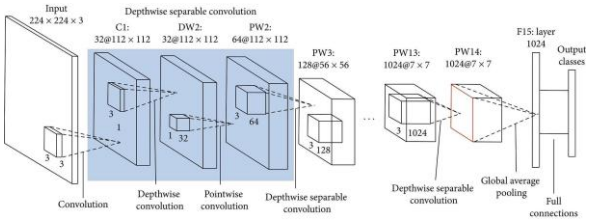


Figure 4. MobileNetV2 architecture used as the pretrained feature extractor in this study.

As illustrated in Figure 4, MobileNetV2 employs depthwise separable convolution and inverted residual blocks to reduce computational complexity while maintaining feature extraction capability. In this study, MobileNetV2 was implemented using transfer learning with pretrained ImageNet weights, while the pretrained layers remained frozen during training.

2.6 Experimental Configuration

The experimental setup was designed to evaluate and compare the performance of CNN and MobileNetV2 models for waste classification. Both models used input images resized to 224×224 pixels and were trained using the same training and validation datasets generated through the ImageDataGenerator framework with an 80:20 split ratio.

The CNN model consisted of three convolutional layers with 32, 64, and 128 filters, respectively, followed by MaxPooling layers for dimensionality reduction. A fully connected layer containing 128 neurons and a dropout rate of 0.5 were applied to reduce overfitting. The CNN model was trained using the Adam optimizer with a batch size of 32 for 10 epochs.

For MobileNetV2, transfer learning was implemented using pretrained ImageNet weights. The pretrained layers were kept frozen during training to preserve previously learned visual features. A Global Average Pooling layer, a dense layer with 128 neurons, and a dropout rate of 0.7 were added to adapt the model to the waste classification task. The MobileNetV2 model was trained using the Adam optimizer with a learning rate of 0.0001 and a batch size of 32 for 15 epochs.

To improve training stability and reduce overfitting, an early stopping mechanism was applied with a patience value of 3 epochs. The training process automatically restored the best-performing model weights based on validation performance.

The experimental configuration used in this study is summarized in Table 3. Both CNN and MobileNetV2 models were trained using the same dataset partition and preprocessing pipeline to ensure a fair comparison.

Table 3. Experimental Configuration

Parameter	CNN	MobileNetV2
Input Size	224×224	224×224
Batch Size	32	32
Epoch	10	15
Optimizer	Adam	Adam
Learning Rate	Default Adam	0.0001
Dropout	0.5	0.7

Parameter	CNN	MobileNetV2
Transfer Learning	No	Yes
Pretrained Weights	No	ImageNet
Early Stopping	Yes	Yes
Patience	3	3

Overall, the selected configuration was designed to balance classification performance, computational efficiency, and model generalization capability. The same experimental settings were maintained throughout the study to ensure a consistent comparison between CNN and MobileNetV2.

III. Result and Discussions

Experimental Results

This study evaluated the performance of two deep learning models, namely CNN and MobileNetV2, for image-based waste classification. Both models were trained using the same dataset, preprocessing procedures, and data augmentation strategies to ensure a fair comparison.

Table 4. Model training and validation Accuracy of CNN and MobileNetV2

Model	Training Accuracy	Validation Accuracy
CNN	54%	48.5%
MobileNetV2	70%	69.8%

Table 4 presents the training and validation accuracy achieved by each model. The results indicate that MobileNetV2 outperformed CNN, achieving a validation accuracy of 69.8%, whereas CNN obtained 48.5%. This finding demonstrates the effectiveness of transfer learning in improving classification performance.

To provide a more comprehensive evaluation, additional performance metrics were calculated for the MobileNetV2 model, including precision, recall, and F1-score. The detailed results are presented in Table 5 and discussed in Section 3.3.. The MobileNetV2 model achieved an overall accuracy of 71.97%, with a precision of 73%, recall of 72%, and F1-score of 71%. These results indicate that MobileNetV2 is capable of maintaining balanced classification performance across different waste categories while achieving reliable prediction accuracy.

The difference between the validation accuracy reported in Table 4 and the overall accuracy presented in Table 5 is attributed to the final evaluation procedure performed on the validation dataset after model training

The training and validation accuracy curves of both models are illustrated in Figure 5 , which shows the learning behavior of the models throughout the training process.

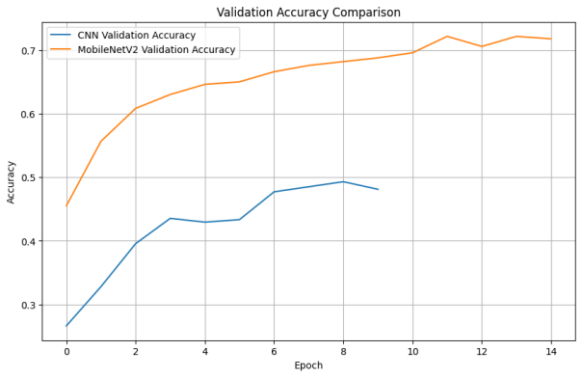


Figure 5. Comparison of validation accuracy between the CNN and MobileNetV2 models during the training process.

As shown in Figure 5, MobileNetV2 consistently achieved higher validation accuracy than CNN throughout the training process. The validation accuracy of MobileNetV2 increased steadily and stabilized at approximately 70%, whereas CNN showed slower improvement and reached a maximum validation accuracy of approximately 49%. This result further confirms the effectiveness of transfer learning in improving feature extraction and classification performance for waste classification tasks.

3.2 Model Performance Analysis

Based on the experimental results, MobileNetV2 demonstrated superior performance compared to the conventional CNN model in waste classification tasks. As shown in Table 4, MobileNetV2 achieved a validation accuracy of 69.8%, significantly outperforming CNN, which obtained 48.5%.

The lower performance of CNN may be attributed to its limited capability in extracting complex visual features from waste images, particularly when trained on a relatively small dataset. Furthermore, the CNN model exhibited indications of underfitting, as reflected by the relatively low accuracy observed in both training and validation phases. This suggests that the CNN architecture employed in this study may not have been sufficiently deep or complex to effectively capture discriminative patterns among waste categories.

In contrast, MobileNetV2 achieved better performance due to the use of transfer learning with pretrained ImageNet weights. The pretrained feature extraction mechanism enables MobileNetV2 to leverage previously learned low-level and high-level visual features, such as texture, shape, and edge information, thereby improving classification capability.

Figure 5 further illustrates that MobileNetV2 consistently achieved higher validation accuracy than CNN throughout the training process. The validation accuracy of MobileNetV2 increased steadily and stabilized at approximately 70%, whereas CNN showed slower improvement and reached a maximum validation accuracy of approximately 49%. This result confirms the effectiveness of transfer learning in improving feature extraction and classification performance for waste classification tasks.

3.3 Model Evaluation Metrics

To evaluate the classification performance more comprehensively, additional evaluation metrics including accuracy, precision, recall, and F1-score were calculated for the MobileNetV2 model. These metrics provide a more detailed assessment of model effectiveness beyond validation accuracy alone.

Table 5. Performance Evaluation Metrics of MobileNetV2

Metric	Score
Accuracy	71.97%
Precision	73%
Recall	72%
F1-score	71%

Table 5 presents the evaluation results obtained from the classification report. The MobileNetV2 model achieved an overall accuracy of 71.97%, indicating that the model correctly classified the majority of waste images in the validation dataset. Furthermore, the model obtained a weighted precision of 73%, recall of 72%, and F1-score of 71%.

The relatively balanced values of precision, recall, and F1-score indicate that the model maintained consistent classification performance across different waste categories. The weighted precision of 73% suggests that most predicted classes were correctly identified, while the recall value of 72% indicates that the model successfully detected a large proportion of actual waste categories. In addition, the F1-score of 71% demonstrates a balanced trade-off between precision and recall.

Overall, these results confirm that MobileNetV2 provides reliable classification performance and exhibits good generalization capability for waste classification tasks.

3.4 The Effect of Data Augmentation and Optimization

Data augmentation and optimization techniques played an important role in improving the performance of the waste classification models. Data augmentation increased the diversity of the training dataset by generating modified versions of existing images through transformations such as rotation, translation, zooming, horizontal flipping, and brightness adjustment. As a result, the models were exposed to a wider range of visual variations, enabling them to generalize better to unseen data.

The use of data augmentation helped reduce the risk of overfitting, particularly because the dataset contained a relatively limited number of images. By simulating different real-world conditions, the model became more robust to variations in object orientation, scale, illumination, and background complexity.

In addition to data augmentation, several optimization techniques were applied during training. Dropout layers were incorporated to reduce excessive reliance on specific neurons and improve model generalization. Furthermore, a lower learning rate was used in MobileNetV2 to ensure more stable parameter updates during training. Early stopping was also employed to prevent overfitting by terminating the training process when no significant improvement in validation performance was observed.

The combination of data augmentation and optimization techniques contributed to the improved performance of MobileNetV2, resulting in more stable training behavior and better validation accuracy compared to the CNN model.

3.5 Confusion Matrix Analysis

The confusion matrix was used to evaluate the distribution of model predictions across the six waste categories. Figure 6 presents the confusion matrix of the MobileNetV2 model on the validation dataset.

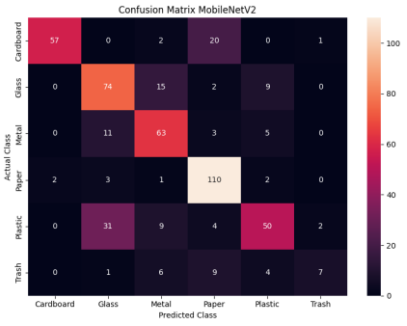


Figure 6. Confusion matrix of the MobileNetV2 model showing the distribution of predicted classes against actual classes in the validation dataset.

As shown in Figure 6, the model achieved strong classification performance for several categories, particularly paper and cardboard, which obtained F1-scores of 0.85 and 0.83, respectively. The high number of correctly classified samples in these categories indicates that the model was able to learn distinctive visual features effectively.

The metal category also demonstrated good performance, achieving a recall of 0.83 and an F1-score of 0.72. However, some misclassifications were observed between visually similar categories such as plastic, glass, and metal, suggesting overlapping visual characteristics among these waste types.

The trash category achieved the lowest performance, with an F1-score of 0.33 and a recall of 0.22. This result may be attributed to the limited number of training samples and the higher visual diversity within the trash category. Consequently, the model encountered greater difficulty in learning representative features for this class.

Overall, the confusion matrix analysis confirms that MobileNetV2 achieved relatively balanced classification performance, with an overall accuracy of 71.97%, weighted precision of 73%, recall of 72%, and F1-score of 71%.

