ISSN 2085-4552

# ULTIMATICS

## Jurnal Teknik Informatika

UMN

UNIVERSITAS
MULTIMEDIA
NUSANTARA

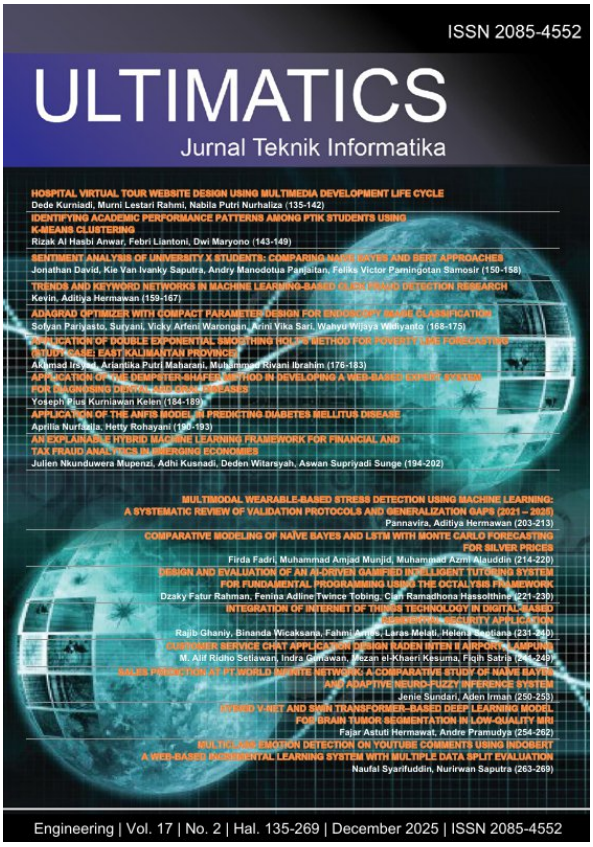
ISSN 2085-4552

# EDITORIAL BOARD



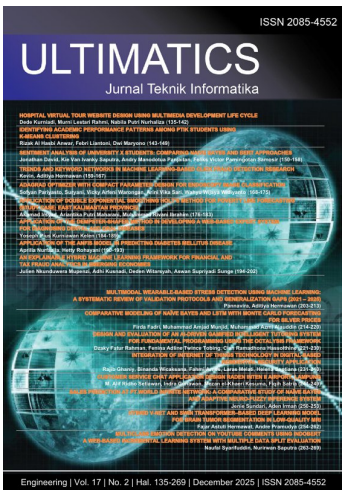
# EDITORIAL ADDRESS

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 December 2025 edition, Ultimatics enters the 2nd Edition of Volume 17. 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: Hospital Virtual Tour Website Design Using Multimedia Development Life Cycle, Identifying Academic Performance Patterns Among PTIK Students Using K-Means Clustering, Sentiment Analysis of University X Students: Comparing Naive Bayes and BERT Approaches, Trends and Keyword Networks in Machine Learning-Based Click Fraud Detection Research, Adagrad Optimizer with Compact Parameter Design for Endoscopy Image Classification, Application Of Double Exponential Smoothing Holt's Method For Poverty Line Forecasting (Study Case: East Kalimantan Province), Application of the Dempster-Shafer Method in Developing a Web-Based Expert System for Diagnosing Dental and Oral Diseases, Application of the ANFIS Model in Predicting Diabetes Mellitus Disease, An Explainable Hybrid Machine Learning Framework for Financial and Tax Fraud Analytics in Emerging Economies, Multimodal Wearable-Based Stress Detection Using Machine Learning: A Systematic Review of Validation Protocols and Generalization Gaps (2021 – 2025), Comparative Modeling of Naïve Bayes and LSTM with Monte Carlo Forecasting for Silver Prices, Design and Evaluation of an AI-Driven Gamified Intelligent Tutoring System for Fundamental Programming Using the Octalysis Framework, Integration of Internet of Things Technology in Digital-Based Residential Security Application, Customer Service Chat Application Design Raden Inten II Airport Lampung, Sales Prediction at PT.World Infinite Network: A Comparative Study Of Naïve Bayes and Adaptive Neuro-Fuzzy Inference System, Hybrid V-Net and Swin Transformer–Based Deep Learning Model for Brain Tumor Segmentation in Low-Quality MRI, Multiclass Emotion Detection on YouTube Comments Using IndoBERT A Web-Based Incremental Learning System with Multiple Data Split Evaluation.

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.

December 2025,

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

ISSN 2085-4552

# DAFTAR ISI

ISSN 2085-4552

# Hospital Virtual Tour Website Design Using Multimedia Development Life Cycle

Dede Kurniadi[1], Murni Lestari Rahmi[2], Nabila Putri Nurhaliza[3]
[1,2,3] Department of Computer Science, Institut Teknologi Garut, Garut, Indonesia
[1] dede.kurniadi@itg.ac.id, [2] 2106035@itg.ac.id, [3] 2106074@itg.ac.id

Accepted 22 December 2025
Approved 06 January 2026

***Abstract*— In 2023, Indonesia recorded 3,155 hospitals spread across the country, comprising 2,636 general and 519 specialized hospitals. Although the number is significant, not all community members have easy, direct access to hospitals. To overcome this challenge, virtual tour technology has emerged as a relevant solution to facilitate access to information and increase the transparency of hospital services. This project aims to develop a virtual tour website for Medina Hospital in Garut Regency. The project uses a systematic development method known as the Multimedia Development Life Cycle (MDLC), which includes stages from concept to distribution. The resulting website allows users to explore various hospital areas, such as the Main Building, Emergency Room, Tulip Building, and Chemotherapy Poly, through a virtual 360-degree panoramic view. Additionally, building and floor selection features are designed to make it easy for users to navigate. This website is also equipped with a chatbot feature that helps users find the location of a specific room and video tutorial guides that provide instructions for using the website. The results of the black box test show that the website functions well without any significant technical problems, so it is ready for public use. This website is expected to increase accessibility and convenience for users in obtaining information about the facilities and rooms available at Medina Hospital.**

***Index Terms*— 360-degree Panorama; Hospital Virtual Tour; Interactive Website; Multimedia Development Life Cycle (MDLC).**

## I. INTRODUCTION

Based on the latest data, the number of hospitals in Indonesia has increased by 9.7% from 2,877 hospitals in 2019 to 3,155 in 2023. This number consists of 2,636 general hospitals and 519 special hospitals [1]. In Garut Regency, several hospitals serve public health needs, one of which is Medina Hospital. It was established in 2021 and is located in Wanaraja. With the increasing number of hospitals, there is a challenge to provide information that is easily accessible to the public regarding the facilities and services available in each hospital. In this case, information technology plays an important role in introducing and promoting the profile of hospital institutions as a transparent and informative public service [2]. Based on interviews with the management of Medina Hospital and observations of the hospital's official website, it was found that information regarding facilities and supporting services is still presented in text form, and there is no information about room locations. This makes it difficult for visitors to understand the building layout and locate specific facilities within the hospital. Virtual tour technology, as one of the latest innovations, offers a solution that can make it easier for patients and visitors to explore the facilities [3] of the hospital virtually, especially considering the complexity of the often-confusing layout of hospital buildings. Conventional websites containing text and static images are unable to provide a complete spatial representation. Users can only view certain angles rather than the full context of the space. However, a 360° virtual tour allows interactive visualization, enabling visitors to see facility areas, directions, and building layouts more clearly, as if they were physically present on-site.

A virtual tour is a simulation that displays a specific location through a series of photos that are put together into a panorama with a 360° perspective [4]. In the virtual tour, several images are combined through a stitching process to create a comprehensive panoramic photo effect. By integrating multiple photos, a wider and more detailed view is produced in the form of a 360° panorama [5]. Virtual tours can be used in various sectors such as education [6], tourism [7], real estate [8], hospitality [9] and the health sector as well as hospitals. The use of virtual tours in hospitals not only makes it easier for patients and visitors to explore the facilities, but can also provide better information regarding the services available, such as inpatient rooms, emergency rooms, operating rooms, polyclinics, and others.

This research is prepared by referring to several previous studies on the topic discussed, namely virtual tours. One of the studies focuses on the development of a Virtual tour application for Family Recreational Objects on 3D Stable which is designed to provide an interactive experience to visitors through the Multimedia Development Life Cycle (MDLC) method, allowing users to explore the 3D Stable area in 360 degrees [10]. With the same method, another study developed a virtual tour application for SMP Negeri 3

Kota Pagar Alam, which aims to help promote the school and facilitate access to information for students, teachers, and the community virtually [6]. In addition, there is also the development of a virtual tour application for SMK Negeri 1 Wajo, which helps prospective new students obtain information about majors and school facilities with interactive features that can be accessed through mobile and desktop devices [3]. Furthermore, the research on the virtual tour of Tourism in Lahendong Village using the prototyping method aims to introduce regional tourism by facilitating virtual access to information [7]. Finally, innovations in virtual tours that modify the Borg and Gall method focus on the development of educational tourism, which provides an interesting exploration experience for visitors and increases public awareness of the importance of the University of Malang Learning Museum [11].

Based on the background and previous research, this research was conducted to develop a virtual tour application at Medina Hospital using the MDLC method. The MDLC method was chosen because it offers a structured approach specifically for developing interactive multimedia content that integrates visual elements, text, and digital objects [12]. This method is also suitable because it supports the development of multimedia applications that combine images, animations, and interactive components in a systematic and efficient workflow [13]. This application is designed to make it easier for patients and visitors to explore various hospital facilities virtually. By utilizing virtual tour technology, it is hoped that the user experience can be improved through interactive features accessed through mobile and desktop devices

## II. METHOD

In order to maintain the timeliness and quality of the research project, the researcher implements the MDLC (Multimedia Development Life Cycle) approach in creating this project because of the appropriate development cycle for multimedia applications from start to finish which includes six stages [14] [15]. The six stages of MDLC include Concept, Design, Material Collection, Assembly, Testing, and Distribution presented in Figure 1.

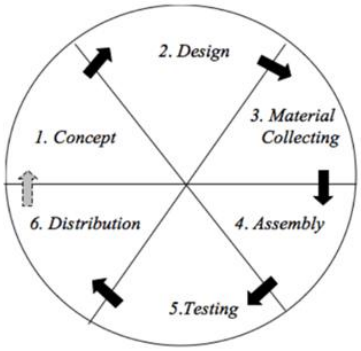



Fig. 1. Stages of the Multimedia Development Cycle

The first stage carried out is the Concept which is the initial formulation stage of the strategy to implement the project. Among them are determining the purpose of creating the project, identifying users, and determining strategies to achieve the project goals. In this project, the process carried out includes setting the goal of making the project, which is to facilitate access to information about Medina Hospital and provide an interesting visual experience to visitors. In addition, this stage also includes identifying the target users, such as patients, patients' families, or medical professionals, as well as formulating strategies to achieve the project objectives.

Furthermore, in the second stage, the design of multimedia project specifications was carried out. In this project, the process carried out includes designing the virtual tour navigation flow, program structure, virtual tour display design, and website user interface (UI). This design aims to allow users to explore the hospital virtually in a structured and easy-to-understand way.

The third stage is Material Collection, where at this stage the necessary assets are collected for the project. In the Medina Hospital virtual tour project, the main asset required is a 360-degree panoramic image of every corner of the hospital location taken in high quality to provide an accurate and compelling representation of the hospital. In addition, text assets are required to label the name of each location.

Then, all assets that have been collected at the Material Collection stage are then assembled in this fourth stage until it becomes a multimedia project. In this project, panorama assets are assembled according to the design created in the Design stage, including the addition of interactive elements and additional information with the help of the Panorama Studio 3 Pro application to create a virtual tour. This process also includes creating a website allowing users to access virtual tours. This process also includes creating a website using Visual Studio Code (VS Code) tools, an open-source source code editor that supports cloud and web development [16]. The website will be a platform for users to access virtual tours.

After the assembly stage is completed, testing is carried out using the black box method, which is a test that focuses on evaluating the functionality of the software based on predetermined criteria [17]. Blackbox testing aims to detect and identify various problems or malfunctions that may arise in the application [18], to ensure that all aspects of the virtual tour are working properly before they are distributed.

In the last stage, namely Distribution, the virtual tour website is hosted using the Koyeb cloud platform [19] to be accessed online through a URL. To facilitate access, QR codes were created and shared through various means, including social media and posters placed at several points of Medina Hospital locations.

This ensures that virtual tours can be accessed by the public easily and efficiently.

## III. Result And Discussion

### A. Concept

The Concept Stage in this study includes four activities carried out. The first activity is the identification of needs, carried out through interviews with the IT team of Medina Hospital which succeeded in revealing the needs of functionality that must be in the virtual tour website, including information about the rooms and facilities at Medina Hospital. In this activity, system requirements are generated which are presented in Table 1.

TABLE I. System Requirement

| No | System Requirements |
|---|---|
| **The system must be able to:** | |
| 1 | Displays 360° panoramic photos that clearly show the hospital area |
| 2 | Display the name of the room at each point of the room |
| 3 | Intuitive navigation features between rooms and floor |
| 4 | Drag to explore panoramic images of the virtual tour |
| 5 | Zoom in/out on panoramic images to see clearer details |
| 6 | There is a video tutorial on conducting a virtual tour |
| 7 | Displays brief profile information of Medina Hospital |
| 8 | Provides information on the location of a room |
| 9 | There is information on the location of Medina Hospital and can direct users via Google Maps |
| 10 | Dynamic website display |
| 11 | Responsive accessibility for a wide range of devices |
| **Nonfunctional Requirements:** | |
| 1 | Colors adjust the philosophy of Medina Hospital/similar to the main website of Medina Hospital |
| 2 | The website must be user-friendly |

Furthermore, the second activity at this stage is designing a concept. Based on the results of the interview, an initial concept was designed that described how this website works. The detailed description of the initial concept of the Medina Hospital virtual tour website is as follows.

- On the website, there are four pages (Home, Profile, Tutorial, and Virtual Tour).
- The home page displays a brief description of the website, and a button to start a virtual tour.
- To start the virtual tour, there are also four buttons based on the building (Main Building, Chemo Poly Building, Emergency Room Building, and Tulip Building) on the main page.
- The Profile page displays a brief history of Medina Hospital.
- The Tutorial page shows you how to use the website in the form of a video.
- The Virtual Tour page displays a 360-degree image display for virtual tours, there is also a menu button to move to a building/floor and a home button.
- There is a hotspot button on the virtual tour that allows users to switch locations.
- There is a chatbot to provide information on the location of a room.
- There is a button that points to Google Maps.

Then the third activity is the identification of the tools needed in this research. In this activity, a list of tools, both hardware and software, is produced which is presented in detail in Table 2.

TABLE II. Tools Used

| No | Tools | Description |
|---|---|---|
| 1 | Smartphone | Used for capturing 360-degree panoramic image assets in various areas of the hospital |
| 2 | Laptop | The main tools for project development and management, from design, and coding to deployment |
| 3 | DrawIO | Tools to create visual diagrams and design virtual tour interfaces and websites |
| 4 | Panorama 360 & Virtual Tours | Used to produce 360-degree panoramic images |
| 5 | Canva | Used for the creation of text assets, icons, and other graphic elements |
| 6 | PanoramaStudio 3 Pro | Used to edit and combine panoramic images into an interactive virtual tour |
| 7 | Visual Studio Code | Text editor and IDE for writing programming code in website development |
| 8 | Koyeb | Free hosting service platform used to deploy websites |

The last activity at this stage is the study of the literature of relevant journals and articles to be used as a reference in building virtual tours, the features that may exist in virtual tours, the latest technology that supports the creation of virtual tours, and the methods used to work on virtual tour projects. In this activity, the researcher obtained 5 journals that were used as a reference for making projects.

### B. Design

The Design stage in this study includes two stages of activities, the result of the design is in the form of a wireframe made using draw.io to provide an initial overview of the layout and navigation flow of the Medina Hospital virtual tour website.

The first activity is to design the virtual tour display. In this activity, the researcher determined the basic structure that would be used to navigate the user through various rooms and hospital floors, which in

this case was the hotspot icon. The hotspot icon used is listed in Table 3.

TABLE III. HOTSPOT IKONS IN USE

| No | Tools | Description |
|---|---|---|
| 1 |  | Used to direct users if they want to go to a specific area, this arrow will make it easier for users to understand the direction to take in a virtual tour. |
| 2 |  | Indicates the name of the specific location of the room in the hospital. Provide users with visual information about important points, such as a building, room, or specific facility. |
| 3 |  | Located on various doors that indicate the interaction area. When the user clicks on this icon, they can enter and exit the room. |

Then the second activity at this stage is to design a website appearance that focuses on the layout and structure of the website pages that will be used to access the virtual tour. Figure 2 presents the navigation flow diagram, which illustrates the relationships between pages and the main interaction paths within the Medina Hospital virtual tour website.

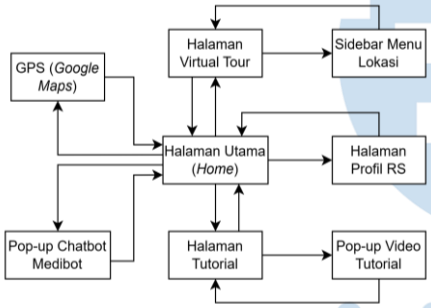



Fig. 2. Website Navigation Flow Design

Based on the navigation flow diagram in Figure 2, users begin their exploration from the Main Page (Home), which serves as the central navigation hub of the website. From this page, users can access several other pages, including the Hospital Profile Page, Tutorial Page, Virtual Tour Page, as well as interactive features such as the Medibot Chatbot Pop-up and the Tutorial Video Pop-up.

This activity also produces several UI layouts that will be used in the website, as presented in Figure 3 - Figure 9.

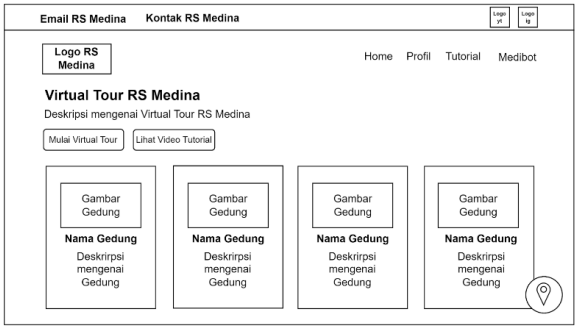



Fig. 3. Main Page Design

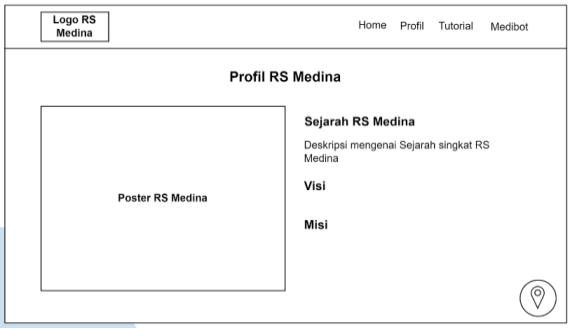



Fig. 4. Profile Page Design

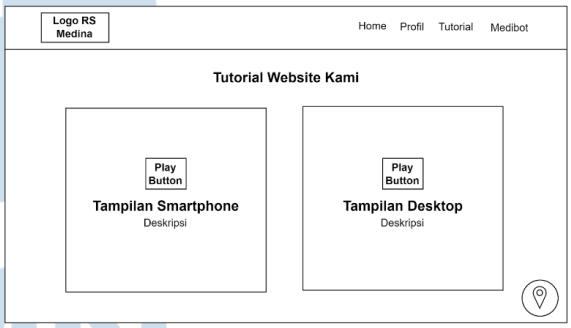



Fig. 5. Tutorial Page Design

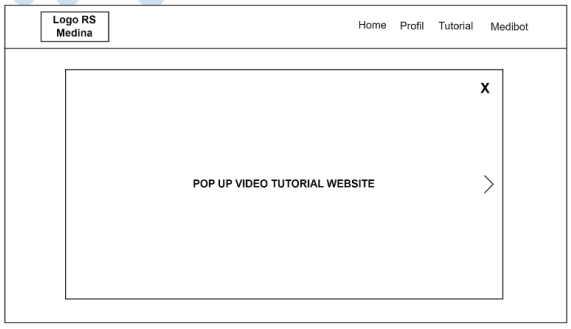



Fig. 6. Tutorial Video Pop-up Design

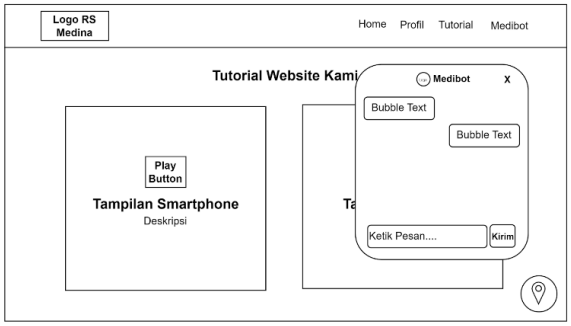



Fig. 7. Chatbot Pop-Up Design

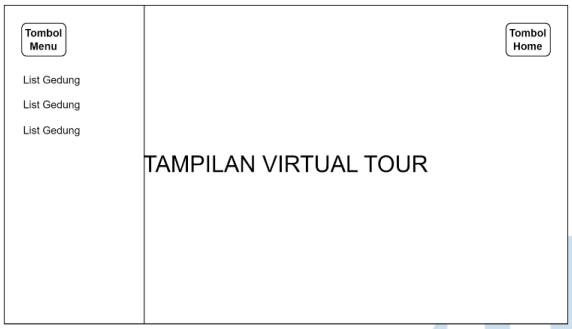



Fig. 8. 360 Degree Panorama Image Sample

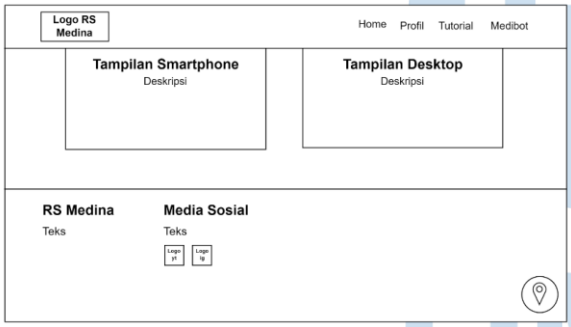



Fig. 9. Web Footer Design

These designs will then be implemented into the website at the Assembly stage.

## C. *Material Collecting*

At the Material Collecting stage, two activities were carried out. The first activity was to take panoramic pictures at 67 location points. Among them are 51 location points in the Main Building, 5 location points in the Chemotherapy Building, 5 location points in the Emergency Room Building, and 6 location points in the Tulip Building. This activity resulted in a 360-degree panoramic image with a total of 67 images. The resulting 360-degree panoramic image sample can be seen in Figure 10.

Fig. 10. 360 Degree Panorama Image Sample

Next, the second activity at this stage is to create text assets and icons using Canva tools. In this activity, a total of 131 text assets and a total of 12 icon assets were produced, both of which were in .png format. The resulting asset sample is shown in Figure 11.


Fig. 11. Text & Icon Assets Sample

The assets that have been collected, namely 67 360-degree panoramic images and 143 images of text assets and icons, will then be implemented at a later stage.

## D. *Assembly*

There are two activities in this Assembly stage. The first activity is to create a virtual tour by integrating 360-degree panoramic images from 67 locations at Medina Hospital into the PanoramaStudio 3 Pro application, including placing interactive elements such as hotspot icons and text. This activity resulted in a virtual tour, which can be seen in Figure 12.

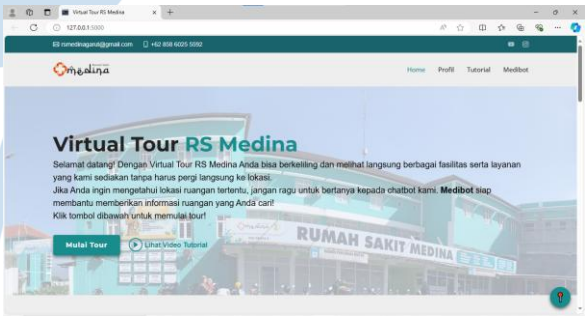



Fig. 12. Virtual Tour Page Display

The second activity is to create a virtual tour website. This website is developed according to the concept that has been designed and uses all the materials that have been collected at the celebrity stage by using Visual Studio Code (VS Code) tools. After that, the virtual tour is integrated into the website. The following are the results of the construction of the Medina Hospital virtual tour website presented in Figure 13 - Figure 20.

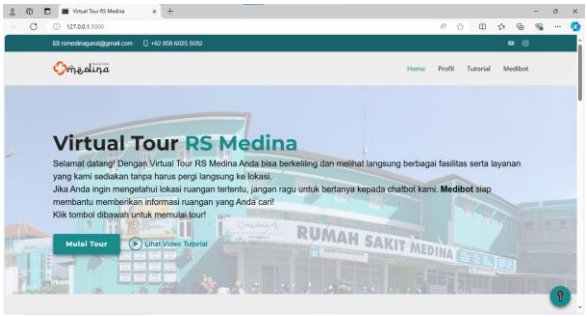



Fig. 13. Main Page Display (Home)

Figure 13 shows the main page (Home) which is the initial display of the website. On this page, users can start the virtual tour by clicking the "Start Tour" button or go to the tutorial page by clicking the "View Video Tutorial" button.

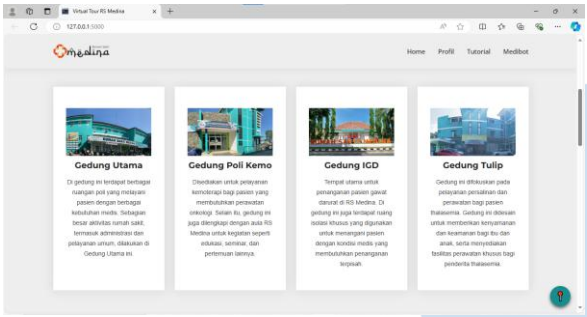



Fig. 14. Main Page Display (Building Options)

Figure 14 shows the main page (Building Selection) that allows users to start the virtual tour heading directly to a specific building.

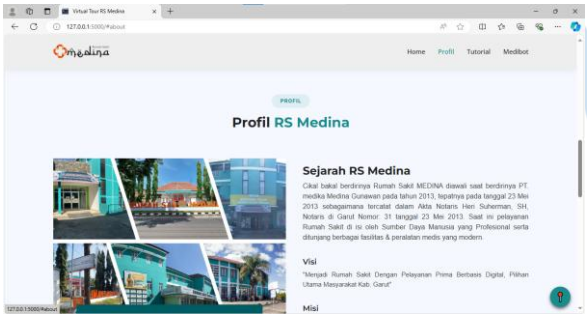



Fig. 15. Profile Page Display

Figure 15 shows the profile page that presents a brief history of Medina Hospital.

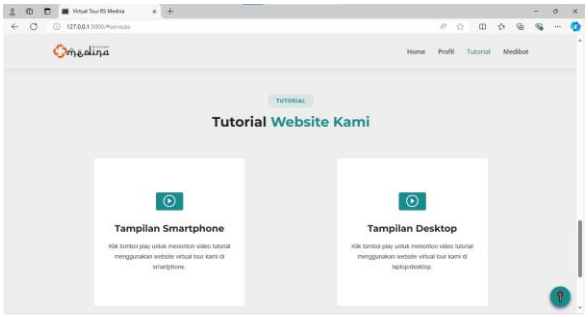



Fig. 16. Tutorial Page Display

Figure 16 shows the tutorial page using the website, both on smartphone and desktop displays.

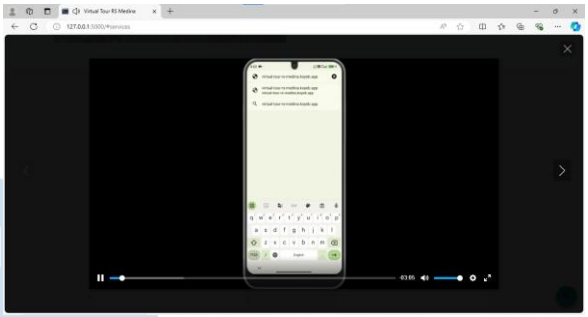


Fig. 17. Tutorial Video Pop-up Display

Figure 17 shows the pop-up video tutorial when clicking the play button.

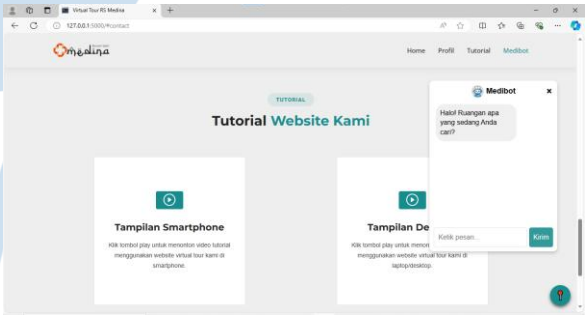



Fig. 18. Chatbot Pop-up Display

Figure 18 shows a pop-up chatbot named Medibot, where users can ask for information related to the location of rooms in Medina Hospital.

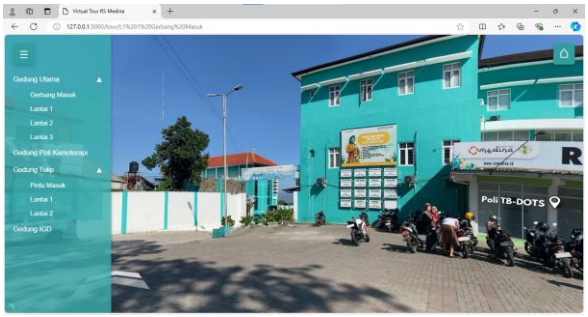



Fig. 19. Virtual Tour & Sidebar Menu Page Display

Figure 19 menunjukkan tampilan melakukan *virtual tour*, halaman ini dilengkapi dengan *sidebar* menu untuk mengakses secara langsung lokasi tertentu.

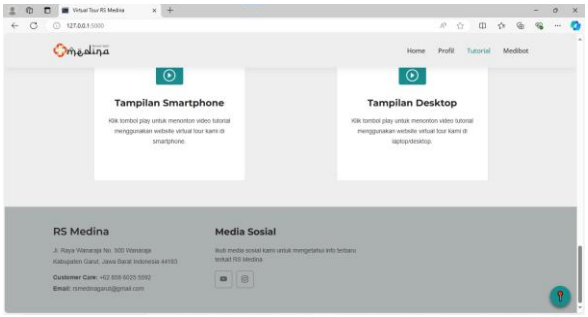



Fig. 20. Footer Display

Figure 20 shows a footer that displays clear contact information and links to Medina Hospital's official social media accounts, such as Instagram and YouTube to make it easier for users to access the latest information about Medina Hospital.

### *E. Testing*

The testing phase was carried out using the black box method. Researchers developed 26 test scenarios to ensure that all website features functioned according to the planned requirements. Some of the key scenarios included: starting the virtual tour from the main page, allowing users to access the virtual tour page after clicking the "Start Tour" button; selecting a specific building and floor, enabling users to navigate to the Main Building, Poli Kemo Building, Emergency Building, and Tulip Building, including floor-to-floor navigation within the Main and Tulip Buildings; using the Medibot chatbot, where users can inquire about the location of a specific room and the chatbot provides the appropriate response; accessing video tutorials, allowing users to watch step-by-step guides on both smartphone and desktop views and correctly close the tutorial pop-up; and Google Maps integration, enabling users to access the location of Medina Hospital directly through the Google Maps icon on the website.

Overall, the 26 blackbox test scenarios covered various user interactions, ranging from main page navigation, building and floor selection, hotspot usage, chatbot interaction, to video tutorial testing and map access. All scenarios resulted in an “OK” status which means the result is in accordance with the expected result. The overall test results show that all features on the Medina Hospital virtual tour website have functioned as expected.

### *F. Distribution*

The distribution stage aims to ensure the accessibility of the Medina Hospital virtual tour website to the public through the Internet platform. The results of the activities in this stage include the implementation and publication of the website. The Medina Hospital virtual tour website was successfully deployed and can be accessed online via the URL https://virtual-tour-rs-medina.koyeb.app/. As for the initial display when the website is visited, it can be seen in Figure 21.

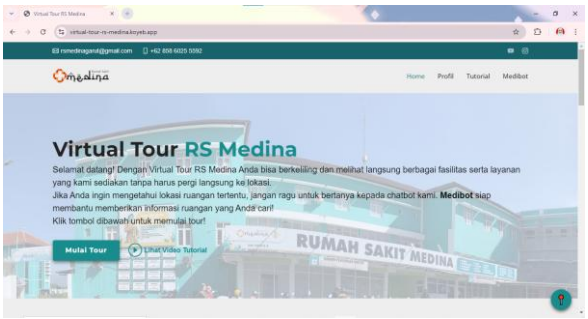



Fig. 21. Medina Hospital Virtual Tour Website Page Display

To promote the website, an online poster was created equipped with a QR Code as shown in Figure 22 which is linked to the virtual tour domain.


Fig. 22. Online Poster for Medina Hospital Virtual Tour Website

By scanning the QR Code, users can easily access the website without typing in the URL manually. This strategy not only facilitates access to information but also increases community engagement with the virtual tour services offered. Posters were distributed on Medina Hospital's official social media as well as placed in several areas of the hospital to introduce and promote the website to patients and visitors.

## IV. CONCLUSIONS

Based on the results and discussion, this study successfully built the Medina Hospital Virtual Tour Website, which is equipped with navigation features to facilitate users in exploring various areas of the hospital virtually displayed in a 360-degree panoramic format. In addition, there are additional features, such as building and floor options designed to make it easier for users to go directly to the desired location, chatbot features to help users find information about the location of the room they are looking for, also on each page is equipped with a Google Maps navigation icon that allows users to find the location of Medina Hospital. We also provide a tutorial page that is accompanied by a step-by-step video guide to using the

website on both smartphone and desktop displays, thus ensuring convenient accessibility for all users. The results of the functionality test using black box testing show that the Medina Hospital Virtual Tour Website functions optimally and meets the planned functional needs. This is evidenced by the test results which found no significant technical problems. The website was presented and directly tested by the management of Medina Hospital, who assessed that the platform is effective in presenting the hospital's facilities and services interactively. So it can be said that the website is ready to be accessed by the public and provides useful information for users. For future development, adding an indoor navigation map would help users determine their real-time location. Additionally, integrating AR (Augmented Reality) or location-based notifications could further enhance the user experience.

## ACKNOWLEDGMENT

The Authors wish to acknowledge Institut Teknologi Garut, which supports and funds this research publication

## REFERENCES

[1] Kementerian Kesehatan Indonesia, "Profil Kesehatan Indonesia Tahun 2023," Jakarta, 2023.

[2] A. E. Permanasari, D. A. Hidayat, S. Wibirama, I. S. Sakkinah, and D. R. A. Rambli, "Development of a hospital virtual tour with virtual reality-based panorama," *International Journal of Innovation and Learning*, vol. 30, no. 2, p. 119, 2021, doi: 10.1504/IJIL.2021.117218.

[3] Safwan Kasma, S. Supriadi, and S. Suhaemi, "Pengembangan Aplikasi Virtual Tour Pengenalan Lingkungan Sekolah SMK Negeri 1 Wajo Sebagai Media Informasi," *BANDWIDTH: Journal of Informatics and Computer Engineering*, vol. 2, no. 2, pp. 121–131, Jul. 2024, doi: 10.53769/bandwidth.v2i2.803.

[4] D. Dairoh, D. I. Af'idah, S. F. Handayani, R. W. Pratiwi, A. Rachman, and D. C. A. Saputra, "Pengenalan dan Pemanfaatan Aplikasi Virtual Tour sebagai Media Promosi Wisata," *JMM (Jurnal Masyarakat Mandiri)*, vol. 7, no. 1, p. 12, Jan. 2023, doi: 10.31764/jmm.v7i1.11734.

[5] Y. Anggara, G. Maulana Zamroni, A. Dahlan, J. Selatan, and D. Istimewa Yogyakarta, "Virtual Reality Tour Menggunakan Metode Gambar Panorama 360° Sebagai Media Informasi dan Pengenalan Gedung Perkuliahan Kampus 4 Universitas Ahmad Dahlan," vol. 9, no. 1, pp. 1–12, 2021, doi: 10.12928/jstie.v8i3.xxx.

[6] P. Utami and J. Jemakmun, "Aplikasi Virtual Tour Sekolah Menengah Pertama (SMP) Negeri 3 Kota Pagar Alam Berbasis Android," *Journal of Software Engineering Ampera*, vol. 2, no. 3, pp. 144–153, Oct. 2021, doi: 10.51519/journalsea.v2i3.124.

[7] Vicky C. Mende, Quido C. Kainde, and Ferdinan I. Sangkop, "Virtual Tour Pariwisata Kelurahan Lahendong Berbasis Web Menggunakan Metode Prototyping," *Jurnal Penelitian Rumpun Ilmu Teknik*, vol. 2, no. 2, pp. 187–199, May 2023, doi: 10.55606/juprit.v2i2.1963.

[8] I. Miljkovic, O. Shlyakhetko, and S. Fedushko, "Real Estate App Development Based on AI/VR Technologies," *Electronics (Basel)*, vol. 12, no. 3, p. 707, Jan. 2023, doi: 10.3390/electronics12030707.

[9] H. Rahaman, E. Champion, and D. McMeekin, "Outside Inn: Exploring the Heritage of a Historic Hotel through 360-Panoramas," *Heritage*, vol. 6, no. 5, pp. 4380–4410, May 2023, doi: 10.3390/heritage6050232.

[10] I. Febrianto and M. S. Putra, "Aplikasi Virtual Tour Sebagai Pengenalan Objek Rekreasi Keluarga Pada 3D Stable," *Jurnal Ilmiah Komputasi*, vol. 23, no. 1, pp. 113–120, 2024.

[11] U. Nafi'ah, A. Sapto, J. Sayono, A. Herdiyani, and G. Smith, "The Innovation of Virtual Tour of Malang State University Learning Museum as an Alternative for Educational Tourism in the Disruptive Era," *KnE Social Sciences*, pp. 117–126, 2023.

[12] M. Gemilang Ramadhan *et al.*, "Pengembangan Aplikasi Monitoring Kondisi Tanaman Berbasis Markerless Augmented Reality dengan Metode MDLC," *Jurnal Ilmiah MEDIA SISFO*, vol. 19, no. 2, 2025, doi: 10.33998/mediasisfo.2019.19.2.2570.

[13] D. Kurniadi, L. Fitriani, E. Satria, and A. Rahman, "Multimedia system model for electrical circuits on android mobile devices," *IOP Conf Ser Mater Sci Eng*, vol. 1098, no. 3, p. 032092, Mar. 2021, doi: 10.1088/1757-899X/1098/3/032092.

[14] A. C. Luther, *Authoring Interactive Multimedia*. in IBM tools series. AP Professional, 1994. [Online]. Available: https://books.google.co.id/books?id=gpULAQAAMAAJ

[15] F. N. Kumala, A. Ghufron, P. P. Astuti, M. Crismonika, M. N. Hudha, and C. I. R. Nita, "MDLC model for developing multimedia e-learning on energy concept for primary school students," *J Phys Conf Ser*, vol. 1869, no. 1, p. 012068, Apr. 2021, doi: 10.1088/1742-6596/1869/1/012068.

[16] M. Plainer, "Practical Study of Visual Studio Code Practical Course—Contributing to an Open-Source Project," 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:227995553

[17] M. Sholeh, I. Gisfas, Cahiman, and M. A. Fauzi, "Black Box Testing on ukmbantul.com Page with Boundary Value Analysis and Equivalence Partitioning Methods," *J Phys Conf Ser*, vol. 1823, no. 1, p. 012029, Mar. 2021, doi: 10.1088/1742-6596/1823/1/012029.

[18] Uminingsih, M. Nur Ichsanudin, M. Yusuf, and S. Suraya, "Pengujian Fungsional Perangkat Lunak Sistem Informasi Perpustakaan dengan Metode Blackbox Testing bagi Pemula," *STORAGE: Jurnal Ilmiah Teknik dan Ilmu Komputer*, vol. 1, no. 2, pp. 1–8, May 2022, doi: 10.55123/storage.v1i2.270.

[19] Koyeb Cloud Platform (Version 5.6.0) [Serverless hosting platform and command-line interface], "Cloud Platform," 2025, 5.6.0. Accessed: Aug. 24, 2024. [Online]. Available: https://www.koyeb.com

ISSN 2085-4552

# Identifying Academic Performance Patterns Among PTIK Students Using K-Means Clustering

Rizak Al Hasbi Anwar[1], Febri Liantoni[2], Dwi Maryono[3]
[1,2,3] Dept. of Informatics Education, Sebelas Maret University, Indonesia
[1]rizakalhasbi_22@student.uns.ac.id, [2]Febri.liantoni@gmail.com, [3]dwimaryono@staff.uns.ac.id

Accepted 23 June 2025
Approved 06 January 2026

***Abstract*— This study explores the identification of academic performance patterns among students in the Informatics and Computer Engineering Education Study Program (PTIK) at Sebelas Maret University, focusing on the 2022 cohort. Using the K-Means clustering method within the scope of Data Mining, this research analyzes student performance data across multiple course categories from the first to fourth semesters. Through the Elbow method, four optimal clusters were established, each representing distinctive patterns of academic achievement. The analysis was conducted using RapidMiner software to reveal nuanced insights into student learning outcomes. Cluster 1 consists of students with moderate achievements in most categories, with a particular strength in Multimedia. Cluster 2 includes students with generally lower academic performance but shows a relative strength in General Courses. Cluster 3 is composed of high-achieving students who excel across categories, particularly in Software Engineering (RPL), Multimedia, and Educational subjects, indicating well-rounded academic proficiency. Cluster 4 comprises students with notable strengths in Software Engineering and Computer Networking, yet demonstrates lower performance in certain specialized subjects. These findings highlight the potential to tailor educational programs to address the specific learning needs and strengths of each student group, facilitating more personalized and effective academic support.**

***Index Terms*— *Clustering; K-Means; Data Mining; Academic Performance; RapidMiner.***

## I. INTRODUCTION

In the rapidly evolving field of Informatics and Computer Engineering, understanding student performance patterns has become increasingly important. Accurate identification of academic achievements and potential areas of improvement among students can significantly enhance educational strategies and contribute to more personalized learning experiences. Academic institutions are now leveraging data mining techniques to identify hidden trends in student performance data, which can help develop targeted interventions and improve overall learning outcomes. The increasing availability of student data has driven the adoption of data mining techniques to support academic decision-making and personalize learning [1]. The advancement of data mining technologies has influenced many researchers to investigate more profound insights into the knowledge dissemination process [2]. The application of data mining methods in the field of education has attracted great attention in recent years [3]. Data mining will certainly be very useful to analyze the activities of students who succeed and those who are at risk of failing, to develop improvement strategies based on students' academic performance, and therefore to assist educators in the development of pedagogical methods [4]. One such technique, K-Means clustering, has shown promise in grouping students based on similar characteristics, allowing for a detailed analysis of their academic tendencies and strengths. The K-Means Clustering algorithm is a widely used data analysis tool that divides data into many clusters according to shared attributes [5]. With the emergence of Big Data and the increased availability of educational datasets, clustering techniques can now be applied to identify distinct groups of students based on their academic performance across multiple subjects. These techniques are increasingly utilized in higher education to uncover hidden patterns in student achievement, enabling more effective academic planning and resource allocation [6].

However, in the context of higher education in Indonesia, particularly in programs such as Informatics and Computer Engineering Education (PTIK), there remains a lack of comprehensive, data-driven analysis regarding student academic achievement patterns. Most academic evaluations still rely on cumulative GPA or individual course performance without identifying group-based trends that can inform strategic interventions. Identifying these patterns is crucial because it allows institutions to detect underperforming groups early, tailor pedagogical approaches, and provide more equitable academic support [7]. Moreover, with the increasing complexity of interdisciplinary courses within PTIK, spanning

software engineering, multimedia, and education science, students often exhibit diverse capabilities across subject categories. Without a structured analysis, such diversity may go unnoticed and unaddressed. Therefore, this research seeks to fill the gap by identifying academic performance patterns among PTIK students through clustering analysis. This will not only inform curriculum improvements but also support institutional decision-making for targeted academic interventions.

Informatics and Computer Engineering Education (PTIK) is a challenging program that combines technical, theoretical, and practical skills, requiring students to perform well across diverse subject areas. With the emergence of Big Data and the increased availability of educational datasets, clustering techniques can now be applied to identify distinct groups of students based on their academic performance across multiple subjects. Clustering method, especially K-Means, is one of the commonly used techniques in data mining to categorize data into groups based on similar features [8]. In particular, K-Means clustering enables a structured analysis of students' grades, providing insights into their strengths and weaknesses across different domains. Similarities between students can be found through clustering analysis of student evaluation scores using the K-means technique [9]. In a study titled "K-Means Algorithm for Grouping Student Thesis Topics," the K-Means algorithm was also used. This led to the grouping of students based on their areas of expertise. The grouping with the highest cluster indicates that the students are proficient in each group of areas of expertise, allowing them to select essay topics that are appropriate for their group of areas of expertise [10]. By clustering students according to academic performance in various categories, institutions can gain a deeper understanding of the distribution of student achievement and the specific challenges faced by certain groups.

One technique that can be used to maximize the k-means method in forming or determining the number of clusters is the silhouette coefficient [11]. To evaluate the quality of clustering, this study also utilizes the Silhouette Score, a metric that measures how well each data point is separated from others in different clusters. Although the Silhouette Scores indicate that some clusters have stronger separations than others, the clustering approach provides a foundational understanding of the varied academic achievements among PTIK students. The results offer valuable insights that can be used to tailor educational programs, such as targeted support for students in lower-performing clusters and advanced projects for high achievers. Ultimately, this research highlights the potential of data mining in identifying and addressing student needs, creating a pathway for more effective and individualized learning interventions.