3.6 Discussion

The experimental results demonstrate that MobileNetV2 significantly outperformed the conventional CNN model in waste classification tasks. MobileNetV2 achieved a validation accuracy of 69.8% and an overall evaluation accuracy of 71.97%, whereas CNN achieved a validation accuracy of only 48.5%. These findings indicate that transfer learning is effective in improving classification performance, particularly when working with relatively limited datasets.

The superior performance of MobileNetV2 can be attributed to its pretrained feature extraction capability. By utilizing weights learned from the large-scale ImageNet dataset, MobileNetV2 is able to capture low-level and high-level visual features such as edges, textures, shapes, and object patterns more effectively than a CNN trained from scratch. Consequently, the model exhibits better generalization capability and classification accuracy.

The confusion matrix analysis further revealed that the model performed particularly well in classifying the paper, cardboard, and metal categories. These classes achieved relatively high recall and F1-score values, indicating that their visual characteristics were sufficiently distinctive for the model to learn. In contrast, the trash category achieved the lowest performance, which may be attributed to the limited number of training samples and the greater visual diversity within the class.

In addition, some misclassifications were observed between plastic, glass, and metal categories. These errors may be caused by similarities in texture, color, reflective surfaces, and object appearance, which make it difficult for the model to distinguish between classes. Such challenges are commonly reported in waste classification studies and highlight the need for larger and more diverse datasets.

The implementation of data augmentation and optimization techniques also contributed positively to model performance. Data augmentation increased the diversity of training samples, while dropout and early stopping helped reduce overfitting and improve training stability. As a result, MobileNetV2 achieved a more balanced performance between training and validation data.

The findings obtained in this study are consistent with recent studies that reported the effectiveness of transfer learning and lightweight deep learning architectures for waste classification tasks [18], [20], [23]. Although the accuracy achieved in this study is lower than some state-of-the-art approaches, the proposed framework demonstrates competitive performance while maintaining computational efficiency.

3.7 Limitations

This study has several limitations. First, the dataset size is relatively limited, which may affect the model's ability to learn highly discriminative features. Second, certain classes, particularly trash, contain fewer samples and exhibit greater visual diversity, resulting in lower classification performance. Third, the pretrained MobileNetV2 layers were kept frozen during training, which may limit adaptation to the waste classification dataset.

IV. CONCLUSION

Based on the experimental results, it can be concluded that MobileNetV2 outperformed the conventional CNN model in image-based waste classification. The MobileNetV2 model achieved a validation accuracy of 69.8% and an overall evaluation accuracy of 71.97%, while the CNN model obtained a validation accuracy of 48.5%. These results indicate that transfer learning using pretrained ImageNet weights can significantly improve classification performance compared to a CNN model trained from scratch.

Additional evaluation metrics further confirmed the effectiveness of MobileNetV2, achieving a precision of 73%, recall of 72%, and F1-score of 71%. The confusion matrix analysis showed that the model

performed particularly well in classifying paper, cardboard, and metal categories, while the trash category remained challenging due to limited training samples and greater visual diversity.

The implementation of data augmentation techniques, including rotation, translation, zooming, horizontal flipping, and brightness adjustment, contributed to improving model robustness and generalization capability. Furthermore, optimization strategies such as dropout, learning rate adjustment, and early stopping helped improve training stability and reduce the risk of overfitting.

Overall, the results demonstrate that MobileNetV2 is an effective and computationally efficient approach for waste classification tasks. Future research may focus on expanding the dataset size, addressing class imbalance issues, and exploring advanced transfer learning and fine-tuning strategies to further improve classification performance.

REFERENCES

[1] A. Kumar et al., "Waste Classification Using Deep Learning," Journal of Environmental Informatics, 2022.

[2] S. Rahman et al., "Automated Waste Detection Using CNN," IEEE Access, 2020.

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[4] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4510–4520, 2018.

[5] A. Howard et al., "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," arXiv:1704.04861, 2017.

[6] Z. Zhang et al., "Lightweight CNN Models for Image Classification," 2021.

[7] S. Shorten and T. M. Khoshgoftaar, "A Survey on Image Data Augmentation for Deep Learning," Journal of Big Data, vol. 6, no. 1, pp. 1–48, 2019.

[8] H. Li et al., "Improving Image Classification Using Data Augmentation," 2023.

[9] M. Ahmed et al., "Transfer Learning for Waste Classification," 2024.

[10] J. Brown et al., "Deep Learning Optimization Techniques," 2022.

[11] G. Thung and M. Yang, "Classification of Trash for Recyclability Status," Stanford CS229 Project Report, 2016.

[12] M. Yang and G. Thung, "TrashNet: A Dataset for Image Classification of Waste," Stanford University Technical Report, 2021.

[13] S. R. Dubey, S. K. Singh, and R. K. Chaudhuri, "Activation Functions in Deep Learning: A Comprehensive Survey and Benchmark," Neurocomputing, vol. 503, pp. 92–108, 2022.

[14] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A Survey of the Recent Architectures of Deep Convolutional Neural Networks," Artificial Intelligence Review, vol. 53, no. 8, pp. 5455–5516, 2021.

[15] M. Tan and Q. Le, "EfficientNetV2: Smaller Models and Faster Training," in Proceedings of the International Conference on Machine Learning (ICML), 2021, pp. 10096–10106.

[16] S. V. Militante and N. V. Dionisio, "Deep Learning Implementation for Waste Segregation and Classification System," International Journal of Advanced Computer Science and Applications, vol. 12, no. 2, pp. 232–240, 2021.

[17] R. M. S. Verma and P. Singh, "Transfer Learning-Based Waste Classification Using MobileNetV2," Proceedings of the International Conference on Smart Computing and Electronic Enterprise, 2022.

[18] A. Alqahtani, M. Alsubaie, and M. Alqahtani, "Automatic Waste Classification Using Deep Transfer Learning Models," Sustainability, vol. 15, no. 4, 2023.

[19] J. H. Park and S. Lee, "Lightweight Deep Learning Models for Smart Waste Management Systems," Sensors, vol. 23, no. 10, 2023.

[20] M. A. Islam, M. R. Hossain, and M. S. Rahman, "Comparative Analysis of CNN and Transfer Learning Models for Waste Classification," IEEE Access, vol. 12, pp. 45612–45625, 2024.

[21] P. Sharma and R. Gupta, "Data Augmentation Strategies for Improving Waste Image Classification Performance," Journal of Imaging, vol. 10, no. 3, 2024.

[22] Y. Chen, X. Li, and H. Wang, "Transfer Learning for Resource-Efficient Waste Classification Systems," Expert Systems with Applications, vol. 245, 2024.

[23] S. Ahmed, M. Khan, and F. Ali, "MobileNetV2-Based Waste Classification for Smart Recycling Applications," Applied Sciences, vol. 14, no. 5, 2024.

[24] T. Nguyen, H. Tran, and K. Le, "Optimization Techniques for Deep Learning-Based Waste Classification," Engineering Applications of Artificial Intelligence, vol. 131, 2025.

[25] R. Kumar and V. Patel, "Advanced Deep Learning Approaches for Automated Waste Classification: A Review," IEEE Access, vol. 13, pp. 11245–11267, 2025.

ISSN 2085-4552

Analysis of Public Concerns Towards COVID-19 Vaccines on X Post-Peak Pandemic Phase

Luthfi Atikah[1], Novrindah Alvi Hasanah[2], Binti Kholifah[3]
[1] Management Informatic: Politeknik Astra, Kabupaten Bekasi, Indonesia
[2] Departement of Electrical Engineering: Universitas Islam Maulana Malik Ibrahim, Malang, Indonesia
[3] Management Informatic Universitas Negeri Surabaya, Surabaya, Indonesia
luthfiatikah @polytechnic.astra.ac.id[1], bintikholifah@unesa.ac.id[2], novrindah@uin-malang.ac.id[3]

Accepted 03 June 2026
Approved 28 June 2026

***Abstract*— The rapid development of social media, particularly platform X, has created both opportunities and challenges in public health discourse. This study aims to analyze public concerns and perceptions toward COVID-19 vaccines in the post-peak pandemic phase. A dataset of 4.519 Indonesian-language tweets collected between September 2024, and September 2025 was processed through four stages of preprocessing and subsequently labeled. The labeled data were then evaluated using two experimental scenarios, namely with and without data augmentation, and analyzed using a seven-class sentiment classification framework.The results indicate that negative sentiment overwhelmingly dominates the discourse at 96.29%, with strongly negative sentiment accounting for 91.88%. Meanwhile, neutral sentiment is only 0.36%, and positive sentiment remains relatively low at 3.35%. These findings suggest that public distrust toward COVID-19 vaccines persists even in the post-peak phase of the pandemic. Therefore, more effective, targeted, and evidence-based health communication strategies are necessary to address public concerns and improve vaccine confidence.**

Index Terms*— BERT; IndoBert; COVID-19 Vaccine; Sentiment Analysis; X; Vaccine Hesitancy.

I. INTRODUCTION

In March 2020, the World Health Organization (WHO) declared COVID-19 a global pandemic with widespread impacts on public health, the environment, human psychology, the global socioeconomy, and education [1]; [2]. The surge in mortality due to the virus placed immense pressure on healthcare systems worldwide [3]. Various measures were implemented to curb the uncontrolled spread of the virus, including mobility restrictions. However, these policies led to increased unemployment and reduced household income due to significant losses faced by many companies. Additionally, major shifts in social interaction patterns widened social inequalities and made mental health a critical concern for policymakers. In fact, widespread psychological distress resulted in increased cases of depression globally [4].

Effective pandemic management became a crucial step in maintaining global health system stability. [5] stated that delayed responses and weak healthcare systems contributed to approximately 14.8 million deaths during the pandemic. In response, experts developed various solutions, including vaccination programs as an effective preventive strategy. Several studies highlight vaccination as a key solution to reducing COVID-19 transmission. For instance, [6] demonstrated that vaccines significantly reduce symptomatic infections. Similarly, B [7] reported 94.1% efficacy in preventing COVID-19 in large-scale trials of the Moderna vaccine.

The success of vaccination programs heavily depends on public trust. Vaccine hesitancy, often driven by misinformation and distrust in health authorities, can significantly reduce vaccination coverage. Social media has become a key platform for expressing public concerns regarding vaccination programs [8] Studies also show that hesitancy is not limited to the public but is also present among healthcare workers, with vaccine side effects being a primary concern [9]. Public rejection of government vaccination programs often stems from low trust in health institutions. This is reinforced by [10] who emphasized the importance of trust and confidence in determining vaccine acceptance. Therefore, successful vaccination programs require not only vaccine availability but also comprehensive communication, wide coverage, equitable distribution, and healthcare system preparedness [11].