Through this research, it is expected that the insights gained will inform the design of learning policies and support programs that meet the specific needs of each student group, fostering a more adaptive and supportive educational environment in the PTIK program.

## II. METHOD

With an emphasis on the 2022 cohort, this study employs a quantitative methodology to examine academic performance trends among students enrolled in Sebelas Maret University's Informatics and Computer Engineering Education (PTIK) program. The main technique is data mining using K-Means clustering to find unique student groups according to their academic achievement in a variety of categories.

### *A. Data Collection*

The dataset used in this study comprises the academic records of all students enrolled in the 2022 cohort of the Informatics and Computer Engineering Education (PTIK) program at Sebelas Maret University. This dataset includes a total of 87 students, covering academic grades from the first to fourth semesters. The data consists of average scores derived from multiple subjects that are grouped into specific course categories, such as Software Engineering (RPL), Multimedia, Education, General Courses, and Expertise-related courses. Each student's performance in these categories serves as the basis for the clustering analysis.

The academic data used in this study were obtained from the PTIK UNS Study Program administrator in the form of student transcripts. Each course was then grouped into one of six predefined categories based on the academic focus of the subject: Software Engineering (RPL), Computer Networking, Multimedia, Expertise, Education, and General Courses. The grouping was conducted by referring to the official course descriptions published on the PTIK Study Program's academic website: https://ptik.fkip.uns.ac.id/akademik/daftar-mata-kuliah/. This classification ensured that related courses were analyzed collectively within their respective domains. Grouping features based on academic content is a commonly applied practice in educational data mining to increase interpretability and analytical relevance [12].

### *B. Data Preprocessing*

To guarantee the precision and dependability of the clustering procedure, data pretreatment was carried out. This involved classifying courses into predetermined groups, managing missing numbers, and normalising grade scales. To prepare the dataset for clustering analysis, several preprocessing steps were conducted: categorizing course types, handling missing values, and normalising grade scales. Incomplete records were excluded to preserve data integrity. This rigorous

preprocessing ensures that the input is suitable for clustering algorithms like K-Means. Proper data preprocessing, such as normalization and removal of incomplete entries, is fundamental in clustering educational datasets to achieve accurate and unbiased student groupings [13]. To preserve the quality of the data, any incomplete records were not included in the study. This method focuses on transforming the dataset to ensure it is suitable for clustering algorithms and data mining tools [14]. Preprocessing steps such as normalization and missing value handling are crucial to ensure data consistency and improve clustering results [15]. Once the data is collected, a data cleaning process is performed to deal with missing or incomplete data [16].

### C. *Clustering Method: K-Means*

The efficiency of the K-Means clustering method in dividing data points into discrete clusters according to similarity led to its selection. For the clustering procedure, the following actions were taken:

1. Finding the Ideal Number of Clusters: The Elbow Method, which examines the sum of squared distances inside clusters, was used to determine the ideal number of clusters. Four clusters were identified as the best arrangement for this dataset using this strategy.
2. Implementation of Clustering: RapidMiner, a data mining program that makes it easier to use clustering algorithms effectively, was used to carry out the clustering. In order to identify performance trends within each cluster, students were categorised according to how similar their academic performance was throughout the designated areas.
3. Interpretation of Cluster Characteristics: After generating the clusters, each was analyzed to identify its specific characteristics based on average performance in each course category. This analysis aimed to provide insights into the academic tendencies of students in each cluster.

### D. *Cluster Evaluation: Silhouette Score*

To assess the quality of the clusters, the Silhouette Score was calculated for each cluster, measuring how similar each data point is to its assigned cluster compared to other clusters. The score ranges from -1 to 1, with higher values indicating better-defined clusters. The Silhouette coefficient is one of the most commonly used metrics to evaluate clustering quality, especially for methods like K-Means [17]. This evaluation provided a measure of how well-separated and homogenous the clusters were, with Cluster 4 exhibiting the highest score (0.458), indicating a stronger separation than the others.

## III. Result And Discussion

The K-Means clustering analysis results are shown in this section along with a discussion of the unique patterns of academic achievement seen among PTIK students in the 2022 cohort. To determine the optimal number of clusters, the Elbow Method was applied by plotting the Within-Cluster Sum of Squares (WCSS) against various values of k, ranging from k=2 to k=10. The Elbow Method helps identify the point at which increasing the number of clusters yields diminishing returns in variance reduction. Based on the plotted curve, a clear "elbow" was observed at k=4, indicating that four clusters offer the best trade-off between clustering accuracy and model simplicity. This approach is widely accepted and commonly used in educational data mining to select the appropriate number of clusters [18]. Therefore, k=4 was chosen as the optimal number of clusters for further analysis. Following this, the initial centroids were randomly selected from the student dataset, and each cluster centroid represents a set of averaged scores across four categorized academic domains.

Although Clusters 1, 2, and 3 produced relatively low Silhouette Scores (ranging from 0.149 to 0.190), this does not necessarily indicate poor clustering quality, but rather reflects overlapping characteristics among some student groups. In educational datasets, especially those derived from multi-dimensional academic records, moderate to low Silhouette Scores are common due to the inherent complexity and interdependence of performance attributes. This phenomenon has been observed in studies applying fuzzy clustering to student data, where overlapping academic profiles result in lower separation metrics [19].

The decision to retain k=4 clusters was made based on the Elbow Method, which clearly indicated a significant inflection point at that value. This choice balances model simplicity and information capture. Nonetheless, to enhance cluster quality and robustness, future studies may consider alternative clustering algorithms. For instance, comparative analyses between K-Means and DBSCAN in educational settings have shown DBSCAN can better identify overlapping clusters and handle noise, often yielding higher Silhouette Scores . Additionally, Fuzzy C-Means allows soft membership assignments, which is useful for datasets where students may exhibit blended performance profiles . A systematic comparison using internal validation metrics like Davies–Bouldin or Calinski–Harabasz could further elucidate the optimal approach for clustering academic performance data.

### A. *Clustering Results*

Based on the clustering analysis conducted in RapidMiner, the Elbow Method indicated that four clusters would provide the most meaningful insights. The clusters were analyzed for distinct performance

characteristics across four academic categories: Software Engineering (RPL), Multimedia, Education, and General Courses. Although this study primarily focuses on descriptive cluster interpretation, the observed mean differences between clusters in each academic category indicate potentially significant group distinctions. These distinctions are further supported by the Silhouette Score values, which reflect intra-cluster homogeneity and inter-cluster separation, as follows:

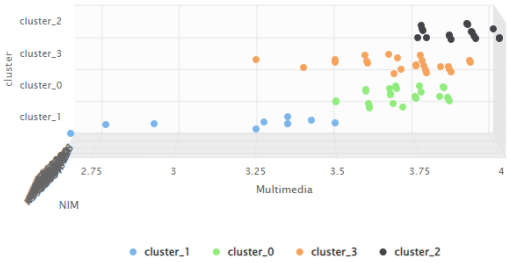



**Fig. 1. Cluster 0 Visualization**

Cluster 0: This cluster consists of students with moderate academic achievements across most categories, with a notable strength in Multimedia. Students in this cluster display average performance in both general and specific informatical courses. Their relatively higher performance in Multimedia suggests a potential interest or aptitude in that field, though they may benefit from additional support in other areas to achieve a more balanced skill set.

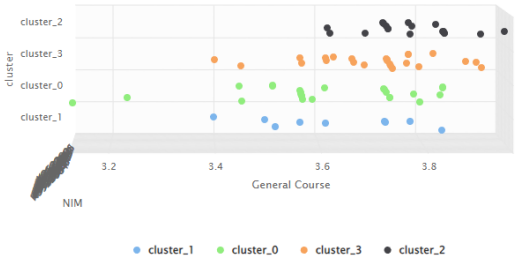



**Fig. 2. Cluster 1 Visualization**

Cluster 1: Students in Cluster 1 score comparatively better in General Courses but have the lowest total academic accomplishments. This pattern suggests that while students in this cluster may struggle in more complex informatics-related courses, they excel in basic or non-technical courses. To bridge performance disparities, this group can profit from focused academic assistance in informatics-related courses like computer networking and software engineering.

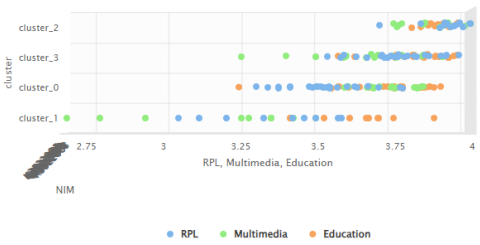



**Fig. 3. Cluster 2 Visualization**

Cluster 2: Students in Cluster 2 perform well in many areas, but especially in education, multimedia, and software engineering (RPL). Students in this cluster have shown a high level of academic proficiency and a broad range of skills, indicating that they can successfully understand both technical and academic topics. To help lower-performing pupils and expand their knowledge, these students can be eligible for advanced projects or peer mentoring.

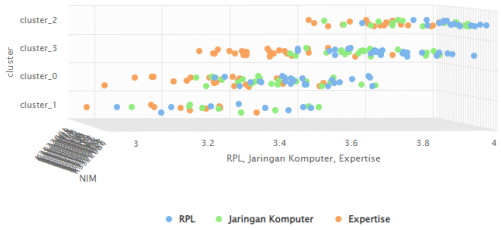



**Fig. 4. Cluster 3 Visualization**

Cluster 4: Cluster 3 includes students who excel in Software Engineering (RPL) and Computer Networking but have relatively lower achievements in the Expertise category. This cluster indicates a preference or strength in specific technical domains rather than a broader expertise across Informatics fields. Students in this cluster might benefit from exposure to additional resources in the Expertise category to develop a more comprehensive skill set.

### *B. Cluster Evaluation Using Silhouette Score*

To assess the internal quality of each cluster, a Silhouette Score was calculated for each, yielding the following results:

**Table I. Silhouette Score**

| | **Silhouette scores** |
|---|---|
| Cluster 0 | 0,149 |
| Cluster 1 | 0,190 |
| Cluster 2 | 0,181 |
| Cluster 3 | 0,458 |

A moderate degree of separation across clusters is shown by the Silhouette Score values. With the highest Silhouette Score (0.458), Cluster 3 showed more uniform traits and distinct distinction between its members. Clusters 0, 1, and 2, on the other hand, received comparatively low ratings, indicating that some of the students in these clusters have traits in common with those of other clusters. This might be because academic performance in other categories is similar, which could indicate that some disciplines have overlapping learning patterns. While K-Means provides an efficient and straightforward approach, it has notable limitations in this context. First, it assumes spherical and equally sized clusters, which may not reflect the actual distribution of student academic performance that often exhibits varied shapes or densities. Second, K-Means is highly sensitive to outliers and noisy data, extreme student scores can disproportionately influence centroid positions and skew cluster assignments [20]. Third, the algorithm requires a predefined number of clusters ($k$), which might not always align with the natural groupings in educational datasets.

These limitations could impact the reliability of the clustering results, particularly in how borderline students or outlier performances are categorized. For instance, students with mixed academic profiles may be forced into clusters that do not accurately reflect their learning trajectories, and outliers may distort cluster centroids. Consequently, findings based solely on K-Means should be interpreted with caution. To mitigate these issues, future research could explore more robust alternatives such as DBSCAN, which can handle non-spherical clusters and identify noise; Fuzzy C-Means, which accommodates overlapping cluster memberships; or K-Medoids, which is less influenced by outliers. Comparative evaluations using internal metrics (e.g., Davies-Bouldin, Calinski-Harabasz) and outlier-aware variants of K-Means would further strengthen the validity of conclusions drawn from cluster analysis.

### C. *Interpretation of Clusters and Educational Implications*

Each cluster provides insights into students' academic needs and potential areas for targeted support or enhancement, as detailed below:

- Cluster 0: The moderate achievements of students in this cluster, coupled with strength in Multimedia, suggest a need for academic support in other Informatics fields. Interventions could focus on strengthening skills in core Informatics subjects to achieve a balanced academic profile, while fostering their interest in Multimedia through specialized projects or resources.
- Cluster 1: As the cluster with the lowest academic performance, yet with relative strength in General Courses, Cluster 1 may benefit from an intensive support program in technical subjects. This might include foundational workshops, tutoring in Software Engineering and Computer Networking, or remedial courses to build a stronger foundation in Informatics-related subjects.
- Cluster 2: Students in this cluster demonstrate balanced, high performance across categories, positioning them as candidates for advanced learning opportunities. Providing mentorship roles or participation in research projects could not only further develop their skills but also enhance the learning experience for other clusters, particularly Clusters 0 and 1.

Cluster 3: This group's high performance in Software Engineering and Computer Networking suggests a focused interest in specific technical areas. Targeted support in Expertise-related courses could be valuable in expanding their academic breadth, while opportunities for deeper specialization in their areas of strength could also be beneficial.

## IV. CONCLUSIONS

The clustering results from this study offer practical insights that can assist academic advisors and program managers in tailoring support strategies for students. By identifying groups of students with similar academic performance patterns, advisors can more effectively design personalized interventions. For instance, students in clusters characterized by strong performance in technical domains but lower results in general or education-related courses may benefit from academic writing workshops or soft-skills enhancement programs. Conversely, students in clusters with strong general course performance but low scores in Multimedia or RPL could be offered targeted tutoring or peer mentoring in those specific areas.

The interpretation of cluster characteristics in this study was conducted based on domain understanding of the PTIK curriculum structure and academic performance expectations across categorized subjects. Although no formal external validation such as expert review or stakeholder consultation was conducted, the course groupings and observed performance patterns were analyzed in reference to official course descriptions and academic benchmarks outlined by the PTIK study program. This approach ensures that the clustering results maintain practical relevance to real-world academic settings. Nevertheless, we acknowledge the importance of involving domain experts, such as curriculum designers or academic advisors, to validate the semantic coherence of each cluster, particularly in confirming whether groupings align with observable student learning trends. Future research could enhance the robustness of interpretation by incorporating expert validation sessions or using labeled data to conduct external cluster validation.

In order to identify discrete student groups based on academic performance across many course categories, this study used K-Means clustering to examine the academic performance patterns of PTIK students at Sebelas Maret University, specifically among the 2022 cohort. Four distinct clusters, each of which represented distinct academic tendencies and strengths, were produced by applying the Elbow Method to determine the ideal number of clusters. The clustering results provide valuable information on the academic profiles of the students, identifying areas that could use focused educational support.

Overall, this study demonstrates the potential of data-driven methods like K-Means clustering to inform academic support strategies, ensuring that student needs are addressed with precision and relevance. By leveraging insights from such clustering analyses, educational institutions can create more targeted learning environments, ultimately contributing to improved student outcomes and more effective curriculum planning.

For future improvements, it is recommended to explore hybrid approaches that combine clustering with classification methods. For example, Oyelade et al. (2010) found that integrating K-Means clustering with deterministic/statistical models significantly enhanced performance prediction accuracy—suggesting a hybrid clustering–classification pipeline could be beneficial in profiling PTIK students [21]. Additionally, alternative clustering techniques such as Fuzzy C-Means or Self-Organizing Maps (SOM) can be tested to capture non-crisp student membership and visualize multi-dimensional academic profiles more intuitively. Rffectiveness of integrating multi-dimensional feature fusion in student performance analysis, suggesting that combining temporal and spatial features can lead to more nuanced clustering insights [22]. Future studies may also incorporate statistical validity checks like silhouette analysis or ANOVA to quantitatively verify cluster separation and include longitudinal data to monitor how student cluster membership evolves over time.

## ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to Sebelas Maret University, particularly the Informatics and Computer Engineering Education Study Program (PTIK), for their support and for providing the data necessary to conduct this research.

## REFERENCES

[1] Romero C, Ventura S. Educational data mining and learning analytics: An updated survey. *WIREs Data Mining Knowl Discov*. 2020; 10:e1355. https://doi.org/10.1002/widm.1355.

[2] Khan, A., Ghosh, S.K. Student performance analysis and prediction in classroom learning: A review of educational data mining studies. *Educ Inf Technol* 26, 205–240 (2021). https://doi.org/10.1007/s10639-020-10230-3.

[3] Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, *9*(1). https://doi.org/10.1186/s40561-022-00192-z

[4] Simangunsong, B. N., Manalu, M. R., & Medan, U. I. (2023). Testing the K-Means Clustering Algorithm in Processing Student Assignment Grades Using the RapidMiner Application. *Journal of Data Science*, *1*, 51–60.

[5] Casquero, O., Ovelar, R., Romo, J., Benito, M., & Alberdi, M. (2016). Students' personal networks in virtual and personal learning environments: a case study in higher education using learning analytics approach. *Interactive Learning Environments*, *24*(1), 49–67. https://doi.org/10.1080/10494820.2013.817441

[6] B. Minaei-Bidgoli, D. A. Kashy, G. Kortemeyer and W. F. Punch, "Predicting student performance: an application of data mining methods with an educational Web-based system," *33rd Annual Frontiers in Education, 2003. FIE 2003.*, Westminster, CO, USA, 2003, pp. T2A-13, doi: 10.1109/FIE.2003.1263284

[7] Khan, A., Ghosh, S.K. Student performance analysis and prediction in classroom learning: A review of educational data mining studies. *Educ Inf Technol* **26**, 205–240 (2021). https://doi.org/10.1007/s10639-020-10230-3

[8] Rahayu, N. D., Anshor, A. H., & Afriantoro, I. (2024). Penerapan Data Mining untuk Pemetaan Siswa Berprestasi menggunakan Metode Clustering K-Means. JUKI : Jurnal Komputer Dan Informatika, 6(1), 71–83. https://doi.org/10.53842/juki.v6i1.474

[9] Liu, R. (2022). Data Analysis of Educational Evaluation Using K-Means Clustering Method. Computational Intelligence and Neuroscience, 2022. https://doi.org/10.1155/2022/3762431

[10] M. R. Muttaqin and M. Defriani, "Algoritma K-Means untuk Pengelompokan Topik Skripsi Mahasiswa," Ilk. J. Ilm., vol. 12, no. 2, pp. 121–129, 2020, doi: 10.33096/ilkom.v12i2.542.121-129

[11] Hartama, D., & Anjelita, M. (2022). Analysis of Silhouette Coefficient Evaluation with Euclidean Distance in the Clustering Method (Case Study: Number of Public Schools in Indonesia). Jurnal Mantik, 6(3), 3667–3677. https://iocscience.org/ejournal/index.php/mantik/article/view/3318.

[12] Mashael A. Al-Barrak and Muna Al-Razgan, "Predicting Students Final GPA Using Decision Trees: A Case Study," *International Journal of Information and*

*Education Technology* vol. 6, no. 7, pp. 528-533, 2016, DOI: 10.7763/IJIET.2016.V6.745

[13] et. al., M. V. . (2021). Evaluating Students Placement Performance Using Normalized K-Means Clustering Algorithm. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *12*(11), 3785–3792. https://doi.org/10.17762/turcomat.v12i11.6488

[14] Mohamed Nafuri, A. F., Sani, N. S., Zainudin, N. F. A., Rahman, A. H. A., & Aliff, M. (2022). Clustering Analysis for Classifying Student Academic Performance in Higher Education. *Applied Sciences (Switzerland)*, *12*(19). https://doi.org/10.3390/app12199467

[15] Lhoucine Bahatti, Omar Bouattane, My Elhoussine Echhibat and Mohamed Hicham Zaggaf, "An Efficient Audio Classification Approach Based on Support Vector Machines" International Journal of Advanced Computer Science and Applications(ijacsa), 7(5), 2016. http://dx.doi.org/10.14569/IJACSA.2016.070530

[16] Al-Hagery, M. A., Alzaid, M. A., Alharbi, T. S., & Alhanaya, M. A. (2020). Data Mining Methods for Detecting the Most Significant Factors Affecting Students' Performance. *International Journal of Information Technology and Computer Science*, *12*(5), 1–13. https://doi.org/10.5815/ijitcs.2020.05.0

[17] Peter J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics, Volume 20, 1987, Pages 53-65, ISSN 0377-0427, https://doi.org/10.1016/0377-0427(87)90125-7.

[18] KETCHEN, D.J. and SHOOK, C.L. (1996), THE APPLICATION OF CLUSTER ANALYSIS IN STRATEGIC MANAGEMENT RESEARCH: AN ANALYSIS AND CRITIQUE. Strat. Mgmt. J., 17: 441-458. https://doi.org/10.1002/(SICI)1097-0266(199606)17:6<441::AID-SMJ819>3.0.CO;2-G

[19] Ovtšarenko, O. Opportunities of machine learning algorithms for education. *Discov Educ* **3**, 209 (2024). https://doi.org/10.1007/s44217-024-00313-5

[20] Iliyas Karim khan, Hanita Binti Daud, Nooraini binti Zainuddin, Rajalingam Sokkalingam, Abdussamad, Abdul Museeb, Agha Inayat. Addressing limitations of the K-means clustering algorithm: outliers, non-spherical data, and optimal cluster selection[J]. AIMS Mathematics, 2024, 9(9): 25070-25097. doi: 10.3934/math.20241222

[21] O. J. Oyelade, O. O. Oladipupo, I. C. Obagbuwa. Application of k Means Clustering algorithm for prediction of Students Academic Performance. International Journal of Computer Science and Information Security, IJCSIS, Vol. 7, No. 1, pp. 292-295. https://doi.org/10.48550/arXiv.1002.2425

[22] Luo, Z., Mai, J., Feng, C., Kong, D., Liu, J., Ding, Y., Qi, B., & Zhu, Z. (2024). A Method for Prediction and Analysis of Student Performance That Combines Multi-Dimensional Features of Time and Space. Mathematics, 12(22), 3597. https://doi.org/10.3390/math12223597

ISSN 2085-4552

# Sentiment Analysis of University X Students: Comparing Naive Bayes and BERT Approaches

Jonathan David[1], Kie Van Ivanky Saputra[1], Andry Manodotua Panjaitan[2], Feliks Victor Parningotan Samosir[3]
[1] Mathematics Department, Faculty of Science and Technology (FaST), Universitas Pelita Harapan, Tangerang, Banten 15811, Indonesia
[2] Industrial Engineering Department, Faculty of Science and Technology (FaST), Universitas Pelita Harapan, Tangerang, Banten 15811, Indonesia
[3] Informatics Department, Faculty of Information Technology (FIT), Universitas Pelita Harapan, Tangerang, Banten 15811, Indonesia
[1] joda48614@gmail.com, [2] kie.saputra@uph.edu, [3] andry.panjaitan@uph.edu, [4] feliks.parningotan@uph.edu

Accepted 12 August 2025
Approved 06 January 2026

***Abstract*— Student satisfaction with university facilities and services requires in-depth analysis to ensure improvements in unsatisfactory facilities or services while maintaining those that meet expectations. This study aims to analyze sentiment in student satisfaction surveys using Natural Language Processing (NLP) methods. Survey data collected from 2022 to 2024 were analyzed using two main approaches: Naive Bayes (NB) with $n$-grams ($n = 1, 2, 3$) employing feature extraction methods such as Term Frequency-Inverse Document Frequency (TF-IDF) and Bag of Words (BoW), and Bidirectional Encoder Representations from Transformers (BERT). The analysis reveals that BERT achieves higher sentiment prediction accuracy than NB, with an F1-score of 0.777 compared to NB's 0.676 (a difference of 0.101), though this improvement margin is not statistically significant. This study also identified keywords for both positive and negative sentiments. These keywords were then analyzed across 11 categories of facilities and services to provide focused insights into aspects that need to be maintained or improved. This study concludes that sentiment analysis provides significant contributions to universities in evaluating and enhancing the quality of facilities and services according to student preferences.**

***Index Terms*— Student Satisfaction; Sentiment Analysis; NLP; NB; BERT; $n$-gram; TF-IDF; BoW; University Facilities and Services.**

## I. INTRODUCTION

Higher education institutions play a pivotal role in cultivating students' soft and hard skills, as well as their competitiveness, by offering a range of facilities and services. When adequate facilities and services are in place, students are empowered to fully actualize their personal potential through various opportunities. The Student Satisfaction Inventory (SSI) is a widely employed instrument in assessing student satisfaction with the array of facilities and services provided by universities. Student satisfaction with university facilities and services is a critical factor that can influence their overall performance and experience [1].