If these aspects are neglected, public concern will increase, becoming a major issue for health authorities. Social media serves as an effective tool to monitor these concerns. During the pandemic, platforms like X became central to public opinion formation due to their accessibility and compatibility with social distancing measures. However, the rapid spread of information on such platforms also accelerates the dissemination of misinformation. [12], in "The COVID-19 Infodemic: Twitter versus Facebook," showed that social media can amplify misinformation, significantly influencing

public perception and decision-making. Thus, while social media has great potential as a health communication tool, it also requires strategies to mitigate misinformation risks.

Social media data analysis enables real-time measurement of public sentiment, as these platforms serve as primary outlets for expressing opinions and emotions. Twitter is widely used to detect public perception trends and monitor reactions to government policies [13]. However, the exponential growth of digital data presents significant challenges for manual analysis. [14] highlighted that current data volumes exceed human processing capacity, necessitating automated and algorithmic approaches. During the pandemic, millions of daily posts made manual sentiment analysis impractical [15]. Therefore, integrating machine learning approaches becomes essential for handling large-scale data efficiently [16].

Text mining techniques allow researchers to extract valuable insights from massive textual datasets [17]. Sentiment analysis plays a key role in understanding public emotions, with deep learning models achieving high accuracy. Machine learning approaches have also proven effective for real-time monitoring of public sentiment during crises [18]. Consequently, integrating natural language processing (NLP) and machine learning is crucial for efficient and scalable analysis of textual data.

Artificial intelligence (AI), especially NLP, has proven effective in automatically understanding and classifying text. Studies such as [19] and [20] demonstrated high accuracy in text classification tasks using deep learning. Transformer-based models like BERT have significantly advanced NLP by improving semantic and syntactic understanding [21]. These models enable tasks such as topic classification, entity recognition, and sentiment analysis Additionally, large-scale datasets like COVID Senti further demonstrate the effectiveness of NLP in analysing pandemic-related sentiment.

BERT (Bidirectional Encoder Representations from Transformers) is a state-of-the-art NLP model known for its bidirectional context understanding. It outperforms traditional models in capturing nuanced emotions and hidden concerns in text. Studies have shown its effectiveness in analysing COVID-19-related tweets and detecting public concerns [22]. Despite its strong performance, previous studies in Indonesia have mostly focused on simple three-class sentiment classification or non-Indonesian datasets.

Based on this gap, this study focuses on analysing public sentiment toward COVID-19 vaccines in the post-pandemic phase using BERT, by classifying sentiment into positive, negative, and neutral categories. This research aims to provide deeper insights into public perception and emotional responses toward vaccination programs in Indonesia, particularly as expressed on Twitter (X).

II. METHOD

A. Reserch Design

At this stage, several data preparation steps are carried out based on Figure 1. Including data collection, Preprocessing, Deep Learning based classification, and Evaluation. The figure illustrates the overall workflow of this study.

This study begins with data collection from X (formerly Twitter) over a one-year period, spanning from September 2024 to September 2025. Some of the collected raw data were preprocessed before proceeding to the next stage, including the removal of irrelevant components. After the preprocessing stage, the dataset was manually labeled into seven main

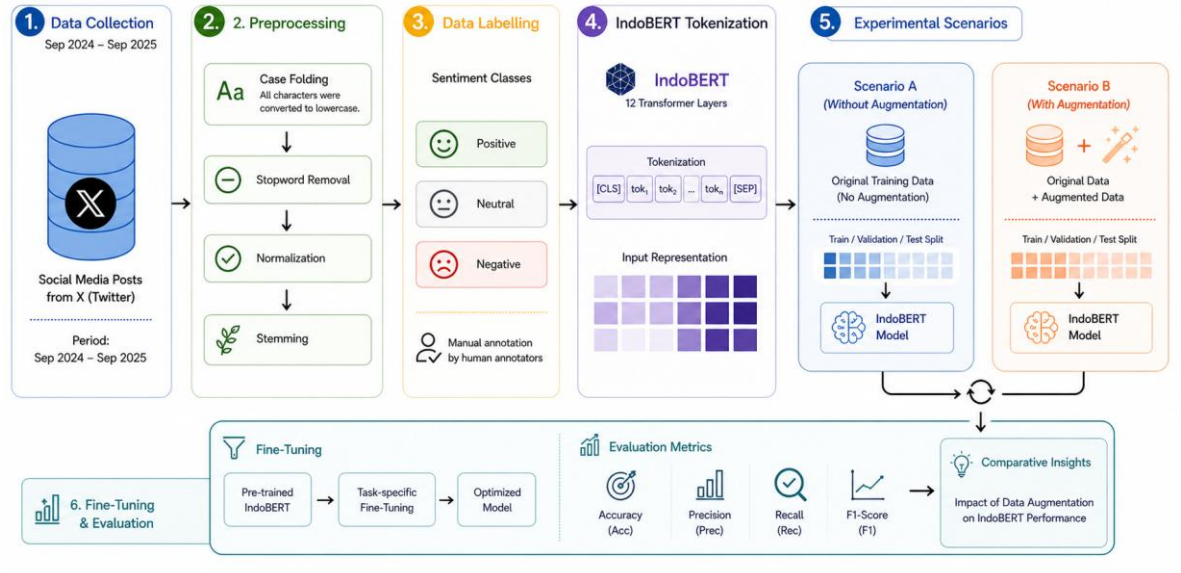


Figure 1. Scenario Research Pipeline

classes: strongly negative, negative, weakly negative, neutral, weakly positive, positive, and strongly positive.

The labeled data are then utilized as input for the IndoBERT model, a pre-trained transformer-based language model designed to capture deep contextual representations of Indonesian text. The output of the IndoBERT model is obtained as a pooled representation with a dimensionality of 768. This representation is then passed through a dropout layer to reduce overfitting, followed by a softmax layer to generate probability distributions over the target classes.

The model was evaluated using several evaluation metrics, including accuracy, precision, recall, and F1-score. The sentiment results were classified into seven classes: strongly negative, negative, weakly negative, neutral, weakly positive, positive, and strongly positive.

The classification into seven sentiment classes in this study was intended to capture the intensity of sentiments expressed regarding the COVID-19 vaccine, thereby providing a more optimal representation of sentiment analysis. For example, tweets expressing mild hesitation have a different representation from tweets expressing strong rejection, even though both would generally be categorized as negative in a three-class classification scheme.

These seven categories were developed from the original three-class manual annotations through a probability-based mapping approach. The mapping process from the three initial classes was further divided into strongly negative (softmax probability > 0.85), negative (0.65–0.85), and weakly negative (0.50–0.65). For neutral tweets, the confidence score threshold was set at > 0.70, while tweets with lower confidence scores were categorized as either weakly positive or weakly negative. A similar approach was also applied to the positive class. The seven-class categorization adopted in this study was based on previous research on fine-grained sentiment analysis [1]. A detailed explanation of each step is provided in the following section.

B. Data Collection

Data were collected from X (formerly Twitter) over one year, spanning from September 2024 to September 2025. The data collection process was conducted using the Twitter Academic Research API v2, which enables systematic and large-scale retrieval of publicly available tweets for research purposes. This approach ensures that the data acquisition process is both structured and reproducible.

During the data collection process, a total of 4,519 tweets in the Indonesian language were collected. The tweets were then filtered based on predefined keywords. In addition to the tweet text, the dataset also consisted of several important attributes, including publication date, number of likes, number of retweets, and other related metadata. Furthermore, bot and spam filtering was conducted by manually removing irrelevant tweets. This stage was carried out to reduce noise and improve the representativeness of the dataset.

All collected data were stored in Comma-Separated Value (CSV) format to facilitate efficient data handling, processing, and reproducibility of the study. The use of a standardized data format also ensures compatibility with various data analysis tools and machine learning frameworks.

Table 1. Keyword Categories

Theme	Keywords (Indonesian)
Side Effects (Efek Samping)	efek samping, reaksi alergi, sakit kepala, demam tinggi
Long-term Effects (Efek Jangka Panjang)	infertilitas, autoimun, risiko kanker
Safety (Keamanan)	risiko kesehatan, anak-anak dan ibu hamil
Efficacy (Efektivitas)	efektif atau tidak, varian baru, perlindungan menurun
Distribution (Distribusi)	ketidakadilan, akses terbatas, negara berkembang
New Technology (Teknologi Baru)	mRNA, teknologi belum teruji, manipulasi genetik

C. Preprocessing

The raw data obtained through the crawling technique are then preprocessed to improve data quality before being used in the modeling process. This preprocessing stage aims to remove noise, simplify text representation, and ensure data consistency, thereby improving the model's ability to understand language context.

In this study, preprocessing consists of four main stages: case folding, stopword removal, normalization, and stemming. In the case folding stage, all characters in the text are converted to lowercase to avoid differences in word representation due to capitalization. Next, in the stopword removal stage, common words that do not contribute meaningful information to the sentiment analysis process, such as conjunctions and prepositions, are removed from the text.

The next stage is normalization, which aims to convert non-standard words, such as slang, abbreviations, or informal spelling variations, into their standard forms in accordance with Indonesian language rules. This is followed by stemming, which involves reducing affixed words to their base forms using the Sastrawi stemmer algorithm, allowing morphological variations of a word to be represented uniformly.

In addition to these four main steps, further cleaning is performed by removing punctuation, URLs, and mention tags that do not contribute to the sentiment analysis. After completing all preprocessing steps, the dataset is reduced from 4,519 tweets to 3,815 tweets, making it ready for use in the subsequent modeling stage.

D. Data labeling

The model was evaluated using several evaluation metrics, including accuracy, precision, recall, and F1-score. The sentiment results were classified into seven categories: strongly negative, negative, weakly negative, neutral, weakly positive, positive, and strongly positive. The classification into seven sentiment categories in this study was intended to capture the intensity of sentiments expressed regarding the COVID-19 vaccine, thereby providing a more optimal representation of sentiment analysis. For example, tweets expressing mild doubt have a different representation from tweets expressing strong rejection, even though both would generally be categorized as negative in a three-class sentiment scheme.

These seven categories were developed from three manually annotated classes through probability-based mapping. The mapping process grouped the three classes into strongly negative (softmax probability > 0.85), negative (0.65–0.85), and weakly negative (0.50–0.65). For neutral tweets, the confidence score threshold was set at > 0.70, while tweets that did not meet this threshold were categorized as either weakly positive or weakly negative. A similar approach was applied to the positive class. The seven sentiment categories were adopted based on previous research in fine-grained sentiment analysis [1].

Table 2 presents representative examples from each sentiment class, illustrating the characteristics and linguistic patterns associated with positive, neutral, and negative labels in the dataset.

Table 2. Sample Data Labeled

Sentiment Label	Example Tweet (Original Indonesian)
Negative	imun gw emg agak kureng tapi semenjak gw vaksin covid gw kena neurodermatitis tiba2 bete bgt.
Neutral	Vaksin covid dan alat2 testnya cba cek skalian
Positive	Covid di RI lu dapet vaksin China, aman.

E. Model Architecture and Classification

Sentiment classification in this study was performed using the IndoBERT model (indobenchmark/indobert-base-p1) [19], a transformer-based language model pre-trained to understand the context of the Indonesian language. In the initial stage, tokenization was performed to split tweet text into subword tokens, with a classification token [CLS] added at the beginning of each sequence to represent the overall text.

Next, a padding process was applied to standardize the input length to a maximum of 128 tokens. If a sequence exceeded this limit, truncation was performed to ensure it conformed to the specified maximum length.

In this study, two experimental scenarios were conducted: data augmentation and non-data augmentation. These scenarios were designed to address the issue of class imbalance in the dataset. The dataset was highly imbalanced, with 94% of the tweets labeled as negative. Data augmentation was applied to the minority classes, namely positive and neutral, to improve the model's generalization capability. The augmentation techniques employed included synonym replacement, random insertion, random swap, and random deletion based on the approach proposed in [21]. For each tweet in the minority classes, three augmented versions were generated, resulting in an augmentation ratio of 1:4. The model output generated probabilities for seven sentiment classes, namely strongly negative, negative, weakly negative, neutral, weakly positive, positive, and strongly positive. The final sentiment label was determined based on the highest probability score. The overall stages of the classification process are illustrated in Figure 2.

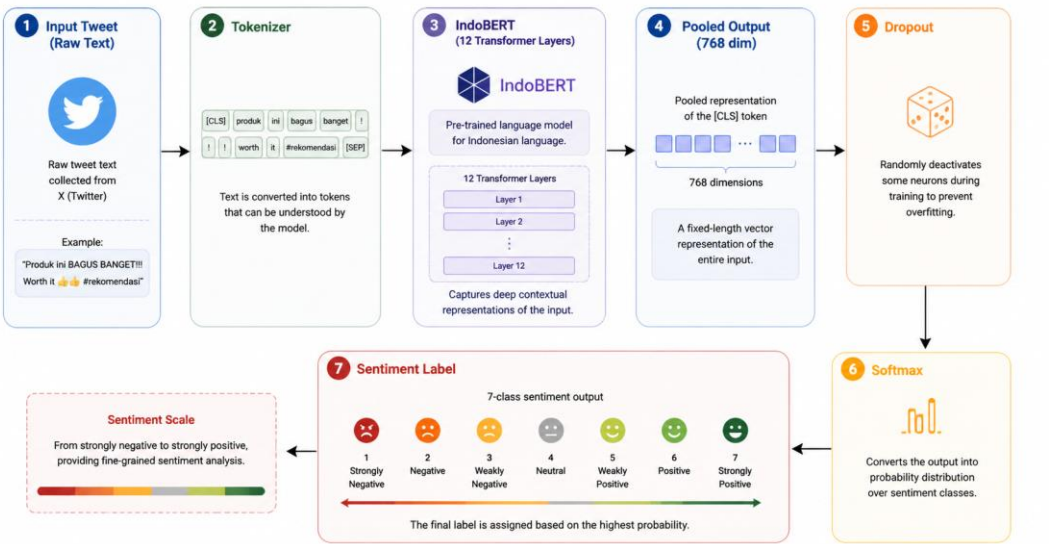


Figure 2. IndoBERT-Based Sentiment Classification Architecture

F. Evaluation Metric

During the evaluation stage, the model was assessed using several evaluation metrics, including accuracy, precision, recall, and F1-score. Accuracy was used to measure the proportion of correct predictions over the entire dataset, while precision and recall were used to evaluate the model's ability to classify data accurately and comprehensively. The F1-score served as a combined metric that balances precision and recall. In addition, a confusion matrix was employed to provide a more in-depth analysis of the classification results in order to identify the strengths and weaknesses of the model for each class.

III. RESULT AND DISCUSSION

In the results and discussion stage, the model was analyzed through the following aspects:

A. Dataset Statistic

After the preprocessing stage, a total of 3,366 tweets were obtained. The preprocessed tweets were then labeled using a semi-supervised approach and categorized into seven sentiment classes: strongly negative, negative, weakly negative, neutral, weakly positive, positive, and strongly positive.

The dataset distribution showed a significant class imbalance, with approximately 90% of the tweets classified as negative sentiment. This imbalance indicates that public opinion was predominantly characterized by concerns, skepticism, and doubts regarding vaccination. The imbalance in label distribution may affect the model's performance; therefore, more in-depth model training and evaluation are required, particularly for the minority classes.

Table 3. Primary Sentimen Distribution

Sentiment Categories	Total Tweets	Percentage (%)
Positive	178	5,29
Neutral	19	0,56
Negative	3.169	94,14
Total	3.366	100,00

B. Sentiment Classification

In the sentiment analysis stage, classification was performed using the IndoBERT model. The results of the classification are presented in Table 4. The IndoBERT model calculated the probability for each predefined sentiment class. The final distribution results were then used to determine the sentiment tendency within each class.

Table 4. Seven Class Sentiment Distribution

Sentiment Categories	Percentage (%)
Strongly Negative	91,88
Negative	3,09
Slightly Negative	1,32
Neutral	0,36
Slightly Positive	0,79
Positive	0,51
Stongly Positive	2,05

Based on the analysis Table 4, it was found that the majority of tweets were classified as negative sentiment. These tweets were dominated by six major issues, namely vaccine side effects, long-term health risks, vaccine safety, the effectiveness of vaccines against new variants, vaccine distribution to remote areas, and concerns regarding mRNA technology. The analysis results indicate that public concern toward vaccination was relatively high.

C. Discussion

Based on the confusion matrix analysis in Scenario B (using data augmentation), the model demonstrated a good classification capability. This was indicated by the successful prediction of Tabel 5 testing data instances correctly. However, the model still experienced difficulties in distinguishing between the negative and weakly negative classes, as well as between the positive and strongly positive classes.

The implementation of data augmentation in Scenario B was considered effective in improving the classification performance for minority classes compared to Scenario A. This can be observed from the increase in the F1-score from 0.9487 to 0.9512, as well as the improvement in overall accuracy from 96.13% to 96.48%, as illustrated in Figure 4.

Based on the experimental results, the sentiment distribution was dominated by negative opinions, accounting for 91.88% of all tweets. The high proportion of negative sentiment in this study may have been influenced by several factors. First, the collected dataset was obtained during the peak phase of the pandemic (September 2024 – September 2025), when vaccination mandates had already ended, leading to an increase in criticism and negative sentiment. Second, the labeling scheme using seven sentiment classes allowed tweets containing negative expressions to be detected more easily compared to neutral categories. Third, numerous negative expressions emerged following reports of vaccine side effects, which further strengthened the spread of negative sentiment within society.

Based on the comparison among the three sentiment categories, a significant difference was observed between the strongly negative sentiment class, which accounted for 91.88%, and the negative sentiment class, which accounted for 1.32%. This indicates that the public not only expressed negative opinions but also showed strong rejection toward vaccination. In contrast, the proportion of positive sentiment in this study was only 3.35%, which is considered significantly lower compared to the positive sentiment proportion reported in previous studies, ranging from approximately 20% to 35%. This indicates the

existence of systematic differences in public trust toward vaccination.

Table 5. Model Performance Evaluation Result

Metric	Scenario A (Without Augmentation)		Scenario B (With Augmentation)	
Class	Precission	Recall	Precission	Recall
Strongly Negative	0.97	0.99	0.97	0.99
Negative	0.41	0.20	0.55	0.35
Slightly Negative	0.38	0.17	0.52	0.30
Neutral	0.00	0.00	0.20	0.14
Slightly Positive	0.33	0.19	0.48	0.32
Positive	0.44	0.23	0.61	0.40
Strongly Positive	0.50	0.31	0.64	0.44
Accuracy	96.13 %		96.48%	

Based on the analysis, this study still has several limitations, particularly the high level of dataset imbalance, which may introduce bias during both the labeling process and model prediction. Therefore, future studies are expected to utilize larger datasets and include more diverse data variations in order to reduce label distribution imbalance.

IV. CONCLUSION

This study shows that COVID-19 vaccination remains a topic of discussion, particularly after the pandemic situation began to subside. In this research, sentiment analysis was conducted on Indonesian-language tweets using the IndoBERT method. The results indicated that public conversations were dominated by negative sentiment, accounting for 96.29% of the overall sentiment classes. Tweets within the negative sentiment categories reflected various emotional expressions, including fear, skepticism, criticism toward vaccination, concerns about vaccine side effects, uncertainty regarding the long-term impact of vaccination, and distrust toward mRNA technology. However, this study also encountered the issue of data imbalance; therefore, data augmentation was applied to the dataset used for modeling. Future studies are expected to utilize larger and more diverse datasets in order to improve model performance and reduce dataset imbalance

REFERENCES

A. Kumar et al., "Waste Classification Using Deep Learning," V. P. Kalanjati *et al.*, "Sentiment analysis of Indonesian tweets on COVID-19 and COVID-19 vaccinations," *F1000Res.*, vol. 12, p. 1007, Aug. 2023, doi: 10.12688/f1000research.130610.1.

[2] Y. Miyah, M. Benjelloun, S. Lairini, and A. Lahrichi, "COVID-19 Impact on Public Health, Environment, Human Psychology, Global Socioeconomy, and Education," 2022, *Hindawi Limited*. doi: 10.1155/2022/5578284.

[3] H. Wang *et al.*, "Estimating excess mortality due to the COVID-19 pandemic: a systematic analysis of COVID-19-related mortality, 2020–21," *The Lancet*, vol. 399, no. 10334, pp. 1513–1536, Apr. 2022, doi: 10.1016/S0140-6736(21)02796-3.

[4] H. Offi cials and S. Galea, "Public Health COVID-19 Impact Assessment: Lessons Learned and Compelling Needs About the NAM series on Emerging Stronger After," 2021. [Online]. Available: https://astho.org/Research/Data-and-Analysis/State-and-Local-Governance-

[5] O. J. Watson, G. Barnsley, J. Toor, A. B. Hogan, P. Winskill, and A. C. Ghani, "Global impact of the first year of COVID-19 vaccination: a mathematical modelling study," *Lancet Infect. Dis.*, vol. 22, no. 9, pp. 1293–1302, Sep. 2022, doi: 10.1016/S1473-3099(22)00320-6.