The SSI has been developed as an instrument specifically designed to measure student satisfaction with various aspects of campus life. The SSI was developed by Ruffalo Noel Levitz, an educational consulting and satisfaction measurement tool development organization. The SSI covers various aspects that students consider important, such as the enrollment process, teaching quality, facilities, campus environment, security, effectiveness of academic advising, and others [2].

The analysis of student comments is a complex process, particularly when dealing with substantial quantities of qualitative data. The diversity in the backgrounds and disciplines of students contributes to the complexity of the task, as each student expresses their thoughts and ideas in a unique manner, which introduces subjectivity into the data [3]. This diversity poses a significant challenge in standardizing the analysis and ensuring the validity and reliability of the results [4].

Sentiment analysis employing the neural network (NN) approach processes sentences that fall into the category of unstructured data [5]. NN is applied to process and analyze data using two main approaches: supervised and unsupervised learning. In the supervised learning approach, NN is built and trained using labeled data to classify sentences into positive or negative sentiment categories. The unsupervised learning approach attempts to classify data without the need for labels [6].

The sentiment analysis research will be conducted using a supervised learning approach, as labeled data has been collected in this research. There are various supervised learning approach methods for sentiment analysis, such as Naive Bayes (NB), Support Vector Machine (SVM), Long Short-Term Memory Networks (LSTM), Bidirectional Encoder Representations from Transformers (BERT), and many more. Therefore, this research will compare and analyze the NB and BERT methods on labeled data collected by SSI University X from 2022 to 2024. The NB and BERT methods are applied by finding the best parameters to achieve the highest level of accuracy in performing sentiment analysis. By using data from SSI University X from 2022 to 2024, it is expected that the results and analysis obtained in the research can be in line with reality.

## II. THEORY

### A. Sentiment Analysis

Sentiment analysis is the process of extracting and assessing the emotional tone of text messages to understand human opinion or behavior. Sentences are analyzed and processed by separating the words to determine whether the sentiment is positive, neutral, or negative. The benefit of sentiment analysis is that it helps in understanding other people's views on a phenomenon [7].

### B. Text Preprocessing

Text preprocessing constitutes a critical step in the analysis process, with the objective of averting substantial deterioration in its performance [8]. Text preprocessing is divided into several stages, such as data cleaning, case folding, tokenizing, and stop words removal. The stages involved in this process are described as follows:

1) Data Cleaning: In this stage, the data is cleaned by removing characters such as symbols. Additionally, punctuations and numbers are also removed. The objective of this step is to minimize disruptions in the classification results [9].
2) Case Folding: This stage involves converting text into a uniform format, specifically by converting all text to lowercase [9].
3) Tokenizing: Sentences are broken down into individual words, known as tokens [9].
4) Stop Words Removal: Stop words are words that occur with high frequency but possess minimal semantic significance. These irrelevant common words are identified, flagged, and removed from the text, resulting in a cleaner text corpus [9].

### C. Term Frequency – Inverse Document Frequency (TF-IDF)

TF-IDF is a combined method of TF and IDF that produces a combined weight for each term in each document [10]. The formula for calculating TF-IDF is as follows:

$$TFIDF(t_i, D_j) = TF(t_i, D_j)IDF(t_i), \quad (1)$$

$$TF(t_i, D_j) = f_{i,j}, \quad (2)$$

$$IDF(t_i) = \log_{10}\left(\frac{N}{df(t_i)}\right), \quad (3)$$

where $TF(t_i, D_j)$ is the TF of term $t_i$ in document $D_j$, $IDF(t_i)$ is the IDF of term $t_i$, $t_i$ is the $i$-th term, $D_j$ is the $j$-th document, and $f_{i,j}$ is the number of occurrences of term $t_i$ in document $D_j$. Index $i$ ranges from 1 to $V$ and index $j$ ranges from 1 to $N$, where $V$ is the number of unique vocabularies in a set of documents and $N$ the total documents.

### D. Bag of Words (BoW)

The Bag-of-Words (BoW) method quantifies word frequencies in a document by disregarding word order. It constructs a dictionary of unique words from a document set and represents each document as a vector, where each element corresponds to a word's frequency. Although BoW ignores word order, it effectively captures topic prevalence and sentiment patterns across documents [11].

### E. n-gram

The n-gram method captures word order by analyzing the frequency of consecutive word sequences (defined by n). Unlike BoW, which tracks single words, n-grams generate a dictionary of unique word combinations. Each document is then represented as a vector, where elements indicate the count of these n-grams. This approach preserves contextual relationships between words, offering richer linguistic insights [12].

### F. Naive Bayes (NB)

NB classification is a classification method that is both simple and efficient. It is known for its ease of implementation. NB classification is based on Bayes' theorem, where the term "naïve" refers to the assumption that the features in the dataset are mutually independent [13]. The formula for calculating NB is as follows:

$$P(Y = y_j | X = x_i) = \frac{P(X=x_i|Y=y_j) \cdot P(Y=y_j)}{P(X=x_i)}, \quad (4)$$

where:

- $x_i$: feature vector of sample $i$, $i \in \{1, 2, \dots, n\}$
- $y_j$: notation of class $j$, $j \in \{1, 2, \dots, n\}$
- $P(Y = y_j | X = x_i)$: the probability of sample $x_i$ given a variable that belongs to class $y_j$.

### G. Bidirectional Encoder Representations from Transformers (BERT)

BERT is a pretrained model for English that has been trained on specialized datasets. The BERT model

has been trained on BookCorpus and English Wikipedia, which contains 11,038 unpublished books. Therefore, the BERT model benefits from its pretraining on a large-scale corpus, enabling it to extract richer linguistic patterns and deeper contextual representations compared to traditional models [14].

BERT's architecture builds on the Transformer model, employing stacked encoder layers to process language. Each encoder integrates a multi-head self-attention mechanism that analyzes all tokens in a sequence bidirectionally, capturing nuanced contextual relationships. This is followed by a feed-forward neural network, which applies non-linear transformations to further refine each token's representation. Depending on the variant, BERT uses 12 (Base) or 24 (Large) such layers, enabling it to generate deep, context-aware embeddings. This design makes BERT exceptionally effective for diverse NLP tasks, including sentiment analysis, question answering, and named entity recognition [14].

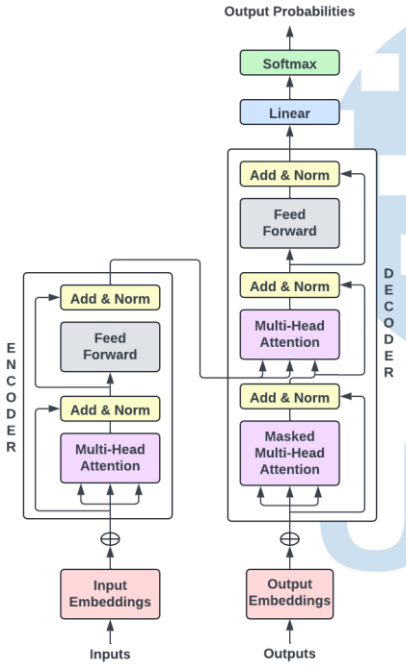



Fig. 1. BERT Architecture

### H. Indo-BERT

The Indo-BERT model is distinguished from the standard BERT model in that it is also a pretrained model on the Indonesian language corpus. This signifies that the Indo-BERT model has been trained on a specialized Indonesian dataset, encompassing diverse sources such as online news, social media, Wikipedia, online articles, and recorded video subtitles. Evidently, the Indo-BERT model is replete with Indonesian-specific information and exhibits remarkable capacity to effectively learn from other data sources [15].

### I. Hyperparameter Tuning

Hyperparameter tuning is defined as the process of identifying the optimal combination of parameters for a machine learning model. The objective of this process is to ascertain the most effective hyperparameter combination to enhance performance and mitigate the risk of overfitting and underfitting [16]. In this study, for naive bayes model, $\alpha$ as smoothing will be optimized using grid search to ensure that the class-conditional probability value does not equal zero, as this could result in the posterior probability value also being zero. For BERT, the following parameters will be optimized using the grid search method:

- **Learning Rate (LR):** 1×10-5, 2×10-5, 3×10-5, 4×10-5, 5×10-5 because the BERT model requires a low LR. The BERT method is a pre-trained model, and if a low LR value is used, the pre-trained information may be lost and the model may become unstable.
- **Epochs**: 3, 4, 5, 6, 7, because the BERT model is a model that is already rich in information. If you use an epoch value that is too high, it will overfit.
- **Batch Size (BS):** 8, 16, 32 because BERT has high memory requirements. Using multiple BS values helped achieve stable accuracy while maintaining reasonable memory usage and computation time.

### J. Confusion Matrix

The Confusion Matrix is a method for calculating the accuracy of a classification model [17]. It is presented as a table showing the number of correct and incorrect classifications for the test data. Then, the accuracy and F1 score can be calculated from the confusion matrix. Accuracy represents the percentage of correctly classified tuples in the test data [18]. It is calculated with the formula:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}. \quad (5)$$

The F1 score is the harmonic mean of precision and recall, providing a balance between the two metrics. Recall is the rate at which positive tuples are correctly identified, and precision is the percentage of tuples labeled positive that are actually positive [18]. It is calculated with the formula:

$$F1score = \frac{2 \cdot precision \cdot recall}{precision+recall}, \quad (6)$$

$$precision = \frac{TP}{TP+FP}, \quad (7)$$

$$recall = \frac{TP}{TP+FN}, \quad (8)$$

where:

- TP: True Positive
- TN: True Negative
- FP: False Positive
- FN: False Negative

## III. METHOD

This study involves several methodologies and processes, as outlined below:

### 1) Data Collection

The data used is a set of text documents containing comments made by students at X University about various facilities and services at X University. Facilities and services such as Career Center (CC), Registrar Office (RO), Finance (FIN), Library (LIB), Sports (OR), General Affair (GA), Student Life (SL), Information Technology Service Desk (ITSD), Wifi, Mobile App (APP), and Learning Experience (STUDY). In the data collection of comment text, sentiment information from the comment is available, sentiment can be 1 (positive), 0 (negative). The data collected were 27,659 comments from the X University Student Satisfaction Survey in 2022 to 2024.

TABLE I. SENTIMENT DATA

| Sentiment | Comment Count |
|---|---|
| Positive | 22475 |
| Negative | 5184 |

TABLE II. CATEGORY DATA

| Category | Comment Count |
|---|---|
| GA | 4084 |
| STUDY | 4285 |
| LIB | 3541 |
| SL | 3551 |
| APP | 3267 |
| RO | 2816 |
| OR | 1637 |
| WIFI | 1259 |
| FIN | 1386 |
| ITSD | 1332 |
| CC | 501 |

TABLE III. LANGUAGE DATA

| Language | Comment Count |
|---|---|
| Indonesia | 27406 |
| English | 253 |

### 2) Data Preprocessing

At this stage of the process, which is referred to as "text preprocessing," a series of critical steps must be taken. Initially, a data cleansing procedure is executed to eliminate non-alphanumeric characters, punctuation marks, and numerals. This is done to avert any potential disruptions in the ensuing classification results. Secondly, case folding involves the conversion of all words and sentences to lowercase, thereby ensuring a uniform text format and eliminating any capitalized words or sentences. The text is then segmented into smaller parts, or tokens, through a process known as tokenization. Finally, in the step of stopword removal, words that are frequently used and have minimal impact on the sentence's meaning, such as "and," "or," and "which," are eliminated. The removal of English stop words will be executed through the utilization of the *NLTK* library, while the removal of Indonesian stop words will be accomplished via the employment of the *re* module.

### 3) Data Undersampling

The undersampling process is implemented exclusively for Indonesian data due to the significant imbalance in the amount of data for each class category following pre-processing. Undersampling is a necessary procedure to ensure data balance, thereby enhancing the efficacy and performance of the model. In the absence of undersampling, the model's predictions are likely to be influenced by the majority class, as the majority class typically has a higher volume of data. In scenarios involving multiple categories for analysis, it is imperative to ensure that the proportion of each category is balanced. This approach facilitates more effective and equitable learning by the model.

### 4) Feature Extraction

Feature extraction constitutes a pivotal step following data pre-processing. This process entails the conversion of text or token data into numerical representations, thereby facilitating machine learning processes. Given the limitation of machines and models in comprehending text directly, but rather, their capacity to interpret numerical values, feature extraction emerges as a crucial step. This enables the subsequent interpretation of data by the machine or model, facilitating more profound prediction and analysis capabilities. The methodologies employed encompass n-grams for the Naive Bayes model and word embeddings for the BERT model.

### 5) Data Splitting

The process of data partitioning is executed in a random manner, involving the allocation of 70% of the data for training, 17.5% for validation, and 12.5% for testing. The data partitioning ratio of 70:17.5:12.5 is employed due to the substantial memory requirements and extended execution time of the BERT model. Subsequent to the undersampling process, the data is segmented into eight non-overlapping datasets, with one dataset allocated for prediction and the remaining datasets utilized for training and validation. It is anticipated that the model will demonstrate the capacity to predict with precision during the testing

phase, contingent on the successful identification of optimal parameters during the training phase.

6) Model Implementation

In this stage, a model is created using data that has undergone the text preprocessing steps. The output from these preprocessing steps is then processed using the Naive Bayes and BERT algorithm. The training of both algorithms is trained to produce the best possible hyperparameter combination for the task at hand. The training process for each model is to be executed independently, given that each model possesses unique steps and characteristics.

7) Keyword Extraction

Keyword extraction will be conducted subsequent to the training and testing process. This procedure will adhere to the same protocol as feature importance, wherein features/words exhibiting the most significant influence on specific class categories will be identified. In the context of NB modeling, keyword extraction can be achieved by calculating the log probability for each feature/word given a class. This will then be incorporated into a vector containing log probabilities for all features. BERT modeling is equipped with an inherent attention mechanism, wherein the attention score is determined during the model training process.

The process of keyword extraction will prioritize the identification of the aspect, disregarding other linguistic elements such as adjectives, verbs, and other non-essential components. To facilitate the sorting of words, an external library will be employed. The library utilized for this purpose is *Spacy* [19] for English and *Stanza* [20] for Indonesian.

8) Evaluation Testing

The trained NB and BERT models have the capacity to utilize the optimal parameters to predict other data. The most effective parameters obtained during the training process are stored in variables. Following the preparation of other data or testing data, the model can be directly applied to predict using the parameters that have been obtained in the training process. Following the execution of the prediction process and the subsequent acquisition of the prediction results, the performance of the two trained models will undergo evaluation. The evaluation of both models will be conducted employing the F1-score metric. The utilization of the F1-score metric is predicated on its capacity to facilitate a fair evaluation, even in scenarios where data is imbalanced.

## IV. Results and Discussion

### A. Data

Data was collected from 2022 to 2024 based on two languages, Indonesian and English. The data used consisted of answers to open-ended questions in the survey. The answers do not represent a direct assessment or evaluation of the facility/service; rather, they reflect personal opinions expressed in free text format. The data collection process also encompassed 11 distinct keyword categories, namely General Affair (GA), Sports (OR), Registrar Office (RO), Library (LIB), Career Center (CC), Student Life (SL), Learning Experience (Study), WiFi, Mobile App (APP), IT Service Desk (ITSD), and Finance (FIN).

TABLE IV. Data Example

| Bahasa | Kategori | Komentar | Sentimen |
|---|---|---|---|
| id | GA | kampus sangat bersih dan tertata dengan rapih. | 1 |
| id | GA | Kurang sentuhan hijau di kampus semanggi | 0 |
| id | OR | Lengkap dan mudah untuk diakses | 1 |
| en | LIB | more hand sanitizers posted please | 1 |
| id | WIFI | Ditingkatkan lagi kualitas Wifi nya | 0 |

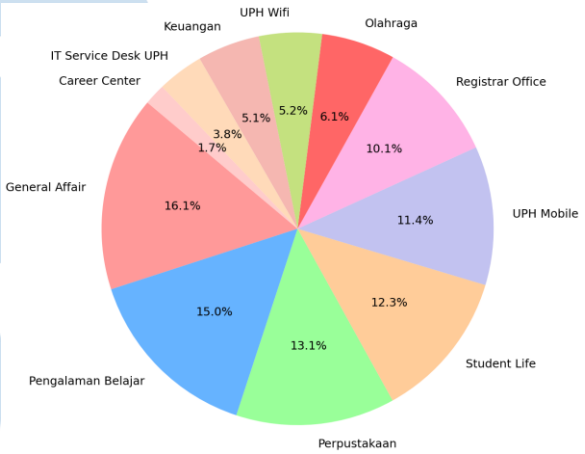



Fig. 2. Category Distribution in the Dataset

### B. Text Preprocessing

The data cleaning stage entails the removal of empty sentiments resulting from errors, as illustrated in Table 4. This results in enhanced organization and richer information content in the comment data compared to previous iterations. During the data cleaning process, irrelevant words are systematically removed, which can lead to the removal of comments due to the presence of empty sentiments. These comments are subsequently removed from the data prior to further processing. It is noteworthy that the data cleaning process is meticulously tailored for both the Indonesian and English data sets. Subsequent to the cleansing process, the individual data sets are then seamlessly integrated.

TABLE V. DATA EXAMPLE

| Before Preprocessing | After Preprocessing |
|---|---|
| Layanan dan fasilitas sudah sangat baik | layanan fasilitas sangat baik |
| The connection sometimes is bad. | connection sometimes bad |

### C. Undersampling Data

Prior to the integration of the NB and BERT models, an undersampling technique is employed to address the class imbalance present in Indonesian data. Class imbalance arises when the dataset exhibits a significant disparity in the proportion of data between classes, with a disproportionate number of instances belonging to class 1 or class 2, as depicted in Figure 4.2. The undersampling method involves the random selection of data points from the majority and minority classes, thereby ensuring a balanced distribution of data. The effectiveness of this method is evident in Tables 4.6 and 4.7, which illustrate the impact of undersampling on the Indonesian data. However, for the English data, the undersampling process is not employed due to its already substantial and balanced nature.

TABLE VI. DATA COUNT

| Indonesian Data Comment Count | | |
|---|---|---|
| **Class** | **Before Undersampling** | **After Undersampling** |
| Positive | 15583 | 2001 |
| Negative | 4217 | 1739 |
| **English Data Comment Count** | | |
| Positive | 147 | |
| Negative | 90 | |

TABLE VII. CATEGORY PROPORTION AFTER UNDERSAMPLING

| Indonesian Data Comment Count | | | | |
|---|---|---|---|---|
| | **Before** | | **After** | |
| **Category** | **Negative** | **Positive** | **Negative** | **Positive** |
| **GA** | 563 | 2634 | 170 | 170 |
| **OR** | 264 | 935 | 170 | 170 |
| **RO** | 329 | 1676 | 170 | 170 |
| **LIB** | 366 | 2225 | 170 | 170 |
| **CC** | 60 | 284 | 60 | 280 |
| **SL** | 268 | 2173 | 170 | 170 |
| **STUDY** | 593 | 2362 | 170 | 170 |
| **WIFI** | 719 | 301 | 170 | 170 |
| **APP** | 667 | 1581 | 170 | 170 |
| **ITSD** | 149 | 606 | 149 | 191 |
| **FIN** | 237 | 782 | 170 | 170 |

### D. Feature Extraction

The subsequent stage of the process involves the extraction or transformation of features from words into numbers. This stage follows the undersampling process. The feature extraction of the NB model will be evaluated through two distinct methods: TF-IDF and BoW, as illustrated in Tables 8 and 9. Subsequently, the feature extraction of the BERT model will be executed using word embeddings.

TABLE VIII. TF-IDF RESULT

| bersih | sangat | … | baca | … | kampus | sentimen |
|---|---|---|---|---|---|---|
| 0.3807 | 0.2145 | … | 0 | … | 0.3702 | 1 |
| 0 | 0.2162 | … | 0 | … | 0 | 1 |
| 0 | 0 | … | 0 | … | 0 | 1 |
| … | … | … | … | … | … | … |
| 0 | 0 | … | 0.3308 | … | 0 | 0 |

TABLE IX. BOW RESULT

| bersih | sangat | … | baca | … | kampus | sentimen |
|---|---|---|---|---|---|---|
| 1 | 1 | … | 0 | … | 1 | 1 |
| 0 | 1 | … | 0 | … | 0 | 1 |
| 0 | 0 | … | 0 | … | 0 | 1 |
| … | … | … | … | … | … | … |
| 0 | 0 | … | 1 | … | 0 | 0 |

### E. Model Training

In the training phase, the GSCV process is executed on both Indonesian and English data, and the optimal hyperparameter combinations obtained can vary between the two datasets. A total of 18 hyperparameter experiments have been selected for the GSCV test of the NB model, including 1e-6, 1e-5, 1e-4, 0.001, 0.01, 0.0 5, 0.1, 0.5, 1, 2, 5, 10, 20, 50, 100, 200, 500, and 1000. The determination of the optimal hyper-parameters for both datasets is achieved by selecting the highest F1-score, as illustrated in Table 10.

TABLE X. TOP 5 PERFORMING HYPERPARAMETER NB

| F1-score | | | | | | | |
|---|---|---|---|---|---|---|---|
| **TF-IDF 1-gram** | | | | **BoW 1-gram** | | | |
| $\alpha$ | ID | $\alpha$ | EN | $\alpha$ | ID | $\alpha$ | EN |
| 5 | 0.6921 | 2 | 0.7639 | 5 | 0.6921 | 2 | 0.7639 |
| 10 | 0.6902 | 5 | 0.7639 | 10 | 0.6902 | 5 | 0.7639 |
| 2 | 0.6896 | 500 | 0.7639 | 2 | 0.6896 | 500 | 0.7639 |
| 0.5 | 0.6887 | 200 | 0.7639 | 0.5 | 0.6887 | 200 | 0.7639 |
| 0.1 | 0.6870 | 1 | 0.7639 | 0.1 | 0.6870 | 1 | 0.7639 |

A total of 125 hyperparameter combination experiments have been selected for the BERT model GSCV test. The BERT model was tested using three different hyperparameters: LR, epochs, and BS. The determination of the optimal hyperparameters for both datasets is achieved the same way like NB, by selecting the highest F1-score, as illustrated in Table 11.