[6] E. J. Haas *et al.*, "Impact and effectiveness of mRNA BNT162b2 vaccine against SARS-CoV-2 infections and COVID-19 cases, hospitalisations, and deaths following a nationwide vaccination campaign in Israel: an observational study using national surveillance data," *The Lancet*, vol. 397, no. 10287, pp. 1819–1829, May 2021, doi: 10.1016/S0140-6736(21)00947-8.

[7] L. R. Baden *et al.*, "Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine," *New England Journal of Medicine*, vol. 384, no. 5, pp. 403–416, Feb. 2021, doi: 10.1056/nejmoa2035389.

[8] S. L. Wilson and C. Wiysonge, "Social media and vaccine hesitancy," *BMJ Glob. Health*, vol. 5, no. 10, Oct. 2020, doi: 10.1136/bmjgh-2020-004206.

[9] A. A. Dror *et al.*, "Vaccine hesitancy: the next challenge in the fight against COVID-19," *Eur. J. Epidemiol.*, vol. 35, no. 8, pp. 775–779, Aug. 2020, doi: 10.1007/s10654-020-00671-y.

[10] L. S. Rozek, P. Jones, A. Menon, A. Hicken, S. Apsley, and E. J. King, "Understanding Vaccine Hesitancy in the Context of COVID-19: The Role of Trust and Confidence in a Seventeen-Country Survey," *Int. J. Public Health*, vol. 66, p. 636255, 2021, doi: 10.3389/ijph.2021.636255.

[11] S. L. Wilson and C. Wiysonge, "Social media and vaccine hesitancy," *BMJ Glob. Health*, vol. 5, no. 10, pp. 1–7, 2020, doi: 10.1136/bmjgh-2020-004206.

[12] M. Cinelli *et al.*, "The COVID-19 social media infodemic," *Sci. Rep.*, vol. 10, no. 1, Dec. 2020, doi: 10.1038/s41598-020-73510-5.

[13] N. El-Bassel, K. R. Hochstatter, M. N. Slavin, C. Yang, Y. Zhang, and S. Muresan, "Harnessing the Power of Social Media to Understand the Impact of COVID-19 on People Who Use Drugs During Lockdown and Social Distancing," *J. Addict. Med.*, vol. 16, no. 2, pp. E123–E132, Mar. 2022, doi: 10.1097/ADM.0000000000000883.

[14] Viktor. Mayer-Schönberger and Kenneth. Cukier, *Big data : a revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt, 2013.

[15] S. Sharma, A. Kundu, S. Basu, N. P. Shetti, and T. M. Aminabhavi, "Sustainable environmental management and related biofuel technologies," Nov. 01, 2020, *Academic Press*. doi: 10.1016/j.jenvman.2020.111096.

[16] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep learning applications and challenges in big data analytics," *J. Big Data*, vol. 2, no. 1, Dec. 2015, doi: 10.1186/s40537-014-0007-7.

[17] X. Dou *et al.*, "Research progress on chitosan and its derivatives in the fields of corrosion inhibition and

antimicrobial activity," May 01, 2024, *Springer*. doi: 10.1007/s11356-024-33351-5.
[18] "icns_20_1-3_53".
[19] N. A. Hasanah, N. Suciati, and D. Purwitasari, "Identifying degree-of-concern on covid-19 topics with text classification of twitters," *Register: Jurnal Ilmiah Teknologi Sistem Informasi*, vol. 7, no. 1, pp. 50–62, 2021, doi: 10.26594/register.v7i1.2234.
[20] L. Atikah, D. Purwitasari, and N. Suciati, "DETEKSI KEJADIAN LALU LINTAS PADA TEKS TWITTER DENGAN PENDEKATAN KLASIFIKASI MULTI-LABEL BERBASIS DEEP LEARNING", doi: 10.25126/jtiik.202295206.
[21] W. J. Li, "Letter for Qiu et al. (2021) regarding 'The distribution and behavioral characteristics of plateau pikas (Ochotona curzoniae),'" 2022, *Pensoft Publishers*. doi: 10.3897/zookeys.1102.79148.
[22] F. Abbas and A. Taeihagh, "Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence," Oct. 15, 2024, *Elsevier Ltd*. doi: 10.1016/j.eswa.2024.124260.

ISSN 2085-4552

Decision Support System Using the Weight Product (WP) Method for Selecting the Best Employee Performance

Budi Permana Putra[1], Eka Sofiati[2], Muhammad Harits Pratama[3], Tomy Nanda Putra[4]
[1,2] Dept. of Information System, Mitra Gama Institute of Technology, Duri, Indonesia
[3,4] Dept. of Information Technology, Mitra Gama Institute of Technology, Duri, Indonesia
[2]ekasfti@gmail.com

Accepted 29 May 2026
Approved 28 June 2026

Abstract— A company cannot be separated from the role of Human Resources (HR) who work in it. The success of a company is supported by employees who can work optimally and optimally, in this case the company must motivate or encourage employees to work optimally and optimally. The problem that often occurs in a company is that in making decisions in the performance selection process, the best employees are needed. For this reason, a fast and accurate decision-making process is needed, so that it can help managers in making decisions. In supporting the **selection of the best employee performance, the method used is the Weight Product (WP) method. Decision Support System (DSS) is an information system that is specifically intended to assist management in making decisions related to issues that are structured and semi-structured. Weighted product (WP) is a decision making method that is quite widely used. The Weighted Product (WP) method uses multiplication to connect attribute ratings, where the rating of each attribute is raised to the first power by the weight of the attribute in question. There are several steps in the Weigh Product method, namely determining criteria, determining suitability ratings for each criterion, normalizing or improving weights, determining vector S values and determining vector V values. In this study, 15 alternative data and 5 criteria data were used, namely attendance, behavior, loyalty. , neatness and performance. in this study produced alternative A-13 as the best alternative with a value of 0.09. In this study it can be concluded that the Weight Product (WP) method can be implemented in a case study of selecting the best employee performance.**

Index Terms— DSS; Weight Product; Employee; PHP; MYSQL.

I. INTRODUCTION

Information technology processes involve data processing, including collecting, organizing, storing, and manipulating data to obtain high-quality information [1]. High-quality information refers to information that is relevant, accurate, and timely, used for personal, business, and government purposes, as well as for decision-making [2]. The development of information technology is currently quite rapid; one example is the existence of communication tools using networks that can span the entire globe, whereas initially, technology was used only for writing, creating graphics, and storing data [3]. Competition among companies is a natural occurrence; this drives companies to compete in improving their quality, requiring them to continuously make improvements—one of which is by enhancing the quality of their employees [4]. Within a company, the role of Human Resources (HR) is indispensable [5]. A company's success is supported by employees who can work optimally and to their full potential [6]; in this regard, companies must motivate or encourage employees to work optimally and to their full potential [7]. Employees or the workforce play a crucial role within a company [8], as employees are the key factor determining a company's success [9]. Without employees, a company would not survive for long and would decline [10].

Rapid technological advancements have enabled many creative and innovative individuals to make significant strides in streamlining various fields, particularly in the workplace [11]. An increasing number of decision-making processes are being assisted or taken over by AI [12] in fields such as warfare, finance, law, healthcare, and insurance. Decision Support Systems have been introduced to improve the selection of the best employees by matching individual characteristics with predetermined criteria and providing assessments to support the company in making decisions [13].

Research on Decision Support Systems using the Weighted Product Method for selecting the best employees demonstrates that a Decision Support System is an information system specifically designed to assist management in making decisions related to structured and semi-structured problems [14]. Decision

Support Systems enable decision-makers to make effective and efficient decisions.

Another study examined a comparative analysis of Decision Support Systems using the Simple Additive Weighting and Weighted Product methods in the performance evaluation of contract workers [15]. The conclusion of this study is that Decision Support Systems employing the Simple Additive Weighting and Weighted Product methods can be applied in the performance evaluation of contract workers using predetermined criteria and weights. In this study, the Simple Additive Weighting and Weighted Product methods produced the same ranking order of alternatives or contract workers.

II. METHOD

A research framework is designed to assist researchers in conducting their studies, achieving the intended results, and ensuring that the research report is easy to understand. A research framework is necessary to ensure that the final research results align with expectations or desired outcomes, as illustrated in Figure 1.

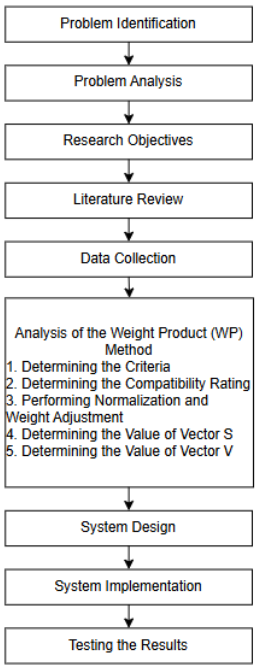


Fig. 1. Research Framework

A. *Problem identification*

Problem identification can be carried out by examining the existing object, starting with how that object can effectively solve the problem. The problem addressed in this study is how PT. Family Raya CRF can determine the selection of top-performing employees. Consequently, a Decision Support System using the Weight Product method is required.

B. *Problem Analysis*

At this stage, the researcher will conduct a problem analysis. This analysis serves to gain a deeper understanding of the problems identified in the previous stage.

C. *Research Objectives*

At this stage, the objectives of the study are defined. This study aims to address issues related to selecting top-performing employees.

D. *Literature Review*

In this study, the researcher identified and analyzed the relevant literature. The literature used in this study consists of scientific articles on decision support systems that discuss the Weight Product (WP) method.

E. *Data Collection*

In this study, several methods were employed to achieve the desired results. The data collection methods used were as follows:

a) Field Research
Field research involves researchers directly collecting data from the research subjects.

b) Observasi
The author conducted an interview with the owner of PT. Family Raya CRF to gather information about employee performance.

c) Library Research
This method involves reading books, literature, and journals related to the research topics to serve as reference material.

d) Laboratory Research
This research method involves analyzing data and designing programs or software tailored to the specific problems at hand. The programs are then tested in practice to verify their effectiveness.

F. *Analysis of the Weight Product Method*

The issues in this study can be analyzed using the second method, namely the WP method, with the aim of determining the criteria and weights for each alternative across all attributes, thereby aiding in the selection of the best-performing employees. The steps of the WP method employed in this study are as follows:

a. Determining the criteria
b. Determining the similarity score
c. Performing normalization and weight adjustment
d. Determining the value of vector S
e. Determining the value of vector V

G. Perancangan Sistem

Following the data analysis and processing using the Weighted Product Method, this stage involves designing a system that meets the criteria and weightings based on the alternatives for all attributes;

the design will be created using the draw.io website and may take the form of a Unified Modeling Language (UML) diagram.

H. System Implementation

The implementation at this stage was designed using the PHP programming language and a MySQL database. The goal is to validate the results of the application design so that users can provide feedback to the application developers.

I. Testing the Results

The results of the testing of the mechanism for data processing can be seen in the following descriptions:

a) Performing manual data processing using Microsoft Excel with the Weight Product (WP) method.

The results of this manual data processing will be compared with the output from the system that has been developed.

III. RESULT AND DISCUSSIONS

The data used in this study consists of employee data collected from PT Family Raya CRF. This data was used as a sample to test the application of the Weight Product method. Data collection was conducted through interviews and direct data collection at the PT Family Raya CRF office. The following employees were included in the sample for this study.

TABLE I. Employee Data

No	Alternative	Attendance	Behavior	Loyalty	Neatness	Performance
1	Fery Andany	Very Poor	Fair	Fair	Good	Very Good
2	Alyusrizal M	Poor	Fair	Very Good	Fair	Fair
3	Amat	Fair	Good	Poor	Fair	Good
4	Budi Hendriyanto	Good	Fair	Good	Very Good	Fair
5	Veronica	Very Good	Fair	Good	Fair	Buruk
6	Hermanto Eka Putra	Poor	Very Good	Fair	Poor	Good
…	…	…	…	…	…	…
…	…	…	…	…	…	…
15	Jekson Putra	Fair	Good	Good	Fair	Very Poor

3.1 Analysis of a Decision Support System Using the Weight Product (WP) Method

This study will analyze the data collected from PT Family Raya CRF Padang. The data will be analyzed using the Weight Product (WP) method to generate a ranking. The Weight Product method involves several analytical steps, including weight normalization or adjustment, determining the value of vector s, and determining the value of vector v. The following outlines the analytical steps of the Weight Product method applied to the case study on selecting the best-performing employees at PT Family Raya CRF. The Weight Product method process is illustrated in Figure 2.

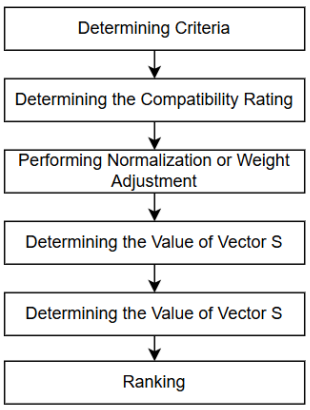


Fig.2. Proses Metode Weight Product

A. Determining Criteria

This study uses five criteria identified through interviews and analysis of data collected at PT Family Raya CRF Padang. The five criteria are attendance, behavior, loyalty, neatness, and performance. These five criteria will be weighted according to their importance. The criteria and their respective weights are presented in Table II.

TABLE II. Criteria and Weightings

No	Criterion Code	Criterion Name	Weight	Category
1	C1	Attendance	0.2	Benefits
2	C2	Behavior	0.2	Benefits
3	C3	Loyalty	0.2	Benefits
4	C4	Neatness	0.2	Benefits
5	C5	Performance	0.2	Benefit

The next step is to determine subcriteria based on the criteria established earlier. The subcriteria will be assigned scores on a scale of 1 to 5, as follows:

1 = Very Poor
2 = Poor
3 = Fair
4 = Good
5 = Excellent

TABLE III. Subcriterion Scores

No	Criterion Name	Attribute	Weight
1	Attendance	Excellent	5
2	Attendance	Good	4
3	Attendance	Fair	3
4	Attendance	Poor	2
5	Attendance	Very Poor	1
…	….	….	….
…	….	….	….
25	Performance	Very Poor	1

B. Determining the Compatibility Rating

The compatibility rating determination stage involves assigning scores to the predetermined criteria—attendance, behavior, loyalty, neatness, and performance—for each identified candidate, namely the employees at PT Family Raya CRF Padang. In the suitability rating determination stage, each alternative is assigned a score that reflects the importance of each criterion. This suitability rating determination stage is performed manually by the user based on an assessment of all employees considered as alternatives. The suitability rating results can be seen in the following table.

TABLE IV. Compatibility Rating Data

No	Alternatives	Criteria				
		C1	C2	C3	C4	C5
1	A1	1	3	3	4	5
2	A2	2	3	5	3	3
3	A3	3	4	2	3	4
4	A4	4	4	4	5	3
5	A5	5	3	4	3	2
…	….	….	…	…	…	…
…	…	…	…	…	…	…
15	A15	3	4	4	3	1

C. Performing Normalization or Weight Adjustment

The normalization or weight adjustment process involves modifying the weights of each criterion. The normalization or weight adjustment process is based on Equation (1)

$$w_j = \frac{wj}{\sum wj} \quad (1)$$

The process of normalization or weight adjustment can be seen in the following solution:

$$w_1 = \frac{0.2}{(0.2+0.2+0.2+0.2+0.2)} = \frac{0.2}{1.00} = 0.2$$

$$w_2 = \frac{0.2}{(0.2+0.2+0.2+0.2+0.2)} = \frac{0.2}{1.00} = 0.2$$

$$w_3 = \frac{0.2}{(0.2+0.2+0.2+0.2+0.2)} = \frac{0.2}{1.00} = 0.2$$

$$w_4 = \frac{0.2}{(0.2+0.2+0.2+0.2+0.2)} = \frac{0.2}{1.00} = 0.2$$

D. Determining the Value of Vector S

The results of the normalization process or weight adjustment will be used to determine the value of vector s. The results of this process will be used to determine the value of vector v. The process of determining the value of vector s is based on equation (2).

$$Si = \prod_j^n Xijwj \prod_j^n Xijwj \quad (2)$$

The process of determining the value of vector s can be seen in the following solution:

$$S_1 = (1^{(0.2)} * 3^{(0.2)} * 3^{(0.2)} * 4^{(0.2)} * 5^{(0.2)}) = (1.00 * 1.24 * 1.24 * 1.31 * 1.37) = 2.83$$

$$S_2 = (2^{(0.2)} * 3^{(0.2)} * 5^{(0.2)} * 3^{(0.2)} * 3^{(0.2)}) = (1.14 * 1.24 * 1.37 * 1.24 * 1.24) = 3.06$$

$$S_3 = (3^{(0.2)} * 4^{(0.2)} * 2^{(0.2)} * 3^{(0.2)} * 4^{(0.2)}) = (1.24 * 1.31 * 1.37 * 1.37 * 1.24) = 3.10$$

$$S_4 = (4^{(0.2)} * 4^{(0.2)} * 4^{(0.2)} * 5^{(0.2)} * 3^{(0.2)}) = (1.31 * 1.31 * 1.37 * 1.37 * 1.24) = 3.95$$

$$S_5 = (5^{(0.2)} * 3^{(0.2)} * 4^{(0.2)} * 3^{(0.2)} * 2^{(0.2)}) = (1.37 * 1.24 * 1.31 * 1.24 * 1.14) = 3.35$$

E. Determining the Value of Vector V

$$w_5 = \frac{0.2}{(0.2+0.2+0.2+0.2+0.2)} = \frac{0.2}{1.00} = 0.2$$

The result of the vector s process will be used to calculate the value of vector v; the resulting value of vector v will be used for the ranking process, in which the values of vector v will be sorted from highest to lowest. The process of determining the value of vector v is based on equation (3).

$$V_1 = \frac{\prod_j^n=1 xijwj}{\prod_j^n=1 (xj)wj} \quad (3)$$

The process of determining the value of vector v can be seen in the following solution:

$$V_1 = \frac{2.83}{(2.83+3.06+3.10+3.73+3.25+2.99+3.73+3.10+2.61+2.49+3.59+3.25+4.37+2.61+2.70)} = 0.06$$

$$V_2 = \frac{3.06}{(2.83+3.06+3.10+3.73+3.25+2.99+3.73+3.10+2.61+2.49+3.59+3.25+4.37+2.61+2.70)} = 0.06$$

$$V_3 = \frac{3.10}{(2.83+3.06+3.10+3.73+3.25+2.99+3.73+3.10+2.61+2.49+3.59+3.25+4.37+2.61+2.70)} = 0.07$$

$$V_4 = \frac{3.95}{(2.83+3.06+3.10+3.73+3.25+2.99+3.73+3.10+2.61+2.49+3.59+3.25+4.37+2.61+2.70)} = 0.08$$

$$V_5 = \frac{3.25}{(2.83+3.06+3.10+3.73+3.25+2.99+3.73+3.10+2.61+2.49+3.59+3.25+4.37+2.61+2.70)} = 0.07$$

F. Ranking

The results of vector v will be ranked, with the values from the vector v process sorted from highest to lowest. The highest value will be labeled as "recommended" and can be used as a reference for the research subjects in selecting the best-performing employees. The following table shows the ranking results based on the vector v values for each alternative.

TABLE V. Ranking Result

No	Alternative	Value of Vector v	Decision
1	A13	0.9	Recommended
2	A4	0.8	Not Recommended
3	A7	0.8	Not Recommended
4	A11	0.8	Not Recommended
5	A5	0.7	Not Recommended
…	….	….	….
…	….	….	….
15	A10	0.5	Not Recommended

Based on the preference results table above, it can be concluded that the highest-ranked alternative is Alternative 13 with a preference score of 0.9, and Alternative 13 is recommended as the top-performing employee at PT Family Raya CRF.

3.2 Design Implementation

A. Form Login

The login form allows administrators and users to access the system by entering their username and password in the provided form. The login form is shown in Fig. 3.

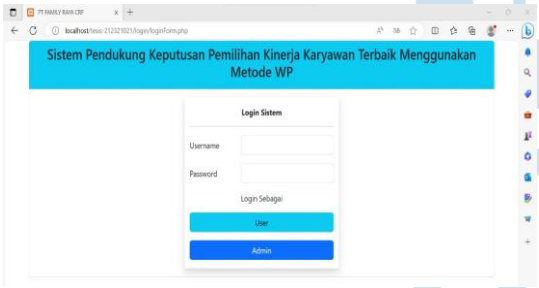


Fig.3. form login

B. Form menu admin

The admin menu form is a form accessed by the administrator after logging in. On this form, the administrator can manage user data. The admin menu form is shown in Fig. 4.



Fig. 4. Admin menu page

C. Form menu user

The user menu form is the form that users access after successfully logging in. The user menu form contains several pages that users can access to use the system. The user menu form is shown in Figure 5.



Fig. 5. User menu page

3.3 System testing

A. Alternative data processing

On this page, users can perform alternative data operations such as adding, modifying, and deleting data. Employee data from PT Family Raya CRF will be used as the alternative data. The alternative data form is shown in Figure 6.