TABLE XI. TOP PERFORMING HYPERPARAMETER BERT

| Data | LR | Epoch | BS | F1 Train | F1 Test |
|---|---|---|---|---|---|
| DataID 1 | $2 \cdot 10^{-5}$ | 4 | 32 | 0.8892 | 0.7856 |
| DataID 2 | $3 \cdot 10^{-5}$ | 5 | 32 | 0.8753 | 0.8278 |
| DataID 3 | $3 \cdot 10^{-5}$ | 5 | 8 | 0.8466 | 0.8361 |
| DataID 4 | $2 \cdot 10^{-5}$ | 5 | 16 | 0.9084 | 0.8047 |
| **DataID 5** | $\mathbf{1 \cdot 10^{-5}}$ | **5** | **16** | **0.9146** | **0.8491** |
| DataID 6 | $5 \cdot 10^{-5}$ | 5 | 32 | 0.8214 | 0.8064 |
| DataID 7 | $3 \cdot 10^{-5}$ | 6 | 8 | 0.8903 | 0.8076 |
| **DataEN** | $\mathbf{1 \cdot 10^{-5}}$ | **3** | **8** | **0.8611** | **0.8308** |
| DataEN | $4 \cdot 10^{-5}$ | 3 | 32 | 0.7938 | 0.8136 |
| DataEN | $5 \cdot 10^{-5}$ | 3 | 16 | 0.8780 | 0.8286 |

### F. *Keyword Extraction*

Keyword extraction is the process of identifying words or tokens that exert the greatest influence on sentiment prediction in a given method. This is conducted subsequent to sentiment prediction.Each method employs distinct techniques to identify keywords that impact sentiment prediction.In this instance, the NB method utilizes log probability, while the BERT method employs a feature from its own model, namely attention and attention score. The objective of keyword extraction is to identify efficiently and quickly which facilities/services have been rated as satisfied and dissatisfied in each category.

The NB method of keyword extraction involves the calculation of the log probability of each word or token. The log probability value obtained for a word or token indicates its importance in sentiment prediction. That is, the higher the log probability value, the more significant the word or token is to sentiment prediction. Conversely, the lower the log probability value, the less relevant the word or token is to sentiment prediction. The results and visualization of the NB method of keyword extraction of GA category can be observed in Figure 2 and Figure 3. Other categories and English versions of the keyword extraction are available at https://github.com/j0daaaa/TA_SentimentAnalysis_NLP.

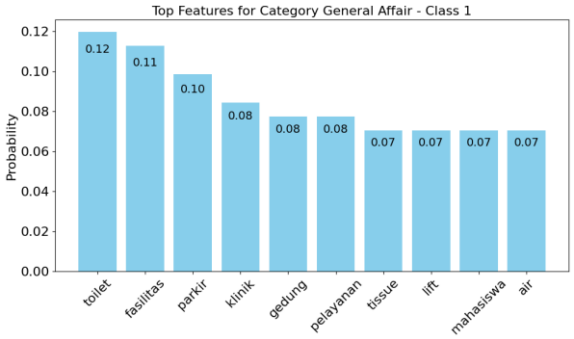



Fig. 3. GA Category Positive Keywords NB 1-gram

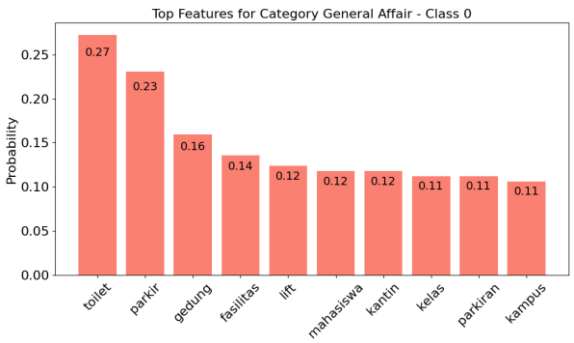



Fig. 4. GA Category Negative Keywords NB 1-gram

The BERT method for keyword extraction entails the extraction of the attention score feature for each word or token in the comment, utilizing the BERT model. Subsequently, the value of the word or token is extracted in its entirety, and the attention score value for a word or token is totaled. The aggregate attention score of a word or token toward positive or negative sentiment is then obtained. This calculation is analogous to the calculation of log probability in the NB method, in that the greater the aggregate attention score value, the more important the word/token is to sentiment prediction. The visualization results of the BERT method keyword extraction of the GA category are presented in Figure 4 and Figure 5.

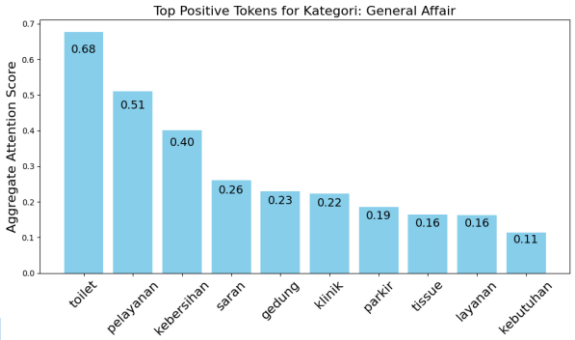



Fig. 5. GA Category Positive Keywords BERT

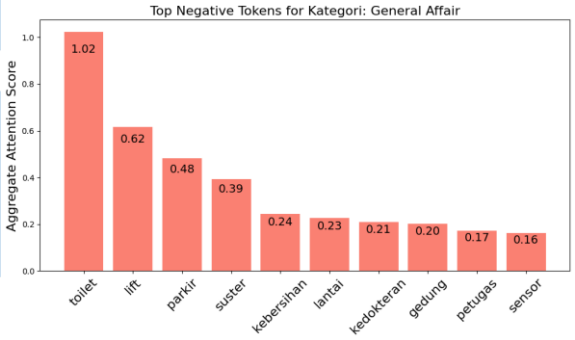



Fig. 6. GA Category Positive Keywords NB BERT

### G. *Model Evaluation*

A total of 508 testing data sets were utilized for sentiment prediction, comprising 53 instances of GA data, 38 instances of OR data, 45 instances of RO data, 44 instances of LIB data, 54 instances of CC data, 46 instances of SL data, 52 instances of STUDY data, 49 instances of WIFI data, 31 instances of APP data, and 57 instances of ITSD data, along with 39 instances of FIN data. Predictions have been made using the TF-IDF method with n = 1, 2, 3, the BoW method with n = 1, 2, 3, and the BERT method. A comprehensive summary of the results obtained from all methods employed is provided in Table 12.

TABLE XII. MODEL SUMMARY

| Model | CM | | F1-score |
|---|---|---|---|
| TF-IDF 1-gram | 27 | 95 | 0.689977 |
| | 39 | 347 | |
| TF-IDF 2-gram | 34 | 88 | 0.670839 |
| | 59 | 327 | |

| Model | CM | | F1-score |
|---|---|---|---|
| TF-IDF 3-gram | 12 | 110 | 0.665953 |
| | 11 | 375 | |
| BoW 1-gram | 27 | 95 | 0.689977 |
| | 39 | 347 | |
| BoW 2-gram | 34 | 88 | 0.670839 |
| | 59 | 327 | |
| BoW 3-gram | 12 | 110 | 0.665953 |
| | 11 | 375 | |
| **BERT** | **83** | **39** | **0.776978** |
| | **116** | **270** | |

As shown in Table 12, BERT's superior performance over NB remains consistent across all tested values of n, reinforcing its robustness in sentiment analysis tasks. This superiority can be attributed to the BERT model's capacity to effectively handle complex language patterns, a capability that is inherently limited in the NB model due to its assumption of feature independence. The NB model demonstrates greater result instability as n-gram levels increase, leading to diminishing reliability in conclusions. In contrast to the NB model, the BERT model demonstrates a notable enhancement in accuracy. This enhancement can be attributed to its ability to learn complex patterns, understand word context bidirectionally, and effectively handle language elements such as negation or sarcasm.

Furthermore, BERT's pre-training on a substantial and varied text corpus enhances its adaptability and efficacy in sentiment analysis, a capability that is lacking in NB. Deep learning models such as BERT have more accurate predictive performance than machine learning models such as NB. This statement is supported by prior research conducted by Braig et al.[21], where it was found that deep learning models such as BERT or RoBERTa achieve higher predictive accuracy compared to machine learning models such as logistic regression, multinomial naive bayes, and others.

The comments in the suggestion column constitute responses to open-ended inquiries. This constitutes a factor that influences the model's comprehension of the context to be acquired. In the development of the BERT model, there was a decline in the F1-score accuracy of approximately 0.07. This decline is presumably attributable to the characteristic nature of comments, which manifest as open-ended responses.

## V. CONCLUSION

The performance effectiveness of the Naive Bayes (NB) and BERT models in sentiment analysis of student satisfaction surveys with mixed and nonstandard languages demonstrates that BERT is superior in capturing sentiment. Following the training of both models and the identification of optimal parameters, BERT attained a prediction accuracy of 0.776978, marginally exceeding the accuracy of 0.689977 achieved by NB with 1-gram, 0.670839 with 2-gram, and 0.665953 with 3-gram. The NB method utilizes the n-gram approach ($n = 1, 2, 3$) with TF-IDF and BoW representations to capture patterns in the data. The primary advantage of BERT lies in its capacity to understand complex language contexts, thereby making it a more reliable choice for sentiment analysis.

The keywords derived from sentiment analysis of student satisfaction surveys, encompassing both positive and negative sentiments, offer a comprehensive representation of students' perceptions regarding various facilities or services. However, the BERT method has been found to outperform the NB method in terms of keyword accuracy. This is primarily due to the presence of equal positive and negative keywords in the NB method, which hinders the ability to draw definitive conclusions. In the context of positive sentiments, keywords such as ”kebersihan”, “layanan”, “court”, “fast respon”, and “pelayanan sangat baik”, reflecting student satisfaction with the facility or service. Conversely, in the case of negative sentiments, keywords such as "toilet", "sinyal", "errors", "kelas karyawan", and "mohon teliti menginput" indicate student dissatisfaction with certain facilities or services. The results of these keywords can be used as evaluation material for the university to identify facilities or services that need to be maintained or improved to increase overall student satisfaction.

Despite BERT’s superior performance, its practical adoption faces challenges. The model’s reliance on large annotated datasets for fine-tuning may limit scalability in resource-constrained scenarios, and its pretraining biases could affect generalizability across domains (e.g., informal text or low-resource languages). Running BERT demands expensive hardware, limiting its use in real-world systems. These constraints suggest that simpler models like Naive Bayes remain viable for tasks where interpretability or efficiency outweighs marginal gains in accuracy.

## REFERENCES


[1] B. B. Agubosim, M. M. Arshad, S. N. Alias, and A. Mousavi, “Job satisfaction and job performance among university staff in Nigeria,” International Journal of Academic Research in Progressive Education and Development, 12 (2): 2620-2631. DOI: 10.6007/IJARPED/viz-i2, vol. 17669, 2023.

[2] R. N. Levitz, “National student satisfaction and priorities report,” Cedar Rapids, Iowa: Ruffalo Noel Levitz. Retrieve from RuffaloNL. com/Benchmark, 2017.

[3] B. L. McCombs, “The learner-centered model: From the vision to the future,” Interdisciplinary applications of the person-centered approach, pp. 83–113, 2013.

[4] M. Y.-P. Peng and C. C. Chen, “The effect of instructor’s learning modes on deep approach to student learning and learning outcomes,” Educational sciences: theory & practice, vol. 19, no. 3, 2019.

[5] P. Chaovalit and L. Zhou, “Movie review mining: A comparison between supervised and unsupervised classification approaches,” in Proceedings of the 38th annual Hawaii international conference on system sciences, 2005, pp. 112c–112c.

[6] K. Dave, S. Lawrence, and D. M. Pennock, “Mining the peanut gallery: Opinion extraction and semantic classification of product reviews,” in Proceedings of the 12th international conference on World Wide Web, 2003, pp. 519–528.

[7] Y. Al Amrani, M. Lazaar, and K. E. El Kadiri, “Random forest and support vector machine based hybrid approach to sentiment analysis,” Procedia Comput Sci, vol. 127, pp. 511–520, 2018.

[8] M. Syarifuddinn, “Analisis sentimen opini publik terhadap efek PSBB pada twitter dengan algoritma decision tree, knn, dan na\"\ive bayes,” INTI Nusa Mandiri, vol. 15, no. 1, pp. 87–94, 2020.

[9] D. Darwis, N. Siskawati, and Z. Abidin, “Penerapan Algoritma Naive Bayes Untuk Analisis Sentimen Review Data Twitter Bmkg Nasional,” Jurnal Tekno Kompak, vol. 15, no. 1, pp. 131–145, 2021.

[10] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. Cambridge University Press, 2008. [Online]. Available: https://nlp.stanford.edu/IR-book/

[11] K. Juluru, H.-H. Shih, K. N. Keshava Murthy, and P. Elnajjar, “Bag-of-words technique in natural language processing: a primer for radiologists,” RadioGraphics, vol. 41, no. 5, pp. 1420–1426, 2021.

[12] W. B. Cavnar, J. M. Trenkle, and others, “N-gram-based text categorization,” in Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval, 1994, p. 14.

[13] S. Raschka, “Naive bayes and text classification i-introduction and theory,” arXiv preprint arXiv:1410.5329, 2014.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” CoRR, vol. abs/1810.04805, 2018, [Online]. Available: http://arxiv.org/abs/1810.04805

[15] B. Wilie et al., “IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding,” in Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, 2020.

[16] R. Ahuja, K. Vats, C. Pahuja, T. Ahuja, and C. Gupta, “Pragmatic Analysis of Classification Techniques based on Hyper-parameter Tuning for Sentiment Analysis,” 2020.

[17] A. M. Zuhdi, E. Utami, and S. Raharjo, “Analisis sentiment twitter terhadap capres Indonesia 2019 dengan metode K-NN,” Jurnal Informa: Jurnal Penelitian dan Pengabdian Masyarakat, vol. 5, no. 2, pp. 1–7, 2019.

[18] M. Y. Aldean, P. Paradise, and N. A. S. Nugraha, “Analisis Sentimen Masyarakat Terhadap Vaksinasi Covid-19 di Twitter Menggunakan Metode Random Forest Classifier (Studi Kasus: Vaksin Sinovac),” Journal of Informatics Information System Software Engineering and Applications (INISTA), vol. 4, no. 2, pp. 64–72, 2022.

[19] Explosion, “spaCy: Industrial-strength Natural Language Processing in Python,” 2020. [Online]. Available: https://spacy.io

[20] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, “Stanza: A Python Natural Language Processing Toolkit for Many Human Languages,” arXiv preprint arXiv:2003.07082, 2020.

[21] N. Braig, A. Benz, S. Voth, J. Breitenbach, and R. Buettner, “Machine learning techniques for sentiment analysis of COVID-19-related twitter data,” *IEEE Access*, vol. 11, pp. 14778–14803, 2023

ISSN 2085-4552

# Trends and Keyword Networks in Machine Learning-Based Click Fraud Detection Research

Kevin[1], Aditiya Hermawan[2]
[1,2] Dept. of Informatics Engineering, Buddhi Dharma University, Tangerang, Indonesia
[1]kevinawi1112@gmail.com, [2]aditiya.hermawan@ubd.ac.id

Accepted 25 June 2025
Approved 30 June 2025

***Abstract*— The rapid advancement of the digital economy has significantly increased the use of online advertising while concurrently giving rise to critical challenges, particularly in the form of click fraud—a manipulative act that harms advertisers by generating fraudulent clicks on digital advertisements. As click fraud attack patterns grow increasingly complex, machine learning (ML)-based research has emerged as a principal approach for detecting and mitigating these threats. This study aims to map the research landscape of ML-based click fraud detection through a bibliometric analysis to identify publication trends, patterns of international and institutional collaboration, and key thematic domains within this field. Employing a bibliometric methodology, the study analyzed 61 publications retrieved from Dimensions.ai spanning the years 2015–2024. The data were collected, refined using OpenRefine, and visualized with VOSviewer to examine keyword co-occurrences and research trends. The findings reveal a marked increase in publication volume since 2019, with dominant contributions from India, China, Saudi Arabia, and the United States. Furthermore, four principal research clusters were identified: cybersecurity, the relationship between click fraud and the digital advertising industry, dataset processing and evaluation techniques, and the development of ML-based detection systems. Each cluster offers practical contributions in areas such as system protection strategies, ad budget optimization, improved detection accuracy, and the development of scalable, real-time detection solutions. Recent trends highlight growing scholarly interest in model performance evaluation and the challenges posed by class imbalance (class skewness). This study concludes that more effective data management and the development of adaptive ML models capable of addressing evolving attack patterns are pivotal for future research. By providing a clearer mapping of current trends, this study aims to support the scientific community in developing more accurate and efficient click fraud detection strategies, thereby strengthening the integrity of the global digital advertising ecosystem.**

***Index Terms*— Click Fraud; Machine Learning; Bibliometric Analysis; Fraud Detection; Digital Advertising**.

## I. INTRODUCTION

Over the past few decades, the digital economy has grown rapidly worldwide [1]. This sector has also become a key transmission hub in the economic system, contributing significantly to global economic growth [2]. One of the main drivers of this growth is technological advancement [3]. The development of technology has enabled companies to reach consumers more effectively through digital platforms.

However, alongside this rapid growth, new challenges related to digital security have emerged, particularly in the form of Click Fraud, a manipulative act that generates fraudulent clicks on advertisements with the intent of harming advertisers [4]. A 2020 study by the University of Baltimore found that click fraud caused losses exceeding $35 billion [5]. Click fraud not only results in substantial financial losses for advertisers but also undermines the integrity of the digital advertising ecosystem as a whole. To address this threat, technology-based solutions are required. Detecting click fraud generally relies on machine learning models, which have become one of the most effective approaches due to their ability to learn complex behavioral patterns and identify subtle anomalies that signal fraudulent activity [6].

The application of ML techniques in detecting click fraud has received significant attention in recent years. ML provides the capability to analyze complex data patterns and identify anomalies that traditional methods might overlook. A study by Aljabri investigated the application of machine learning models to distinguish between human and bot click behaviors in pay-per-click (PPC) advertising. The results showed that while all models achieved strong performance, the Random Forest algorithm consistently outperformed others across all evaluation metrics, indicating its robustness in detecting fraudulent ad-click activity [7]. Additional research also highlighted that ensemble methods can further enhance detection performance [8] .

Despite the increasing number of studies applying machine learning for click fraud detection [9], there is a lack of systematic synthesis regarding how research in this field has evolved, what methods are predominantly used, and which conceptual domains remain underexplored. Previous reviews have tended to focus on algorithmic performance or case-specific implementations, rather than providing a macro-level mapping of the intellectual structure of the field. In contrast, bibliometric studies in other fraud detection domains, such as financial fraud or healthcare fraud, have provided broader overviews of research trends. These studies often focus on general approaches or dominant techniques, such as decision trees, SVMs, or neural networks, but fail to provide a detailed mapping of publication trends or global research collaboration patterns. Therefore, this study fills an important gap by conducting a bibliometric analysis to uncover research trends, influential themes, and methodological patterns in the intersection of machine learning and click fraud detection. The findings are expected to inform both academic research agendas and practical implementations in digital advertising security.

This study aims to address the identified research gap by conducting a bibliometric analysis of scholarly literature focused on click fraud detection using machine learning (ML) techniques. Bibliometric analysis is a widely adopted approach with high methodological validity for examining large bodies of academic literature. This method enables researchers to trace the historical development of a scientific discipline and to identify emerging directions and novel themes within the field [10]. Unlike general fraud detection bibliometrics, which primarily examine techniques applied to broader domains like finance or healthcare, this study focuses on the specific context of click fraud, offering in-depth insights into research trends unique to the digital advertising ecosystem. By mapping the evolution of research in this domain, the study seeks to provide in-depth insights into the current and prospective trajectories of scholarship in click fraud detection. Furthermore, a better understanding of prevailing trends and research patterns may contribute to the development of more effective strategies for detecting and preventing click fraud, thereby reinforcing the integrity of the digital advertising ecosystem.

To achieve these objectives, this study seeks to answer the following research questions:

1. How have publication trends in Click Fraud detection using Machine Learning evolved over time, both in terms of the number of publications and collaboration patterns among countries and institutions?
2. What are the key research topics and keyword co-occurrence patterns in ML-based Click Fraud detection studies, as identified through bibliometric analysis?

How has the research focus on ML-based Click Fraud detection shifted over time, particularly in terms of keyword relationships and emerging topics in recent years?

## II. METHOD

The procedure used to conduct this research consists of five stages. These stages are as follows: Data Collection, Data Cleaning, Data Visualization, Data Analysis, and Report Writing. Fig 1 illustrates how this procedure should be carried out in more detail.

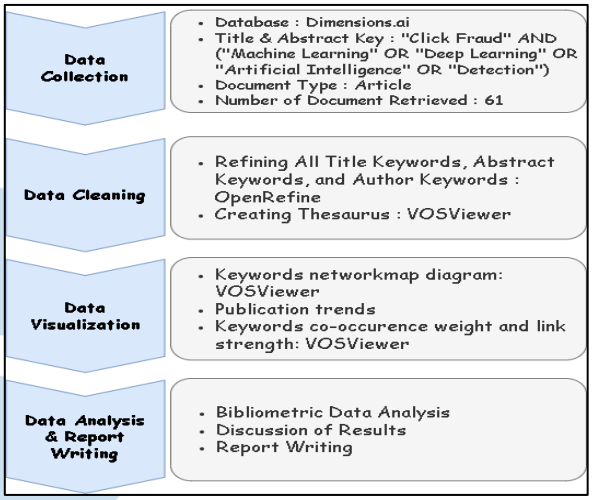


Fig 1. Research methodology

### A. Data Collection

The data was obtained from the Dimensions.ai website as part of the data collection phase in the form of a CSV file, the selection of Dimensions.ai as the primary data source was based on its open and freely accessible nature, which facilitates the independent execution of bibliometric analysis. The applied publication year restriction spans from 2015 to 2024, covering a 10-year range. The search query used was: TITLE-ABSTRACT ("click fraud" OR "ad fraud") AND ("machine learning" OR "deep learning" OR "artificial intelligence" OR "detection"). This search was restricted to journal articles only. Using this search technique, a total of 61 articles were retrieved. Although relatively limited in quantity, the 61 publications included in this study were manually screened and curated to ensure high thematic relevance to the specific domain of click fraud detection using machine learning. Broader search queries using general terms such as 'fraud' produced a considerable number of irrelevant results—covering areas like financial fraud, healthcare fraud, and identity theft—which would have diluted the semantic focus of the analysis. Therefore, this study deliberately prioritized semantic precision over corpus size, a methodological trade-off commonly accepted in bibliometric analyses of niche or emerging topics. Moreover, Glänzel and Moed [11] suggest a rule-of-thumb minimum of 50 documents to ensure approximate properties such as normality in the

distribution of means and relative frequencies. With 61 highly relevant articles, this study meets that threshold and maintains sufficient statistical integrity for meaningful co-word and thematic mapping.