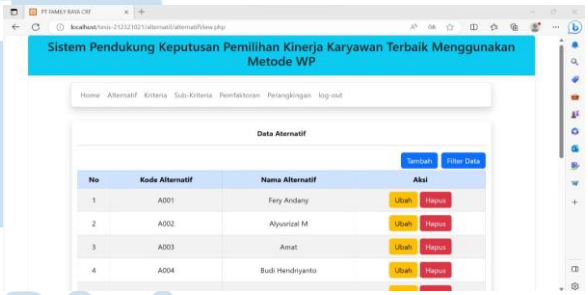


Fig. 6. Alternative Data

As shown in Figure 6 above, users can perform alternative data processing, such as adding alternative code data and alternative name data.

B. Criteria data processing

On this page, users can manage criteria data, such as adding, editing, and deleting data. The criteria data form is shown in Figure 7.

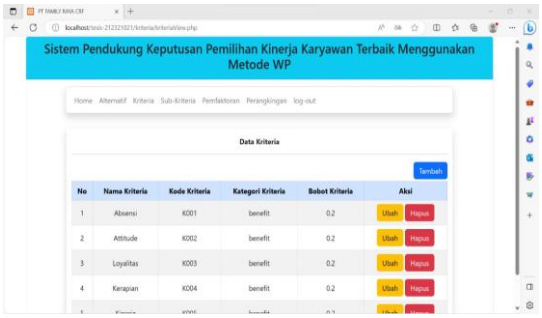
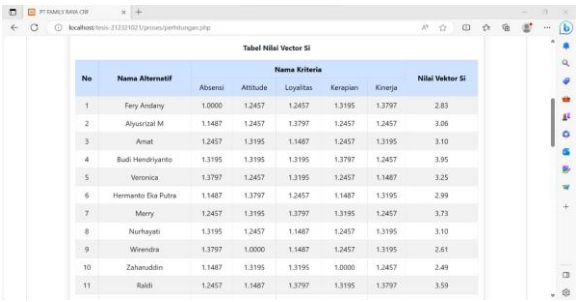


Fig. 7 *Criteria data processing*

As shown in Figure 7 above, users can process criteria data such as criteria codes, alternative names, criteria categories, and criteria weights.

C. Vector S Value

The result of the vector S calculation will be displayed for the user to view. The result of the vector S calculation can be seen in Figure 8 below



Tabel Nilai Vector Si

No	Nama Alternatif	Nama Kriteria					Nilai Vektor Si
		Absensi	Attitude	Loyalitas	Kerapian	Kinerja	
1	Fery Andany	1.0000	1.2457	1.2457	1.3195	1.3797	2.83
2	Alyusrizal M	1.1487	1.2457	1.3797	1.2457	1.2457	3.06
3	Amat	1.2457	1.3195	1.1487	1.2457	1.3195	3.10
4	Budi Hendriyanto	1.3195	1.3195	1.3195	1.3797	1.2457	3.95
5	Veronica	1.3797	1.2457	1.3195	1.2457	1.1487	3.25
6	Hermanto Eka Putra	1.1487	1.3797	1.2457	1.1487	1.3195	2.99
7	Merry	1.2457	1.3195	1.3797	1.3195	1.2457	3.73
8	Nurhayati	1.3195	1.2457	1.1487	1.2457	1.3195	3.10
9	Wirendra	1.3797	1.0000	1.1487	1.2457	1.3195	2.61
10	Zaharuddin	1.1487	1.3195	1.3195	1.0000	1.2457	2.49
11	Raldi	1.2457	1.1487	1.3797	1.3195	1.3797	3.59

.

Fig. 8. S Factor Value.

Based on Fig. 8, the value of vector s is obtained using equation (2), where the weight adjustment values in the figure are processed using the factorization values.

D. Vector V Value

The result of the vector V calculation will be displayed for the user to view. The result of the vector V calculation can be seen in Figure 9 below



Tabel Nilai Vector Vi

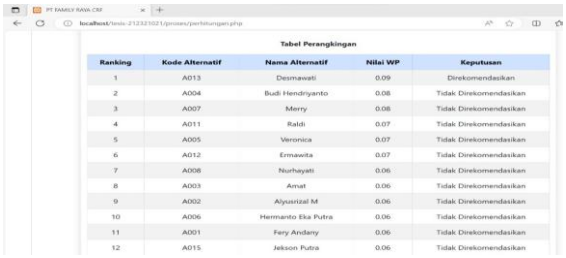
No	Nama Alternatif	Nilai Vektor Si	Total Nilai Vektor Si	Nilai Vektor Vi
1	Fery Andany	2.83	48.63	0.06
2	Alyusrizal M	3.06	48.63	0.06
3	Amat	3.10	48.63	0.06
4	Budi Hendriyanto	3.95	48.63	0.08
5	Veronica	3.25	48.63	0.07
6	Hermanto Eka Putra	2.99	48.63	0.06
7	Merry	3.73	48.63	0.08
8	Nurhayati	3.10	48.63	0.06
9	Wirendra	2.61	48.63	0.05
10	Zaharuddin	2.49	48.63	0.05
11	Raldi	3.59	48.63	0.07
12	Ermawita	3.25	48.63	0.07

Fig. 9. Value of Vector V

From Fig. 9, the value of vector v is obtained through data processing of the results from Fig. 8 using a formula based on Equation (3).

E. Ranking Results

The results of vector v will be ranked based on the highest value; the highest value will be recommended as the top employee. The ranking results are shown in Fig. 10 below.



Tabel Perangkingan

Ranking	Kode Alternatif	Nama Alternatif	Nilai WP	Keputusan
1	A013	Desmawati	0.09	Direkomendasikan
2	A004	Budi Hendriyanto	0.08	Tidak Direkomendasikan
3	A007	Merry	0.08	Tidak Direkomendasikan
4	A011	Raldi	0.07	Tidak Direkomendasikan
5	A005	Veronica	0.07	Tidak Direkomendasikan
6	A012	Ermawita	0.07	Tidak Direkomendasikan
7	A008	Nurhayati	0.06	Tidak Direkomendasikan
8	A003	Amat	0.06	Tidak Direkomendasikan
9	A002	Alyusrizal M	0.06	Tidak Direkomendasikan
10	A006	Hermanto Eka Putra	0.06	Tidak Direkomendasikan
11	A001	Fery Andany	0.06	Tidak Direkomendasikan
12	A015	Jekson Putra	0.06	Tidak Direkomendasikan

Fig. 10. Ranking Results

Based on Figure 10 above, the ranking results obtained using the system show that Alternative 13 is the best alternative. Alternative 13 is marked as "recommended.".

IV. CONCLUSION

The Weight Product (WP) method can be applied in the decision-making process for selecting the best-performing employees. In this study, the data was processed using 15 alternatives and 5 criteria: attendance, behavior, loyalty, neatness, and performance. Based on the data processing using the Weight Product (WP) method, alternative A13—an employee named Desmawati—was identified as the best alternative with a Weight Product (WP) score of 0.9.

A Decision Support System for selecting the best-performing employee using the Weight Product method can be developed into a web-based information system using the PHP programming language and a MySQL database. This system is designed to assist in accurately and precisely selecting the best-performing employee at PT. Family Raya CRF.

REFERENCES

[1] Andoyo, A., Angraeni, E.Y. and Khumaidi, A. (2021) SISTEM PENDUKUNG KEPUTUSAN Konsep, Implementasi & Pengembangan. Penerbit Adab. Available: https://books.google.co.id/books?id=YTgmEAAAQBAJ.

[2] Sofiati, E., Gusti, D., Candra, A., & Meiditra, I. (2024). Data Mining Motorbike Sales Classification Using a Combination of K-Means and Naïve Bayes Algorithms. Ultimatics : Jurnal Teknik Informatika, 16(2).

[3] Sabandar, V. P., & Ahmad, R. (2023). Sistem Pendukung Keputusan Penentuan Produk Terbaik Menggunakan Weighted Product Method. Jurnal Ilmiah Computer Science, 1(2), 58–68. https://doi.org/10.58602/jics.v1i2.7

[4] Rizky, A., & Cipta, H. (2024). Application Of The Weighted Product (WP) Method In The Selection Of Prospective New Employees Using Assessment Indicators At PT. Delta Sukses Sejahtera. Journal of Computer Networks, Architecture and High Performance Computing, 6(4), 1809–1817. https://doi.org/10.47709/cnahpc.v6i4.4785

[5] Sinaga, A., & Maulana, D. (2022). Implementation of Weighted Product Method for Evaluating Performance of Technicians. International Journal of Modern Education and Computer Science, 14(4), 30–42. https://doi.org/10.5815/ijmecs.2022.04.03

[6] Rezagustini, N., Adisti, A., Royadi, D., & Sunggono, N. T. (2025). Implementation of Weighted Product Method for Teacher Selection at an Islamic Boarding School. Bit-Tech, 8(1), 809–819. https://doi.org/10.32877/bt.v8i1.2736

[7] Stefanus M Patanduk, & A.Sidiq Purnomo. (2025). Sistem Pendukung Keputusan Menentukan Kelayakan Kenaikan Gaji Karyawan Menggunakan Metode Weighted Product (WP) (Studi Kasus : Gading Mas Yogyakarta). JEKIN - Jurnal Teknik Informatika, 5(2), 707–721. https://doi.org/10.58794/jekin.v5i2.1446

[8] Sunardi, D., Zein, A. ., & Rozali, C. . (2025). Implementasi Metode Weighted Product (WP) Sistem Penunjang Keputusan Da-lam Menentukan Prestasi Dosen Tebaik Fakultas Ilmu Komputer Universitas Pamulang. Jurnal Informatika Utama, 3(1), 32–41. https://doi.org/10.55903/jitu.v3i1.270

[9] Hidayat, R. N., Santoso, B., & Sumirat, L. P. (2024). Sistem Pendukung Keputusan Pemilihan Siswa Berprestasi Menggunakan Metode Simple Additive Weighting dan Weighted Product: Decision Support System for Selecting Outstanding Students Using Simple Additive Weighting and Weighted Product Methods. MALCOM: Indonesian Journal of Machine Learning and Computer Science, 5(1), 379-390. https://doi.org/10.57152/malcom.v5i1.1787

[10] Halawa, D., Nasution, H. N. ., & Ermawita, E. (2025). SISTEM PENDUKUNG KEPUTUSAN PEMILIHAN KARYAWAN TERBAIK MENGGUNAKAN METODE WEIGHTED PRODUCT (WP) DI PT. COLUMBUS PADANGSIDIMPUAN. JURNAL VINERTEK (Vokasional Informatika Edukasi Riset Dan Teknologi), 5(1). https://doi.org/10.37081/vn.v5i1.1852

[11] D. Rezekika, "Penerapan Algoritma Naïve Bayes Untuk Memprediksi Penjualan Spare Part Sepeda Motor," *J. Pelita Inform.*, vol. 8, no. 3, pp. 326–329, 2020.