### B. Data Cleaning

This data cleaning phase aimed to ensure more accurate exploration of bibliometric and bibliographic data, as well as to enable improved visualization and interpretation of the results [10]. All keywords used in the Title and Abstract fields were standardized using OpenRefine. OpenRefine facilitated the detection of semantically similar keywords by identifying lexical variations within the dataset, thereby supporting the standardization and consolidation of terms that are conceptually identical but expressed differently. This process had a significant impact on enhancing the accuracy and integrity of the keyword co-occurrence network structure, as the merging of redundant terms prevented the fragmentation of thematic clusters that could otherwise distort the conceptual mapping. Consequently, the resulting network visualizations more accurately reflect the dominant themes within the literature and strengthen the validity of interpretations regarding topical interconnections within the analyzed research corpus. Table I presents examples of keyword standardizations performed during the data cleaning process using OpenRefine.

TABLE I.KEYWORD STANDARDIZATION EXAMPLES

| Original Keyword | Standardized Keyword |
|---|---|
| Prediction, predicting, predict, predicted | prediction |
| Demonstrate, demonstrated, demonstrates | demonstrate |
| Classifier, classfiers, classifies, classfier's | classifier |
| Fraudster, fraudsters, fraudster's | fraudsters |

### C. Data Visualization

The Data Visualization Phase was carried out by constructing a network map based on keyword co-occurrence from the analyzed articles using VOSviewer. This phase aimed to identify relationships between keywords in the dataset and explore conceptual linkages within this research field.

At this stage, the minimum keyword co-occurrence threshold was set at 6, resulting in the selection of 79 keywords from a total of 1,697 available terms. The selection of a threshold of six was not arbitrary; rather, it aligns with established bibliometric practices and is theoretically grounded in the thresholding formula introduced by Donohue [12], as operationalized in subsequent studies such as [13], [14]. This method estimates the optimal boundary for distinguishing high-frequency keywords based on the distribution of singleton terms within the corpus. By applying this threshold, the present study adheres to a well-documented standard in co-word analysis, which ensures analytical consistency and avoids distortions caused by low-frequency noise.

To confirm its appropriateness, a limited sensitivity trial was conducted by comparing alternative thresholds. When the threshold was reduced to four, 97 keywords were retained—exceeding the recommended upper boundary of 67 high-frequency terms as per the Donohue model, and introducing considerable lexical noise. Conversely, raising the threshold to eight and ten produced only 57 and 6 keywords, respectively—both falling below the recommended inclusion range and omitting key conceptual terms. Consequently, additional sensitivity testing was deemed unnecessary, as the threshold has been validated and widely adopted in comparable bibliometric investigations.

Furthermore, out of the 79 identified keywords, only 60% (47 keywords) were used as the final threshold. In bibliometric analysis, the 60% threshold is the default setting in VOSviewer and is considered a best practice [15].

Additionally, Python with the Plotly library was used to generate bibliometric data visualizations, such as the total publication distribution by country, which helps identify the most productive countries in this research domain. Moreover, publication trends over the years were analyzed to examine research developments within a specific time frame.

### D. Data Analysis and Report Writing

The final phase of this study consists of data analysis and report writing. The bibliometric data presented in the data visualization phase is then evaluated and interpreted based on the articles included in this study. The interpretation of results is based on the bibliometric data visualizations generated in the previous phase, including the analysis of the network map diagram, which was constructed using the co-occurrence of article keywords. The findings, discussion, and conclusions of this research are then summarized in a comprehensive report, ensuring a clear understanding of the trends and conceptual linkages identified in the study.

## III. Result And Discussion

This section presents the results of the bibliometric analysis on Click Fraud Detection research. The analysis was conducted to identify publication trends over time, publication distribution by country, and the most frequently used machine learning methods for click fraud detection. These findings provide a comprehensive overview of the research developments in this field, including the number of publications, citation impact, and the dominant techniques in current scientific approaches.

### A. Publication Trends Over Time

The publication trend analysis indicates that research on Click Fraud Detection has experienced a significant increase after 2019. As shown in Fig. 2, the

highest number of publications was recorded in 2022, with 14 articles published that year. Specifically, the number of publications grew by 55.56% from 2020 (9 articles) to 2021 (14 articles), reflecting a sharp surge in interest in this area. The trend reveals a steady increase, with an annual growth rate of approximately 30% from 2019 to 2022, followed by a slight decline in 2023 and 2024. This growth indicates the expanding importance of click fraud detection in the context of the digital economy. Despite an average annual growth rate of +14.9%, the trend is heavily skewed by extreme outliers, indicating that the field's growth is non-linear and highly volatile. This pattern may reflect a combination of dataset limitations (e.g., incomplete indexing for 2024), external disruptions (such as funding shifts or academic redirection), and possible saturation in the core area of click fraud detection.

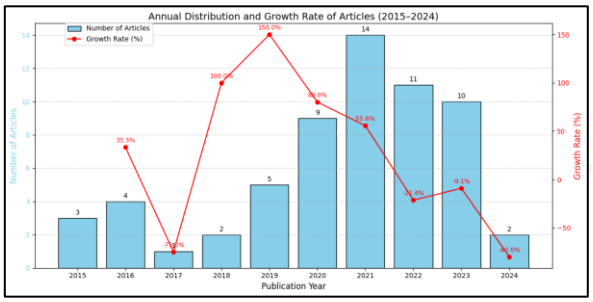



Fig 2. Annual distribution of articles on click fraud detection (2015–2024)

The period from 2018 to 2019 saw research in this area still in its early exploratory phase, with a relatively low number of publications. However, a sharp increase in publications occurred from 2020 to 2022, marked by significant growth in research output, a rise in citation impact, and stronger interconnections among studies in this domain.

### *B. Geographic Trends of Publications*

The analysis results indicate that publications on this topic are globally distributed, with certain countries contributing more significantly than others. Fig. 3 presents the top eight countries with the highest number of publications in click fraud detection research, along with the exact number of publications per country.

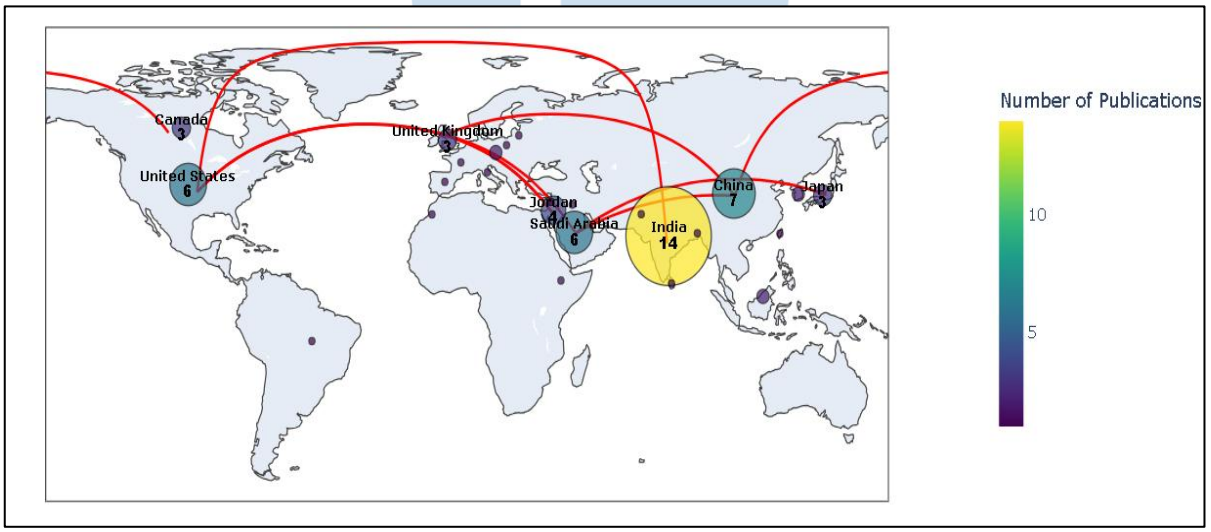



Fig 3. Top eight countries with the highest number of publications in click fraud detection research

Fig. 3. Top eight countries with the highest number of publications in click fraud detection research. The number of publications is indicated for each country: India (14), China (7), Saudi Arabia (6), the United States (6), and others including Jordan, and several European nations.

India emerges as the leading country in this research domain, contributing a total of 14 publications, which accounts for 23% of the total publications in this domain. India's dominance in this field can be attributed to its rapidly growing information technology industry. China, contributing 7 publications (approximately 11% of the total), and Saudi Arabia with 6 publications (about 10%), also demonstrate significant interest in click fraud detection, particularly in the context of protecting their digital advertising ecosystems. The United States, as a global hub for digital technology and advertising, has contributed 6 publications (around 10%), indicating continued academic and industrial engagement in this research area.

Other countries, such as Jordan and several European nations, have also made notable contributions to this field, though their contributions are smaller in scale. The overall distribution of publications highlights a growing global interest in ML-based click fraud

detection, with nations not only from large digital advertising markets but also those prioritizing cybersecurity and the efficiency of digital advertising systems.

### C. Network Map Diagram Analysis

The network map diagram based on keyword co-occurrence in the analyzed articles was generated using VOSviewer. The term "keyword co-occurrence" refers to the frequency with which a keyword appears across multiple publications. The minimum occurrence threshold can vary significantly depending on the research objectives. A lower threshold results in more keywords being displayed, while a higher threshold reduces the number of displayed keywords.

Researchers extracted 1,697 keywords from a total of 61 articles. The minimum threshold for keyword co-occurrence was set at 6 times, resulting in 79 keywords meeting the minimum requirement. Fig 4 illustrates that only 60% of the total connections among the 79 keywords—equivalent to 47 keywords—were included in the final visualization.

The weight of an item determines the size of its label and circle in the network map. The larger the weight of an item, the bigger its label and circle. The color of each item is determined by the cluster to which it belongs, reflecting thematic groupings within the dataset.

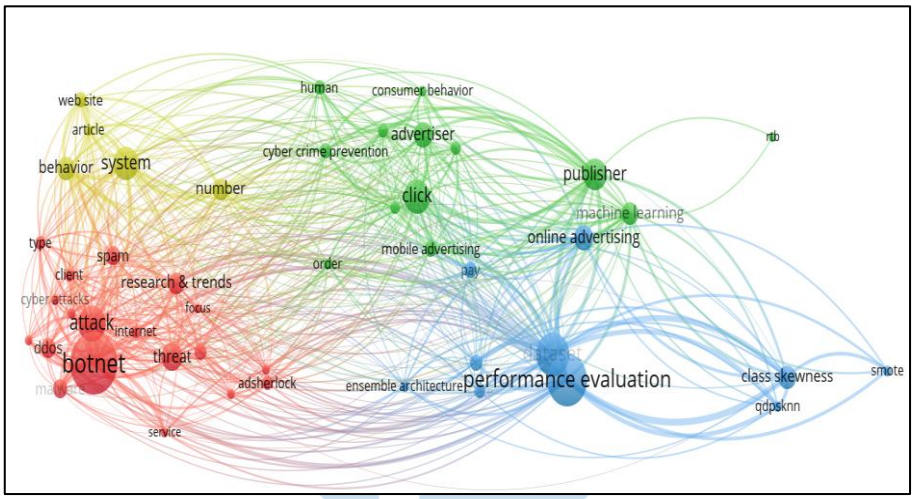



Fig 4. Network visualization of keyword co-occurrence in click fraud detection studies

Each of the four colors visible in Fig 4 represents a different cluster. The clustering approach is based on keyword co-occurrence, as detected across all analyzed articles. This indicates that elements grouped within the same cluster are more closely related to one another compared to elements outside their respective clusters. Therefore, it can be inferred that elements within the same cluster likely share a similar research focus. The details of the keywords in each cluster are summarized in Table II.

Keywords such as "botnet" and "performance evaluation" also exhibit high link strength, suggesting that detection models frequently correlate with botnet-based attack patterns and performance evaluation techniques. The presence of the term "dataset" with strong connections further indicates that data quality and dataset processing methods are key factors in the effectiveness of click fraud detection systems. Fig 5 presents the top 15 keywords with the highest co-occurrence values and total link strength.

While the co-word and cluster analyses have yielded useful thematic structures, it is important to acknowledge several limitations inherent in the bibliometric approach used. First, this study exclusively utilized the Dimensions.ai database, which although extensive in scope aggregates a wide variety of publication types and disciplines. This heterogeneity may influence the consistency and interpretive clarity of the resulting thematic patterns, particularly when compared with more curated and domain-specific bibliographic databases. To mitigate potential coverage bias and ensure methodological triangulation, future studies are encouraged to cross-validate findings using established repositories such as Scopus or Web of Science, which offer more standardized indexing criteria and peer-reviewed literature emphasis.

Second, the process of keyword normalization conducted using OpenRefine may introduce semantic ambiguity. Decisions on merging or standardizing keywords (e.g., “click fraud” vs “ad fraud”) rely partly on subjective judgment, which could influence the resulting cluster composition. Moreover, relying on author-assigned keywords may bias the analysis toward how authors frame their work, rather than capturing the conceptual content in full.

Third, while thresholding co-occurrence at six ensured analytical clarity, this choice may have excluded emerging but low-frequency terms that are thematically significant. This reflects a broader limitation of co-word analysis itself: its tendency to privilege frequency over novelty.

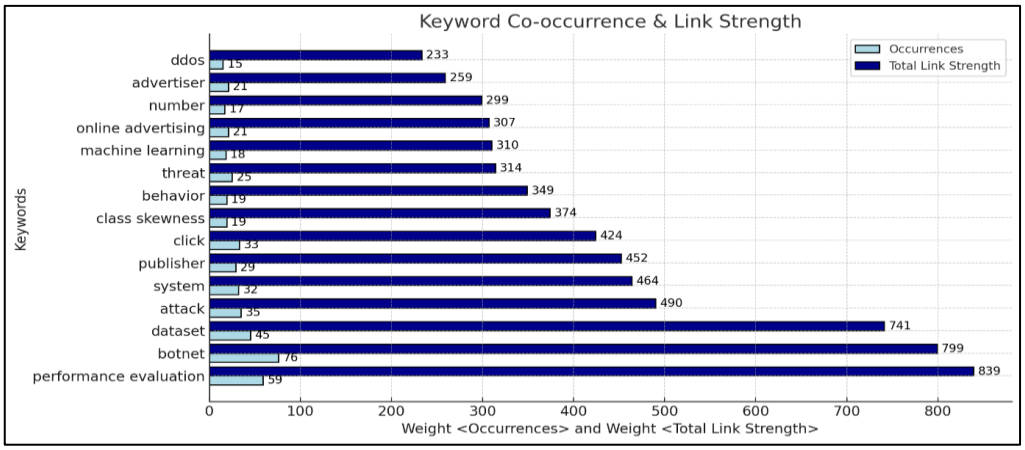



Fig 5. Top 15 keywords ranked by co-occurrence and total link strength in click fraud detection research

TABLE II. KEYWORD CLUSTERS

| Cluster | Keywords | Issue |
|---|---|---|
| 1 (19 Keywords) | adsherlock, attack, botnet, client, cyber attacks, day, ddos, focus, internet, malicious code, malware, mobile app development, online detection, phishing, research & trends, service, spam, threat, type | This cluster is highlighting the foundational concerns of network-based attack vectors and systemic vulnerabilities. This cluster aligns with real-world challenges in identifying sophisticated bot traffic and suggests the need for integrating behavioral analytics into fraud detection pipelines. Studies such as Sadeghpour and Vlajic have shown that botnets often mimic legitimate user behavior, making this cluster crucial for developing resilient detection mechanisms [16] |
| 2 (13 Keywords) | advertiser, click, consumer behavior, cyber crime prevention, fraudulent click, human, machine learning, mobile advertising, order, parameter, publisher, revenue, rtb | This cluster comprising keywords like represents the commercial and economic dimension of the field. The prominence of these terms underscores the growing concern from advertisers and platforms over financial losses. This cluster suggests a need for models that not only detect fraud but also estimate its economic impact [17] |
| 3 (10 Keywords) | class skewness, classification, dataset, ensemble architecture, online advertising, pay, performance evaluation, qdpsknn, smote, state | This cluster revolves around technical modeling issues, with keywords such as class imbalance, SMOTE, and ensemble learning. These terms underscore the methodological challenges in handling skewed datasets—an inherent characteristic of fraud detection tasks, where legitimate instances significantly outnumber fraudulent ones. The prominence of these terms highlights the growing emphasis on developing resilient models capable of maintaining predictive performance under such imbalance. Notably, G.S. T. et al. [18] demonstrated the effectiveness of ensemble-based methods in addressing class imbalance by leveraging the combined strengths of multiple classifiers, thereby supporting the broader adoption of hybrid learning architectures in this domain. This cluster, therefore, opens further avenues for research into meta-learning strategies and cost-sensitive algorithms tailored for rare-event prediction in click fraud detection |
| 4 (5 Keywords) | article, behavior, number, system, web site | Cluster 4 includes terms such as model architecture, detection system, and anomaly detection, reflecting a focus on the system-level implementation of click fraud detection frameworks. This cluster serves as a bridge between theoretical algorithm development and practical engineering deployment, emphasizing the importance of scalable and explainable AI models capable of operating in real-time environments. Notably, a study by Neeraja et al. supports this direction by demonstrating that real-time ad-click fraud can be effectively identified using elementary classifiers [19] |

### *D. Overlay Visualization*

An overlay visualization was created to identify the latest research topics, as illustrated in Fig 6. The color gradient from dark to light represents the publication year, ranging from the earliest to the most recent studies. Darker blue shades indicate older research topics, while yellow shades highlight more recent discussions.

Keywords appearing in the yellow spectrum, such as "performance evaluation" and "dataset", suggest that these topics have gained increased attention in recent studies. This indicates a shift in research focus towards

evaluating the performance of click fraud detection models, emphasizing how dataset quality and characteristics influence model performance.

Additionally, the presence of the keyword "class skewness" reinforces that class imbalance in datasets has become a critical issue with direct implications for detection performance. In the early stages of research, the primary concern was to develop models capable of distinguishing between fraudulent and legitimate clicks, often evaluated using balanced or synthetically constructed datasets. However, as the field has matured, scholars have increasingly acknowledged that imbalanced class distributions are the rule rather than the exception in real-world advertising environments. This recognition has led the research community to treat class skewness not as a peripheral modeling concern, but as a core methodological and operational challenge, particularly in the context of rare-event classification and cost-sensitive decision-making.

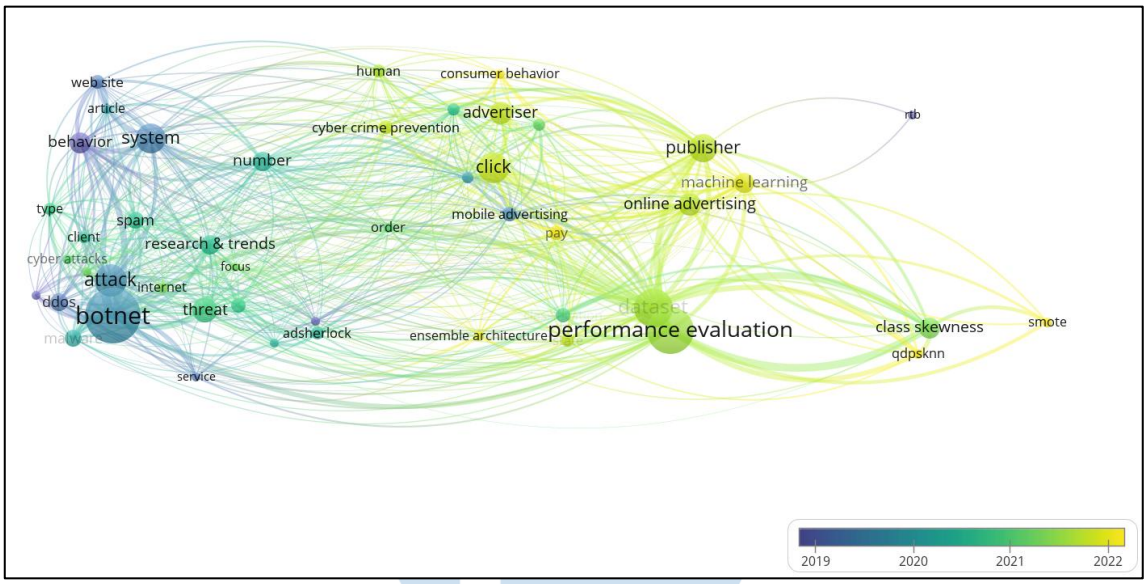



Fig 6. Overlay visualization of keyword co-occurrence in click fraud detection research

The bibliometric analysis conducted in this study reveals that research on Click Fraud Detection has experienced significant growth since 2019. This publication trend aligns with the increasing demand within the digital industry to address fraudulent activities in online advertising. Prior to 2019, the number of publications in this field remained relatively low, indicating that the research was still in an early exploratory phase. However, a notable surge in publications occurred in 2020 and 2021, signaling a heightened academic interest in Click Fraud Detection as a major challenge within the digital advertising industry.

From a geographical perspective, the study demonstrates that India leads in terms of the number of publications in this area, followed by China, Saudi Arabia, and the United States. India's dominance not only reflects the rapid expansion of its information technology and digital marketing sectors but also corresponds with the substantial growth of its digital advertising economy. The high prevalence of click fraud underscores the urgent need for more effective machine learning-based detection methods and helps explain the considerable academic focus on this issue in the region.

The network map analysis conducted in this study identified four major clusters within the field of Click Fraud Detection, reflecting the conceptual evolution and interconnections in the research domain. These clusters not only represent distinct conceptual foci within the literature but also offer substantial practical implications for the digital advertising ecosystem.

- The first cluster, centered on cybersecurity, indicates that click fraud is frequently integrated with broader digital threats such as botnets and phishing, necessitating the development of ML-based mitigation strategies that can be embedded into the IT security infrastructures of advertising firms.
- The second cluster, which explores the relationship between click fraud and digital advertising business models such as real-time bidding (RTB), has direct implications for financial risk management, campaign budget optimization, and the selection of more credible publishing partners.
- The third cluster underscores the critical role of data processing techniques and model evaluation in addressing challenges such as class imbalance (class skewness), which is common in digital advertising datasets and can

degrade the performance of detection models. This highlights the need for advertising service providers to invest in machine learning systems capable of handling real-world data in a more representative manner.

- The fourth cluster, which focuses on system development and user behavior, emphasizes the importance of building adaptive detection architectures based on behavioral profiling. Such systems can be integrated into advertising platforms to monitor user activity in real-time and identify suspicious clicking patterns.

The identified thematic clusters reveal not only the current structure of research in click fraud detection but also highlight unresolved challenges that require sustained scholarly attention. For instance, the emergence of class skewness and SMOTE in Cluster 3 points to a persistent data imbalance problem that undermines model generalization—particularly when fraudulent instances represent a small fraction of total user behavior. This is a real-world constraint in ad ecosystems, where genuine traffic far exceeds malicious activity. Addressing this issue will require future research to move beyond oversampling techniques toward advanced solutions such as cost-sensitive learning, meta-learning, and anomaly-aware classifiers optimized for rare-event detection.

Similarly, the prominence of botnet, malware, and traffic pattern in Cluster 1 signals the growing sophistication of automated fraud actors that mimic legitimate click behavior. These trends call for detection systems that integrate behavioral profiling and network anomaly detection, capable of adapting to adversarial tactics in real time. Meanwhile, Cluster 4's focus on system, behavior, and architecture highlights a translation gap between algorithmic models and their deployment in production environments. This emphasizes the need for scalable, explainable models that can operate within latency-sensitive systems such as real-time bidding (RTB) platforms.

As a whole, these trends suggest that future research must be multidimensional: advancing algorithmic resilience, integrating domain-specific behavioral cues, and aligning model performance with operational constraints. A promising direction includes the development of end-to-end fraud detection pipelines that fuse unsupervised anomaly detection, explainable AI (XAI), and economic impact modeling, thereby enabling fraud mitigation strategies that are not only accurate but also actionable and transparent in commercial advertising environments. For instance, explainable AI frameworks such as LIME introduced for model-agnostic interpretability in general classification tasks [20] have since been widely adopted across domains requiring transparency and trust, including fraud detection and high-stakes automated decision-making. These developments underscore the growing feasibility of integrating transparency, adaptability, and interpretability into real-time fraud mitigation pipelines.

The overlay visualization in Fig 6 reveals a shift in focus within Click Fraud Detection research using Machine Learning. Keywords such as "performance evaluation" and "dataset", appearing in the yellow spectrum, indicate a growing emphasis on model evaluation and dataset quality in recent studies. This trend suggests that the scientific community is becoming increasingly aware of the importance of proper data management to enhance the performance of Click Fraud Detection models.

Additionally, the term "class skewness" has emerged as one of the high co-occurrence keywords in recent studies. This indicates that challenges related to class imbalance in datasets are becoming a major concern, as Click Fraud datasets typically exhibit an imbalanced class distribution [4], [5], [21], [22], [23], [24], [25], [26], [27]. Consequently, Machine Learning methods need to be adapted to effectively handle this issue.

## IV. CONCLUSIONS

The bibliometric analysis in this study reveals that research on Click Fraud Detection has grown significantly in recent years, as evidenced by the increasing number of publications and the broadening scope of international collaboration. The keyword network mapping indicates that research in this domain can be categorized into four major clusters: cybersecurity, the digital advertising industry, dataset evaluation and processing, and the development of more adaptive detection systems. Recent research trends have shifted toward improving dataset quality and model evaluation, suggesting that data validity and the effectiveness of detection methods are becoming central concerns in current scholarly investigations.

Based on these findings, this study offers several practical recommendations. For researchers, it is essential to develop click fraud detection models that can operate in real-time and to implement approaches based on explainable AI (XAI) in order to enhance the transparency and accountability of detection systems. Furthermore, future research agendas should include the exploration of blockchain technology as a foundation for building more secure and decentralized digital advertising systems.

For regulators and policymakers, there is a need for stricter regulations regarding ad traffic verification, as well as the development of policy frameworks that facilitate ethical data sharing between digital platforms and research institutions. For the digital advertising industry, the adoption of real-time detection systems based on machine learning and behavioral profiling can strengthen resilience against click fraud manipulation while simultaneously improving the efficiency of advertising budget management.

## ACKNOWLEDGMENT

The authors would like to thank the Buddhi Dharma University and the supervisors who have supported this research.

## REFERENCES

[1] D. S. Soemarwoto, "PEMANTAPAN EKONOMI DIGITAL GUNA MENINGKATKAN KETAHANAN NASIONAL," *Jurnal Lembaga Ketahanan Nasional Republik Indonesia*, vol. 8, no. 1, pp. 1–6, Oct. 2022, doi: https://doi.org/10.55960/jlri.v8i1.299.

[2] W. Wang *et al.*, "Digital economy sectors are key CO2 transmission centers in the economic system," *J Clean Prod*, vol. 407, 2023, doi: 10.1016/j.jclepro.2023.136873.

[3] C. M. Simamora, R. Ningsih, P. Pendidikan, P. Perdagangan, P. Pengkajian, and P. L. Negeri, "INKLUSIVITAS EKONOMI DIGITAL DI INDONESIA: PERSPEKTIF GENDER DAN PENCIPTAAN LAPANGAN KERJA (STUDI KASUS KAMPUNG MARKETER)," 2020

[4] G. S. Thejas, S. Dheeshjith, S. S. Iyengar, N. R. Sunitha, and P. Badrinath, "A hybrid and effective learning approach for Click Fraud detection," *Machine Learning with Applications*, vol. 3, p. 100016, Mar. 2021, doi: 10.1016/j.mlwa.2020.100016

[5] M. Aljabri and R. M. A. Mohammad, "Click fraud detection for online advertising using machine learning," *Egyptian Informatics Journal*, vol. 24, no. 2, pp. 341–350, Jul. 2023, doi: 10.1016/j.eij.2023.05.006.

[6] D. Sisodia and D. S. Sisodia, "Quad division prototype selection-based k-nearest neighbor classifier for click fraud detection from highly skewed user click dataset," *Engineering Science and Technology, an International Journal*, vol. 28, Apr. 2022, doi: 10.1016/j.jestch.2021.05.015.

[7] M. Aljabri and R. M. A. Mohammad, "Click fraud detection for online advertising using machine learning," *Egyptian Informatics Journal*, vol. 24, no. 2, pp. 341–350, Jul. 2023, doi: 10.1016/j.eij.2023.05.006.

[8] R. Oentaryo *et al.*, "Detecting Click Fraud in Online Advertising: A Data Mining Approach Ghim-Eng Yap," *Journal of Machine Learning Research*, vol. 15, pp. 99–140, 2014, [Online]. Available: http://palanteer.sis.smu.edu.sg/fdma2012/.

[9] L. F. Cardona, J. A. Guzmán-Luna, and J. A. Restrepo-Carmona, "Bibliometric Analysis of the Machine Learning Applications in Fraud Detection on Crowdfunding Platforms," Aug. 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/jrfm17080352

[10] N. Donthu, S. Kumar, D. Mukherjee, N. Pandey, and W. M. Lim, "How to conduct a bibliometric analysis: An overview and guidelines," *J Bus Res*, vol. 133, pp. 285–296, Sep. 2021, doi: 10.1016/j.jbusres.2021.04.070.

[11] W. Glänzel and H. F. Moed, "Opinion paper: Thoughts and facts on bibliometric indicators," *Scientometrics*, vol. 96, no. 1, pp. 381–394, 2013, doi: 10.1007/s11192-012-0898-z.

[12] J. C. Donohue, *Understanding Scientific Literatures: A Bibliometric Approach*. Cambridge, MA: MIT Press, 1974. [Online]. Available: https://mitpress.mit.edu/9780262040396/understanding-scientific-literatures/

[13] D. Guo, H. Chen, R. Long, H. Lu, and Q. Long, "A co-word analysis of organizational constraints for maintaining sustainability," *Sustainability (Switzerland)*, vol. 9, no. 10, pp. 1–19, Oct. 2017, doi: 10.3390/su9101928.

[14] A. Lis, "Keywords Co-occurrence Analysis of Research on Sustainable Enterprise and Sustainable Organisation," *Journal of CorporAte Responsibility and Leadership*, vol. 5, pp. 48–66, 2018, doi: 10.12775/JCRL.2018.011.

[15] A. Klarin, "How to conduct a bibliometric content analysis: Guidelines and contributions of content co-occurrence or co-word literature reviews," *Int J Consum Stud*, vol. 48, no. 2, Mar. 2024, doi: 10.1111/ijcs.13031

[16] S. Sadeghpour and N. Vlajic, "Click fraud in digital advertising: A comprehensive survey," *Computers*, vol. 10, no. 12, Dec. 2021, doi: 10.3390/computers10120164.

[17] S. Nagaraja and R. Shah, "Clicktok: Click fraud detection using traffic analysis," in *WiSec 2019 - Proceedings of the 2019 Conference on Security and Privacy in Wireless and Mobile Networks*, Association for Computing Machinery, Inc, May 2019, pp. 105–116. doi: 10.1145/3317549.3323407.

[18] T. G.S., S. Dheeshjith, S. S. Iyengar, N. R. Sunitha, and P. Badrinath, "A hybrid and effective learning approach for Click Fraud detection," *Machine Learning with Applications*, vol. 3, p. 100016, Mar. 2021, doi: 10.1016/j.mlwa.2020.100016.

[19] Neeraja, Anupam, Sriram, S. Shaik, and V. Kakulapati, "Fraud Detection of AD Clicks Using Machine Learning Techniques," *J Sci Res Rep*, vol. 29, no. 7, pp. 84–89, Jun. 2023, doi: 10.9734/jsrr/2023/v29i71762.

[20] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?' Explaining the predictions of any classifier," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, Aug. 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.

[21] A. Batool and Y. C. Byun, "An Ensemble Architecture Based on Deep Learning Model for Click Fraud Detection in Pay-Per-Click Advertisement Campaign," *IEEE Access*, vol. 10, pp. 113410–113426, 2022, doi: 10.1109/ACCESS.2022.3211528.

[22] Y. Mhaske, A. Gupta, V. Bhosale, and K. Nair, "Click Fraud Detection Of Advertisements using Machine Learning," *International Research Journal of Engineering and Technology*, vol. 09, pp. 908–912, Apr. 2022, [Online]. Available: www.irjet.net

[23] Neeraja, Anupam, Sriram, S. Shaik, and V. Kakulapati, "Fraud Detection of AD Clicks Using Machine Learning Techniques," *J Sci Res Rep*, vol. 29, no. 7, pp. 84–89, Jun. 2023, doi: 10.9734/jsrr/2023/v29i71762.

[24] D. Sisodia and D. S. Sisodia, "A transfer learning framework towards identifying behavioral changes of fraudulent publishers in pay-per-click model of online advertising for click fraud detection," *Expert Syst Appl*, vol. 232, p. 120922, Dec. 2023, doi: 10.1016/j.eswa.2023.120922.

[25] D. Sisodia and D. S. Sisodia, "Quad division prototype selection-based k-nearest neighbor classifier for click fraud detection from highly skewed user click dataset," *Engineering Science and Technology, an International Journal*, vol. 28, p. 101011, Apr. 2022, doi: 10.1016/j.jestch.2021.05.015

[26] D. Sisodia and D. S. Sisodia, "Feature space transformation of user-clicks and deep transfer learning framework for fraudulent publisher detection in online advertising," *Appl Soft Comput*, vol. 125, p. 109142, Aug. 2022, doi: 10.1016/j.asoc.2022.109142

[27] F. Zhu, C. Zhang, Z. Zheng, and S. Al Otaibi, "Click Fraud Detection of Online Advertising-LSH Based Tensor Recovery Mechanism," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 9747–9754, Jul. 2022, doi: 10.1109/TITS.2021.3107373.

ISSN 2085-4552

# Adagrad Optimizer with Compact Parameter Design for Endoscopy Image Classification

Sofyan Pariyasto[1], Suryani[2], Vicky Arfeni Warongan[3], Arini Vika Sari[4], Wahyu Wijaya Widiyanto[5]
[1,2,3] Medical Informatics, Sekolah Tinggi Ilmu Kesehatan Mitra Sejati, Medan, Indonesia
[4]Information Technology Education, Universitas Budi Darma, Medan, Indonesia
[5]Health Information Management (D4), Politeknik Indonusa Surakarta, Surakarta, Indonesia
[1]spariyasto@gmail.com, [2]suryani90harahap@gmail.com, [3]vickyarfeni@gmail.com,
[4]arinivika1@gmail.com, [5]dewawijaya@poltekindonusa.ac.id

Accepted 19 November 2025
Approved 06 January 2026

***Abstract*— Research on CNN Model and Adagrad Optimizer is expected to help identify diseases in the medical world. Especially in the field of image classification in Gastrointestinal endoscopic procedures . The research is specifically for the process of classifying medical images of Diverticulosis, Neoplasm, Peritonitis and Ureters . Previously, there have been quite a lot of studies on CNN and its various optimizers. However, those who have studied the Adagrad optimizer are not too many, especially those discussing the use of minimum parameters. The use of minimum parameters is expected to be one of the contributions of researchers in the fields of computing and medicine. The research was conducted to determine the use of the best parameters and obtain the highest level of accuracy. The research was conducted using minimum epochs starting from epoch 1, epoch 5, and epoch 10. Then the combination process between epoch and the number of convolution layers between 1 and 5 was carried out, resulting in 15 combinations. The test was carried out using 4000 images with 1000 images in each class. From the results of the test, the highest accuracy value was obtained, namely 82.875%. Then the highest average accuracy value was 81.625%. The average CPU usage ranges from 30.42% to 32.69%. And the average computation time ranges from 24.22 seconds to 229.542 seconds. From the research conducted with the use of minimum parameters, short computation time and little resource usage can produce a model with an average accuracy level above 70%.**

***Index Terms*— Adagrad Optimizer; CNN; Endoscopy Image; Image Classification; Minimum Parameters**.

## I. INTRODUCTION

The number of optimizers used in Convolutional Neural Networks is one proof of the development of science[1] . Many optimizers used in *Convolutional Neural Network* (CNN) models include Adadelta, Adagrad, Adam, RMSprop and SGD. Each type of optimizer exhibits different performance and efficiency depending on the characteristics of the data and the architectural model. Previous research has described the applicability and accuracy of various optimizers on specific datasets[2], [3], [4] . However, there is no one that focuses on discussing the Adagrad optimizer with minimum parameters. Research related to CNN models usually uses many hypter parameters to get the best results.

Adagrad (Adaptive Gradient Algorithm) offers the advantage of dynamically adjusting the learning rate for each parameter based on its gradient history, enabling faster convergence on sparse data and efficient handling of diverse feature magnitudes[5], [6]. Compared to Adam or RMSprop, Adagrad requires fewer hyperparameters and computational resources, which makes it suitable for applications that require efficiency in both computation and memory usage. Previous works have shown its potential for resource-limited systems[7], [8], [9].

Despite these advantages, limited research has explored the use of Adagrad with compact parameter design in medical imaging contexts. Most studies rely on heavy computational resources or complex optimizers, leading to inefficiency in deployment. This study aims to evaluate the performance of Adagrad with minimal parameters for classifying gastrointestinal endoscopy images, emphasizing computational efficiency, accuracy, and model simplicity.

The research conducted this time focused on one optimizer to obtain the highest accuracy results with the shortest computing time and the least resource usage. This study applies a strategy of using minimum parameters on the optimizer to achieve efficiency in the model training process. The use of minimum parameters in the optimizer is expected to reduce the workload of the hardware used. The use of CNN models in deep learning, especially in terms of classification, is expected to help identify diseases in the medical world. Health care in the field of *Gastrointestinal endoscopic procedures* is one of the topics in this study. Gastrointestinal endoscopic procedures are medical procedures that use a special tool called an endoscope to examine, diagnose, or treat problems in the digestive tract, including the esophagus, stomach, small intestine, and large intestine[10], [11]. An endoscope is a long, flexible, and thin tool equipped with a camera at the end[12]. This

tool allows doctors to see the inside of the digestive tract directly without the need for major surgery.

The medical image classification process carried out in this study focuses on several medical conditions only, namely, *Diverticulosis, Neoplasm, Peritonitis, Ureters* . *Diverticulosis* is a condition in which there are small pockets or balloons in the intestinal wall, especially in the large intestine. These pockets can form due to excessive pressure in the intestine[13]. *Neoplasm* is the medical term for a tumor or abnormal tissue growth. These tumors can be benign (harmless) or malignant (cancerous)[14], [15]. So, neoplasms include all types of abnormal cell growth in the body. *Peritonitis* is an inflammation of the peritoneum, which is a thin layer that lines the inner wall of the abdomen and protects the organs in the abdomen[16], [17], [18]. This inflammation is usually caused by infection, which can occur because an organ in the abdomen ruptures, such as a perforated intestine. Peritonitis is a serious condition and requires immediate medical attention. *The ureter* is a tube that connects the kidney to the bladder[19], [20], [21], [22], [23], [24]. Its job is to carry urine produced by the kidney to the bladder, where it is stored before being excreted from the body.

The next process is the process of creating models using minimum parameters, the parameters used are in Epochs and Convolution Layers. The use of epochs starts from the smallest epoch, namely 1, 5 and the largest epoch, namely 10. The computation process is carried out with a combination of epochs and convolution layers, each computation is carried out and produces 1 model. Then the number of Convolution Layers in this study starts from 1 convolution layer to 5 convolution layers. Where there are 5 combinations of layers and 3 combinations of epochs.

The last process carried out is the evaluation of the algorithm performance. The algorithm performance evaluation process is carried out using the Confusion Matrix. The algorithm performance is carried out on each model produced, so that the Precision, Accuracy, Recall and F-1 Score values can be calculated from the resulting models. The algorithm performance evaluation is carried out to obtain the best minimum parameter combination results. With the best parameters, it is expected that the use of the optimizer Adagrad can be more optimal in the classification process.

## II. THEORY

The literature study process was conducted to explore more deeply what previous researchers have done in the image classification process. Several studies that discuss the CNN model and its optimizer include. The literature study process was conducted to explore more deeply what previous researchers have done in the image classification process. Some studies that discuss CNN models and optimizers include the following.

The study that discusses the CNN literature study in the cat image classification process by comparing 16 studies from previous researchers[25]. This study focuses on comparing the results of using the CNN model on the same object, namely cats. There is also another study[26]regarding CNN which also discusses the breed classification process in cats. This study discusses CNN and RMSprop optimizer. Another study that discusses CNN was also conducted for the caterpillar pest detection process in Aquaponic plants[27]. This study focuses on the caterpillar identification process with the CNN model which produces an accuracy value of 89%.

From several studies that have been conducted, some focus on hypter parameters, and some focus on increasing model accuracy against datasets. However, there is no specific research on the use of minimum parameters. In previous studies, there was a lot of focus on default parameters and the use of optimizers that were generally not detailed. One study that discussed the optimizer used was RMSprop. While this study focuses on the Adagrad optimizer to find the best parameters with the fastest computer time and the least resource usage

Endoscopy is a medical procedure performed to examine parts of the human body, especially the upper digestive tract or small intestine. Endoscopy is performed using a tool that is inserted into the digestive tract to capture images of the small intestine.[28]. With endoscopy it is possible to examine the digestive tract without undergoing surgery[29], this is certainly easier to do compared to examinations that involve surgery.

The model used in the study is CNN using the scikit learn library and using the python programming language. Classification is done using the Adagrad optimizer to find out the best parameters that can be used for the classification process.

## III. METHOD

There are several stages carried out in this research, including literature review, data preprocessing, creating CNN models and performance evaluation. The flow of the research process carried out is shown in Figure 1 below.

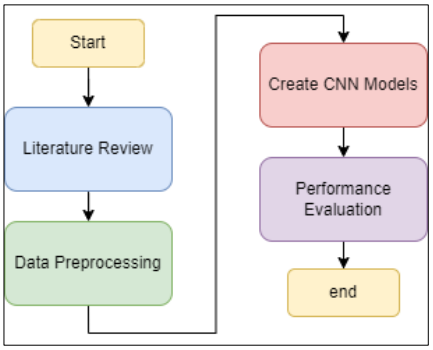



Figure 1. Flow of the research process carried out

### A. *Data Preprocessing*

In preprocessing is done to ensure that the research can run smoothly. One of the important stages carried out is to prepare data in the research. Without processed data, research cannot be carried out. The data used in this study is a public dataset taken from kaggle. The dataset used is an image taken from the *Gastrointestinal endoscopic procedures process* , there are 4000 images in the dataset[30]

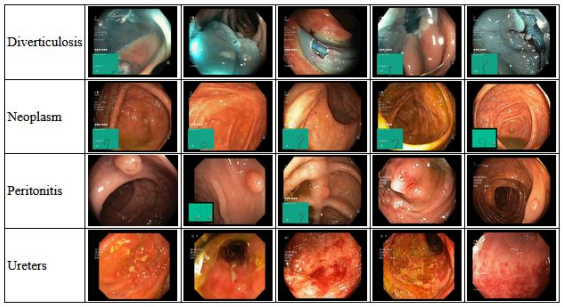



Figure 2. Gastrointestinal endoscopic dataset

The dataset consists of 4 classes with each class consisting of 1000 images. The classes in the dataset are Diverticulosis, Neoplasm, Peritonitis, Ureters. Each image used will later be divided into training data and testing data with a composition of 80% training data and 20% testing data. In detail, the flow of the data preprocessing process is seen in Figure 3 below.

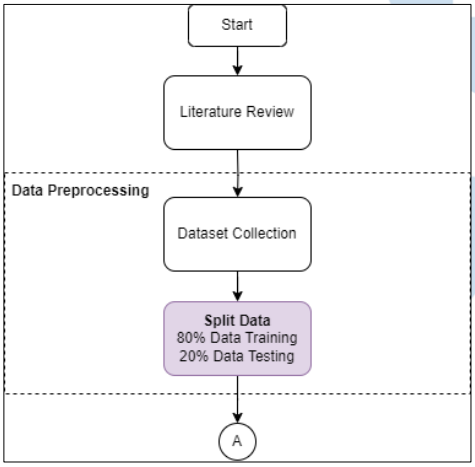



Figure 3. Initial research stage flow

From figure 3, it can be seen that 80% of the dataset used will be used as training data, which means the amount of training data used is 3200 images, and the testing data used is 800 images.

### B. *CNN Model*

Convolutional Neural Networks (CNN) are models that are specialized for image recognition and have high accuracy in classifying objects in images[31], [32]. Convolutional Neural Networks (CNN) are a type of computer network that uses a mathematical technique called convolution. This technique allows the network to find patterns in data, such as images, by examining small parts of the data. In this way, even if certain patterns are not directly described in the data used to train it, the network can still recognize those patterns[33]. An illustration of the CNN model looks like the following figure 4:

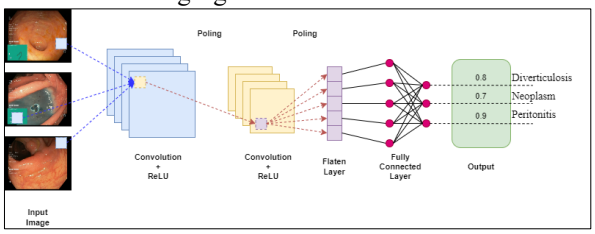



Figure 4. Illustration of CNN Model

The general equation regarding the convolutional layer operation looks like equation 1[29], [30] below.

$$S(i,j) = (I * K)(i,j) = \sum_{m}\sum_{n} I(i+m, j+n) \cdot K(m,n) \quad (1)$$

$K(m,n)$describes the convolution operation between an image $I$ and a kernel $K$at position $(i,j)$. The process is carried out to calculate the value of each element in the convolution result $S$ based on the interaction between the image and the kernel. In this formula, $S(i,j)$shows the convolution result at a point $(i,j)$calculated as the sum of the products of the elements that are interconnected between the image $I$and the kernel $K$. $I(i+m, j+n)$is the image element $I$at position $(i+m, j+n)$, where mm and nn are indices that run along the kernel dimension $K$ . The kernel $K(m,n)$is the element at position $(m,n)$in the kernel used to calculate the convolution.