[12] Hidayat, S., Candra Kirana, H., Halim Saputra, R., Nugroho, K., Ardhianto, E., Informasi, T., Teknologi dan Industri, F., Stikubank, U., & Artikel, R. (2025). INFO ARTIKEL ABSTRAK. 4(3), 250–257. https://doi.org/10.55123

[13] Odelia, & Dicky Pratama. (2025). PENERAPAN METODE WEIGHT PRODUCT PADA SISTEM PENDUKUNG KEPUTUSAN KARYAWAN TERBAIK. Jurnal Rekayasa Sistem Informasi Dan Teknologi, 2(3), 910–920. https://doi.org/10.70248/jrsit.v2i3.1798

[14] Arif, S. M., & Irsan, M. (2026). Sistem Pendukung Keputusan Rekrutmen Karyawan Menggunakan Metode WP PT. Lucky Indah Keramik. Jejak Digital: Jurnal Ilmiah Multidisiplin, 2(2), 3040-3048. https://doi.org/10.63822/43rt7f24

[15] Alriza, A., & Supatman, S. (2025). PENERAPAN METODE WEIGHTED PRODUCT DALAM SELEKSI PENERIMAAN KARYAWAN BARU. Jurnal Informatika Dan Teknik Elektro Terapan, 13(1). https://doi.org/10.23960/jitet.v13i1.5889

AUTHOR GUIDELINES

1. Manuscript criteria

- The article has never been published or in the submission process on other publications.
- Submitted articles could be original research articles or technical notes.
- The similarity score from plagiarism checker software such as Turnitin is 20% maximum.
- For December 2021 publication onwards, Ultimatics : Jurnal Teknik Informatika will be receiving and publishing manuscripts written in English only.

2. Manuscript format

- Article been type in Microsoft Word version 2007 or later.
- Article been typed with 1 line spacing on an A4 paper size (21 cm x 29,7 cm), top-left margin are 3 cm and bottom-right margin are 2 cm, and Times New Roman's font type.
- Article should be prepared according to the following author guidelines in this template. Article contain of minimum 3500 words.
- References contain of minimum 15 references (primary references) from reputable journals/conferences

3. Organization of submitted article

The organization of the submitted article consists of Title, Abstract, Index Terms, Introduction, Method, Result and Discussion, Conclusion, Appendix (if any), Acknowledgment (if any), and References.

- Title
  The maximum words count on the title is 12 words (including the subtitle if available)
- Abstract
  Abstract consists of 150-250 words. The abstract should contain logical argumentation of the research taken, problem-solving methodology, research results, and a brief conclusion.
- Index terms
  A list in alphabetical order in between 4 to 6 words or short phrases separated by a semicolon (;), excluding words used in the title and chosen carefully to reflect the precise content of the paper.
- Introduction
  Introduction commonly contains the background, purpose of the research, problem identification, research methodology, and state of the art conducted by the authors which describe implicitly.
- Method
  Include sufficient details for the work to be repeated. Where specific equipment and materials are named, the manufacturer's details (name, city and country) should be given so that readers can trace specifications by contacting the manufacturer. Where commercially available software has been used, details of the supplier should be given in brackets or the reference given in full in the reference list.
- Results and Discussion
  State the results of experimental or modeling work, drawing attention to important details in tables and figures, and discuss them intensively by comparing and/or citing other references.
- Conclusion
  Explicitly describes the research's results been taken. Future works or suggestion could be explained after it
- Appendix and acknowledgment, if available, could be placed after Conclusion.
- All citations in the article should be written on References consecutively based on its' appearance order in the article using Mendeley (recommendation). The typing format will be in the same format as the IEEE journals and transaction format.

4. Reviewing of Manuscripts

Every submitted paper is independently and blindly reviewed by at least two peer-reviewers. The decision for publication, amendment, or rejection is based upon their reports/recommendations. If two or more reviewers consider a manuscript unsuitable for publication in this journal, a statement explaining the basis for the decision will be sent to the authors within six months of the submission date.

5. Revision of Manuscripts

Manuscripts sent back to the authors for revision should be returned to the editor without delay (maximum of two weeks). Revised manuscripts can be sent to the editorial office through the same online system. Revised manuscripts returned later than one month will be considered as new submissions.

6. Editing References

- **Periodicals**
  J.K. Author, "Name of paper," Abbrev. Title of Periodical, vol. x, no. x, pp. xxx-xxx, Sept. 2013.

- **Book**
  J.K. Author, "Title of chapter in the book," in Title of His Published Book, xth ed. City of Publisher, Country or Nation: Abbrev. Of Publisher, year, ch. x, sec. x, pp xxx-xxx.

- **Report**
  J.K. Author, "Title of report," Abbrev. Name of Co., City of Co., Abbrev. State, Rep. xxx, year.

- **Handbook**
  Name of Manual/ Handbook, x ed., Abbrev. Name of Co., City of Co., Abbrev. State, year, pp. xxx-xxx.

- **Published Conference Proceedings**
  J.K. Author, "Title of paper," in Unabbreviated Name of Conf., City of Conf., Abbrev. State (if given), year, pp. xxx-xxx.

- **Papers Presented at Conferences**
  J.K. Author, "Title of paper," presented at the Unabbrev. Name of Conf., City of Conf., Abbrev. State, year.

- **Patents**
  J.K. Author, "Title of patent," US. Patent xxxxxxx, Abbrev. 01 January 2014.

- **Theses and Dissertations**
  J.K. Author, "Title of thesis," M.Sc. thesis, Abbrev. Dept., Abbrev. Univ., City of Univ., Abbrev. State, year. J.K. Author, "Title of dissertation," Ph.D. dissertation, Abbrev. Dept., Abbrev. Univ., City of Univ., Abbrev. State, year.

- **Unpublished**
  J.K. Author, "Title of paper," unpublished. J.K. Author, "Title of paper," Abbrev. Title of Journal, in press.

- **On-line Sources**
  J.K. Author. (year, month day). Title (edition) [Type of medium]. Available: http:// www.(URL) J.K. Author. (year, month). Title. Journal [Type of medium]. volume(issue), pp. if given. Available: http://www.(URL) Note: type of medium could be online media, CD-ROM, USB, etc.

7. Editorial Adress

Universitas Multimedia Nusantara
Jl. Scientia Boulevard, Gading Serpong
Tangerang, Banten, 15811
Email: **ultimatics@umn.ac.id**

Paper Title
Subtitle (if needed)

Author 1 Name[1], Author 2 Name[2], Author 3 Name[2]
[1] Line 1 (of affiliation): dept. name of organization, organization name, City, Country
Line 2: e-mail address if desired
[2] Line 1 (of affiliation): dept. name of organization, organization name, City, Country
Line 2: e-mail address if desired

Accepted on mmmmm dd, yyyy
Approved on mmmmm dd, yyyy

***Abstract*—This electronic document is a "live" template which you can use on preparing your ULTIMATICS paper. Use this document as a template if you are using Microsoft Word 2007 or later. Otherwise, use this document as an instruction set. Do not use symbol, special characters, or Math in Paper Title and Abstract. Do not cite references in the abstract.**

***Index Terms*—enter key words or phrases in alphabetical order, separated by semicolon (;)**

I. INTRODUCTION

This template, modified in MS Word 2007 and saved as a Word 97-2003 document, provides authors with most of the formatting specifications needed for preparing electronic versions of their papers. Margins, column widths, line spacing, and type styles are built-in here. The authors must make sure that their paper has fulfilled all the formatting stated here.

Introduction commonly contains the background, purpose of the research, problem identification, and research methodology conducted by the authors which been describe implicitly. Except for Introduction and Conclusion, other chapter's title must be explicitly represent the content of the chapter.

II. EASE OF USE

A. Selecting a Template

First, confirm that you have the correct template for your paper size. This template is for ULTIMATICS. It has been tailored for output on the A4 paper size.

B. Maintaining the Integrity of the Specifications

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them.

III. PREPARE YOUR PAPER BEFORE STYLING

Before you begin to format your paper, first write and save the content as a separate text file. Keep your text and graphic files separate until after the text has been formatted and styled. Do not add any kind of pagination anywhere in the paper. Please take note of the following items when proofreading spelling and grammar.

A. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Abbreviations that incorporate periods should not have spaces: write "C.N.R.S.," not "C. N. R. S." Do not use abbreviations in the title or heads unless they are unavoidable.

B. Units

- Use either SI (MKS) or CGS as primary units (SI units are encouraged).
- Do not mix complete spellings and abbreviations of units: "Wb/m2" or "webers per square meter," not "webers/m2." Spell units when they appear in text: "...a few henries," not "...a few H."
- Use a zero before decimal points: "0.25," not ".25."

C. Equations

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

Number the equations consecutively. Equation numbers, within parentheses, are to position flush right, as in (1), using a right tab stop.

$$\int_0^{r_2} F(r,\phi)\, dr\, d\phi = [\sigma r_2 / (2\mu_0)] \qquad (1)$$

Note that the equation is centered using a center tab stop. Be sure that the symbols in your equation have been defined before or immediately following the equation. Use "(1)," not "Eq. (1)" or "equation (1)," except at the beginning of a sentence: "Equation (1) is ..."

D. Some Common Mistakes

- The word "data" is plural, not singular.

B. *Identify the Headings*

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include ACKNOWLEDGMENTS and REFERENCES, and for these, the correct style to use is “Heading 5.”

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles, named “Heading 1,” “Heading 2,” “Heading 3,” and “Heading 4”, are prescribed.

C. *Figures and Tables*

Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation “Fig. 1,” even at the beginning of a sentence.

TABLE I. TABLE STYLES

Table Head	**Table Column Head**		
	Table column subhead	***Subhead***	***Subhead***
copy	More table copy		

Fig. 1. Example of a figure caption

V. CONCLUSION

A conclusion section is not required. Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions.

APPENDIX

Appendixes, if needed, appear before the acknowledgment.

ACKNOWLEDGMENT

The preferred spelling of the word “acknowledgment” in American English is without an “e” after the “g.” Use the singular heading even if you have many acknowledgments. Avoid expressions such as “One of us (S.B.A.) would like to thank ... .” Instead, write “F. A. Author thanks ... .” You could

also state the sponsor and financial support acknowledgments here.

REFERENCES

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

[1] G. Eason, B. Noble, and I.N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529-551, April 1955. (*references*)

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[3] I.S. Jacobs and C.P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740-741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.



UMN

UNIVERSITAS
MULTIMEDIA
NUSANTARA

ISSN 2085-4552





Universitas Multimedia Nusantara
Scientia Garden Jl. Boulevard Gading Serpong, Tangerang
Telp. (021) 5422 0808 | Fax. (021) 5422 0800