In this study, CNN can be used as a solution for the visual data classification process in mapping neural network models using minimum parameters on epochs and the number of convolution layers. The use of dynamic optimization algorithms such as ADAGRAD can help adjust learning values continuously without the need to make adjustments at the beginning. The stages of the process carried out in creating a CNN model are shown in Figure 5 below.

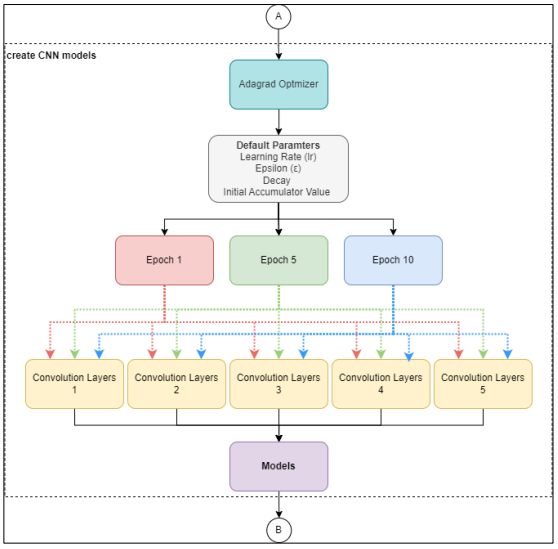



Figure 5. Flow of the CNN Model creation process with each parameter

The stages carried out in the process of creating CNN models start from selecting the use of optimizers (Adagrad Optimizer). The next stage is determining the parameter values of the optimizer used, namely learning rate (lr), Epsilon (ε), Decay, and Initial Accumulator value. The default parameters used at this stage contain the default values available in the pytorch library. The next stage is to determine the number of iterations (epochs) used starting from the minimum value available, namely 1, then 5 and 10. After the batch value is determined, it is continued by determining the number of convolution layers performed . The number of convolution layers used starts from 1 to 5 convolution layers.

*1) Adagrad Optimizer*

Adagrad (Adaptive Gradient Algorithm) optimizer is an algorithm used to improve the way computer models are trained, such as convolutional neural networks (CNNs). Adagrad's main advantage lies in its ability to automatically adjust the learning rate for each element of the model, according to how often the element is updated during training[36], [37], [38]. Using Adagrad, elements that are updated less frequently will experience faster learning improvements, while elements that are updated more frequently will experience slower learning[39].

This allows the model to adapt to different types of data and avoid learning too fast or too slow in some parts[40], [41]. Although Adagrad can speed up training on rare data, its drawback is that the learning rate decreases over time due to the accumulation of gradients from previous iterations, which can cause the model to stop learning earlier than desired. The equation used in the Adagrad optimizer looks like equation 2 below[42], [43].

$$\theta_{t+1,i} = \theta_{t,i} - \frac{\eta}{\sqrt{G_{t,ii} + \varepsilon}} \cdot g_{t,i} \quad (2)$$

From equation 2, it is known that $\theta_{t+1,i}$ is the parameter result of the update from the adagrad optimizer. While $\theta_{t,i}$is the value of the previous parameter update result. η is the learning rate, ε is a small scalar parameter to avoid division by zero. The Adagrad optimizer works by calculating the gradient to adjust the learning rate, then changing the learning rate based on the calculated gradient. The Adagrad optimizer is designed to optimize the learning rate at each iteration, so it can improve the model accuracy without having to change the initial values.

*2) Adagrad Parameters*

Adagrad has several parameters used in the computational process in producing models. Commonly used parameters are Learning rate (lr), Epsilon (ε), Decay, and Initial Accumulator value. The learning rate (lr) parameter has a value of 0.01 which is the default value. Then Epsilon (ε) has a value of 1e-8 or 0.00000001, and decay has a value of 0.0, and the Initial Accumulator value has a value of 0.0[44]. The parameters that will be optimized in this study are Epoch and the number of convolution layers. The use of the epoch parameter starts from the smallest parameter, namely epoch 1, then epoch 5 then epoch 10. Then the number of convolution layers used in this study starts from 1 convolution layer, up to 5 convolution layers. With a total combination of epoch and convolution layer as many as 15 combinations.

## C. Performance Evaluation

To measure the performance of the adagrad optimizer, an analysis was carried out using the confusion matrix method. Measurement of algorithm performance is carried out to determine the Precision, accuracy, recall and F1-Score values from the research that has been done. By using the confusion matrix, it is expected to know the performance of each model produced in the form of numbers. And the best classification model with the highest accuracy value and the shortest computing time, as well as the least resource usage, will be known. The flow carried out in the evaluation process is seen as in Figure 6 below.

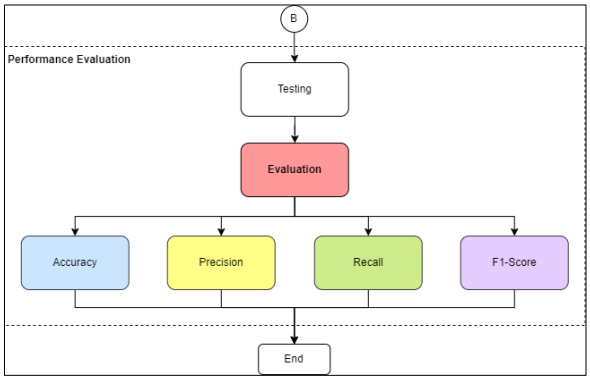



Figure 6. Performance evaluation process flow with confusion matrix

From figure 6, it can be seen that there are 4 measurement components that will be calculated in the evaluation process, namely accuracy, precision, recall and F1-Score. To be able to calculate these values, the TP, TN, FP and FN values must first be determined. The confusion matrix is arranged in the form of a table containing actual values and predicted values[45]as shown in table 1 below.

TABLE 1. CONFUSION MATRIX

| | **Positive Prediction (P)** | **Negative Prediction (N)** |
|---|---|---|
| **Positive Actual (P)** | True Positive (TP) | False Negative (FN) |
| **Negative Actual (N)** | False Positive (FP) | True Negative (TN) |

True Positive (TP) is the number of images that actually belong to a certain class and are correctly predicted by the model. False Positive (FP) is the number of images that are predicted to belong to a certain class, but actually belong to another class. False Negative (FN) is the number of images that actually belong to a certain class, but are predicted to be of another class. True Negative (TN) is an image that actually does not belong to a certain class and is correctly predicted[46], [47], [48].

Precision measures how many positive predictions are actually positive. In image classification, it indicates how many images are correctly predicted and are true. The *precision equation* looks like equation 3[49]below.

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

Recall (Sensitivity) is used to measure how many truly positive images the model successfully predicted correctly. This is also known as *True Positive Rate* (TPR). The *recall equation* looks like the following equation 4.

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

*F1-Score* Is the harmonic mean of precision and recall. *F1-Score* provides a better overview when we need a balance between the two. The *F1-Score equation* looks like the following equation 5.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (5)$$

## IV. RESULT AND DISCUSSIONS

The test was conducted using 800 images as testing data with 200 images for each class . The testing process displays the predicted image and the actual image, as shown in Figure 7 below.

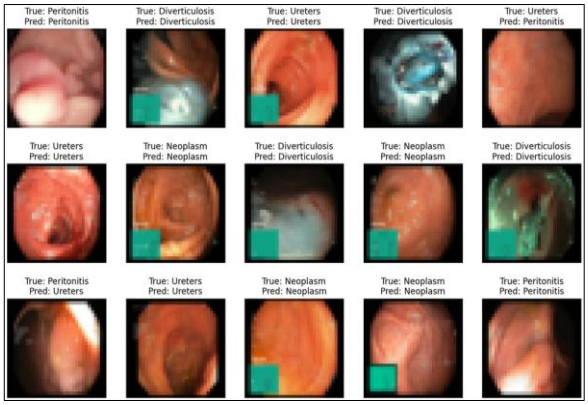



Figure 7. Endoscopic image classification process

Figure 7 shows the endoscopic image classification process by displaying the actual (true) values and predicted values from the Adagrad optimizer.

### *A. Adagrad Optimizer with Epoch 1*

The process starts from the minimum epoch value of 1 combined with the number of convolution layers with a value of 1 to 5 convolution layers. The test results are shown in Table 2 below.

TABLE 2. TEST RESULTS OF EPOCH 1

| **Epoch** | **Number of Convolution Layers** | **Support** | **Precision** | **Recall** | **F1-score** | **Accuracy** | **Time (seconds)** | **CPU Usage (%)** |
|---|---|---|---|---|---|---|---|---|
| **1** | 1 | 800 | 0.77 | 0.76 | 0.77 | 76,375 | 24.09 | 31.5 |
| **1** | 2 | 800 | 0.79 | 0.78 | 0.78 | 77.5 | 23.46 | 31.75 |
| **1** | 3 | 800 | 0.74 | 0.69 | 0.67 | 68,875 | 23.97 | 32.15 |
| **1** | 4 | 800 | 0.69 | 0.69 | 0.67 | 69 | 24.25 | 32.1 |
| **1** | 5 | 800 | 0.69 | 0.69 | 0.69 | 69.125 | 25.33 | 32.45 |

From table 2, it can be seen that the highest accuracy value is 77.5 with 2 convolution layers. The fastest computing time is 23.46 seconds and the lowest CPU usage is 31.5%.

### *B. Adagrad Optimizer with Epoch 5*

The second process is testing with an epoch value of 5 then combined with the number of convolution layers with a value of 1 to a convolution layer of 5. The test results are shown in Table 3 below.

TABLE 3. TEST RESULTS OF EPOCH 5

| **Epoch** | **Number of Convolution** | **Support** | **Precision** | **Recall** | **F1-score** | **Accuracy** | **Time (seconds)** | **CPU Usage (%)** |
|---|---|---|---|---|---|---|---|---|

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **5** | **Layers** 1 | 800 | 0.84 | 0.83 | 0.83 | 82,875 | 113.21 | 31.85 |
| **5** | 2 | 800 | 0.81 | 0.81 | 0.8 | 80.5 | 123.73 | 31.1 |
| **5** | 3 | 800 | 0.83 | 0.83 | 0.82 | 82,625 | 158.03 | 23.35 |
| **5** | 4 | 800 | 0.71 | 0.71 | 0.66 | 70,625 | 110.17 | 33.2 |
| **5** | 5 | 800 | 0.78 | 0.76 | 0.75 | 76,375 | 116.33 | 32.6 |

From table 3, the highest accuracy value is 82,875 with 1 convolution layer. The fastest computing time is 110.17 seconds and the lowest CPU usage is 23.35%.

### *C. Adagrad Optimizer with Epoch 10*

The final process is testing with an epoch value of 10 then combined with the number of convolution layers with a value of 1 to a convolution layer of 5. The test results are shown in Table 4 below.

TABLE 4. TEST RESULTS OF EPOCH 10

| Epoch | Number of Convolution Layers | Support | Precision | Recall | F1-score | Accuracy | Time (seconds) | CPU Usage (%) |
|---|---|---|---|---|---|---|---|---|
| **10** | 1 | 800 | 0.81 | 0.81 | 0.8 | 80.75 | 212.97 | 33 |
| **10** | 2 | 800 | 0.82 | 0.81 | 0.8 | 80,625 | 213.56 | 33.45 |
| **10** | 3 | 800 | 0.83 | 0.82 | 0.81 | 82.125 | 235.3 | 32.15 |
| **10** | 4 | 800 | 0.82 | 0.82 | 0.82 | 82.5 | 236.59 | 32.55 |
| **10** | 5 | 800 | 0.83 | 0.82 | 0.82 | 82.125 | 249.29 | 32.3 |

From table 4, the highest accuracy value is 82,625 with 2 convolution layers. The fastest computing time is 212.97 seconds and the lowest CPU usage is 32.15%.

### *D. Average Performance*

Based on the epoch, the average performance value of the model generated from CNN and Adagrad optimizer is shown in Table 5 below.

TABLE 5. AVERAGE OF TEST RESULTS

| Epoch | Support | Precision | Recall | F1-score | Accuracy | Time (seconds) | CPU Usage (%) |
|---|---|---|---|---|---|---|---|
| **1** | 800 | 0.736 | 0.722 | 0.716 | 72.175 | 24.22 | 31.99 |
| **5** | 800 | 0.794 | 0.788 | 0.772 | 78.6 | 124,294 | 30.42 |
| **10** | 800 | 0.822 | 0.816 | 0.81 | 81,625 | 229,542 | 32.69 |

From table 5, it can be seen that the highest average accuracy value is 81.625 with the number of epochs 10. Meanwhile, the fastest average computing time is 24.22 seconds and the lowest average CPU usage is 30.42%.

## V. CONCLUSION

From the research conducted using the CNN model and Adagrad optimizer, it was concluded that the highest accuracy value was 82.875%. The average CPU usage ranged from 30.42% to 32.69%. And the average computing time ranged from 24.22 seconds to 229.542 seconds. The average accuracy value is still above 70%, this can be seen in table 5, where the lowest average accuracy value is 72.175%. From the tests carried out, it can be concluded that a model with an accuracy level above 70% can be produced with minimum parameters on the Adagrad optimizer and CNN model. These results show that applying minimum parameters to the optimizer not only maintains a good level of accuracy, but also significantly speeds up the computation time, with the fastest average time being 24.22 seconds.

## REFERENCES


[1] K. Usman, N. K. C. Pratiwi, N. Ibrahim, H. Syahrian, and V. P. Rahadi, "Evaluasi Optimizer pada Residual Network untuk Klasifikasi Klon Teh Seri GMB Berbasis Citra Daun," *ELKOMIKA*, vol. 9, no. 4, p. 841, Oct. 2021, doi: 10.26760/elkomika.v9i4.841.

[2] N. A. M. Pauzi, S. M. Mustaza, N. Zainal, and M. F. Bukhori, "Transfer Learning-based Weed Classification and Detection for Precision Agriculture," *ijacsa*, vol. 15, no. 6, 2024, doi: 10.14569/IJACSA.2024.0150646.

[3] S. Shedriko and M. Firdaus, "Perbandingan Optimizer Adagrad, Adadelta dan Adam dalam Klasifikasi Gambar Menggunakan Deep Learning," *STRING*, vol. 8, no. 1, p. 103, Aug. 2023, doi: 10.30998/string.v8i1.16564.

[4] W. Huang, "Implementation of Parallel Optimization Algorithms for NLP: Mini-batch SGD, SGD with Momentum, AdaGrad Adam," *ACE*, vol. 81, no. 1, pp. 226–233, Nov. 2024, doi: 10.54254/2755-2721/81/20241146.

[5] R. Loganathan and S. Latha, "THE EMPIRICAL COMPARISON OF DEEP NEURAL NETWORK OPTIMIZERS FOR BINARY CLASSIFICATION OF OCT IMAGES," *PES*, vol. 7, no. 1, pp. 547–554, Mar. 2025, doi: 10.24874/PES07.01D.011.

[6] K. Kumar Singh *et al.*, "Deep Learning Capabilities for the Categorization of Microcalcification," *IJERPH*, vol. 19, no. 4, p. 2159, Feb. 2022, doi: 10.3390/ijerph19042159.

[7] S. Kılıçarslan, H. A. Aydın, K. Adem, and E. K. Yılmaz, "Impact of optimizers functions on detection of Melanoma using transfer learning architectures," *Multimed Tools Appl*, vol. 84, no. 14, pp. 13787–13807, June 2024, doi: 10.1007/s11042-024-19561-6.

[8] I. Kandel, M. Castelli, and A. Popovič, "Comparative Study of First Order Optimizers for Image Classification Using Convolutional Neural Networks on Histopathology Images," *J. Imaging*, vol. 6, no. 9, p. 92, Sept. 2020, doi: 10.3390/jimaging6090092.

[9] M. N. Halgamuge, E. Daminda, and A. Nirmalathas, "Best optimizer selection for predicting bushfire occurrences using deep learning," *Nat Hazards*, vol. 103, no. 1, pp. 845–860, Aug. 2020, doi: 10.1007/s11069-020-04015-7.

[10] Zahrah Malidia, Yuni Susilowati, and Siti Nurhasanah, "Pengaruh Edukasi Persiapan Endoskopi Terhadap Kepatuhan Pasien Melaksanakan Persiapan Endoskopi,"

*kesehatan*, vol. 8, no. 1, pp. 87–99, May 2019, doi: 10.37048/kesehatan.v8i1.155.

[11] M. Rutter and C. Rees, "Quality in gastrointestinal endoscopy," *Endoscopy*, vol. 46, no. 06, pp. 526–528, Apr. 2014, doi: 10.1055/s-0034-1365738.

[12] S. F. Elahi and T. D. Wang, "Future and advances in endoscopy," *Journal of Biophotonics*, vol. 4, no. 7–8, pp. 471–481, Aug. 2011, doi: 10.1002/jbio.201100048.

[13] P. C. R. Kuta and Y. Pintaningrum, "Diverticulosis and Diverticulitis: A Rarely Known Anatomical Changes of The Colon," *jk*, vol. 12, no. 2, pp. 190–194, June 2023, doi: 10.29303/jk.v12i2.4395.

[14] I. Mahrani and J. S. Lukito, "Hubungan Ekspresi Imunohistokimia Cyclooxygenase-2 (COX-2) dengan Derajat Histopatologi Meningioma".

[15] Y. Wisudarma, H. Agustina, S. Suryanti, and B. S. Hernowo, "Validitas Pemeriksaan Imunositokimia HMGA2 dalam Penegakan Diagnosis Nodul Jinak dan Ganas Tiroidpada Sediaan Biopsi Aspirasi Jarum Halus".

[16] M. Setiawan, "HUBUNGAN ANTARA KEJADIAN ASITES PADA CIRRHOSIS HEPATIS DENGAN KOMPLIKASI SPONTANEOUS BACTERIAL PERITONITIS".

[17] M. Sayuti, "KARAKTERISTIK PERITONITIS PERFORASI ORGAN BERONGGA DI RSUD CUT MEUTIA ACEH UTARA," *AVERROUS*, vol. 6, no. 2, p. 68, Dec. 2020, doi: 10.29103/averrous.v6i2.3089.

[18] A. Japanesa, A. Zahari, and S. Renita Rusjdi, "Pola Kasus dan Penatalaksanaan Peritonitis Akut di Bangsal Bedah RSUP Dr. M. Djamil Padang," *JKA*, vol. 5, no. 1, Jan. 2016, doi: 10.25077/jka.v5i1.470.

[19] L. C. Herlitz *et al.*, "Development of Focal Segmental Glomerulosclerosis after Anabolic Steroid Abuse," *Journal of the American Society of Nephrology*, vol. 21, no. 1, pp. 163–172, Jan. 2010, doi: 10.1681/ASN.2009040450.

[20] K. Adi, F. Safriadi, S. Sugandi, Z. Haroen, B. S. Noegroho, and T. Tjahjodjati, "LITOTRIPSI LASER HOLMIUM YAG UNTUK TERAPI BATU URETER," *JURI*, vol. 15, no. 2, July 2008, doi: 10.32421/juri.v15i2.352.

[21] S. S. Saputra, S. G. Ningrum, and Y. A. Prakoso, "Pengaruh Obstruksi Ureter Terhadap Kadar Uric Acid Dan Urine Chloride Pada Tikus Sprague Dawley," vol. 2, Oct. 2024.

[22] L. M, A. Alvarino, T. Tofrizal, and E. Erkadius, "Pengaruh Pemberian Valsartan Dan Kurkumin Terhadap Pembentukan Fibrosis Di Tubulus Proksimal Ginjal Akibat Obstruksi Ureter Unilateral pada Tikus Wistar.," *JKA*, vol. 2, no. 1, p. 01, Jan. 2013, doi: 10.25077/jka.v2i1.53.

[23] I. Alini and A. Rizaldi, "PENILAIAN LABORATORIS DAN RADIOLOGIK PADA KASUS NYERI KOLIK RENAL AKIBAT BATU GINJAL DAN BATU URETER DI IGD RSU PUTRI BIDADARI STABAT," vol. 6, no. 4, 2022.

[24] I. S. A. Kompiyang and S. Ketut, "RReeffllluukkss VVeessiikkoo UUrreetteerr RReeffllluukkss VVeessiikkoo UUrreetteerr Refluks Vesiko Ureter," *Sari Pediatri*, vol. 8, no. 3, 2006.

[25] K. D. Linda, K. Kusrini, and A. D. Hartanto, "Studi Literatur Mengenai Klasifikasi Citra Kucing Dengan Menggunakan Deep Learning: Convolutional Neural Network (CNN)," *JEECOM*, vol. 6, no. 1, pp. 129–137, May 2024, doi: 10.33650/jeecom.v6i1.7480.

[26] M. A. A. Fawwaz and K. N. Ramadhani, "Klasifikasi Ras pada Kucing menggunakan Algoritma Convolutional Neural Network(CNN)," Feb. 2021.

[27] M. R. S. Erwin, A. G. Putrada, and M. A. Triawan, "Deteksi Hama Ulat Pada Tanaman Selada Berbasis Aquaponic Menggunakan CNN (Convolutional Neural Network)," Oct. 2021.

[28] U. E. Novyanauli, B. Dedi, and N. E. Mauliku, "PENGARUH HEALTH EDUCATION TERHADAP PENGETAHUAN DAN PERSEPSI PASIEN SEBELUM CAPSULE ENDOSKOPI DI RS PLUIT," vol. 10, no. 2, 2024.

[29] A. H. Kuspranoto and M. Ulin Nuha Aba, M.Si, "PENDAMPINGAN INSTALASI BARU PERALATAN KESEHATAN ENDOSCOPY DI PMI KOTA SEMARANG," *JPMPM*, vol. 1, no. 1, pp. 19–23, May 2024, doi: 10.59485/abdikestrada.v1i1.44.

[30] "Medical imaging." Accessed: Jan. 02, 2025. [Online]. Available: https://www.kaggle.com/datasets/f6172f6cdc67a72a5fc61da1f417992aabda1519478230fc5c6617ddb4e1c0d7

[31] E. Rasywir, R. Sinaga, and Y. Pratama, "Analisis dan Implementasi Diagnosis Penyakit Sawit dengan Metode Convolutional Neural Network (CNN)," *Jurnal Sistem Informasi, Teknik Informatika, Software Engineering, dan Multimedia*, vol. 22, no. 2, pp. 117–123, Sept. 2020, doi: 10.31294/p.v22i2.8907.

[32] M. Arsal, B. Agus Wardijono, and D. Anggraini, "Face Recognition Untuk Akses Pegawai Bank Menggunakan Deep Learning Dengan Metode CNN," *TEKNOSI*, vol. 6, no. 1, pp. 55–63, June 2020, doi: 10.25077/TEKNOSI.v6i1.2020.55-63.

[33] Z. Rguibi, A. Hajami, D. Zitouni, A. Elqaraoui, and A. Bedraoui, "CXAI: Explaining Convolutional Neural Networks for Medical Imaging Diagnostic," *Electronics*, vol. 11, no. 11, p. 1775, June 2022, doi: 10.3390/electronics11111775.

[34] "Deep Learning by Ian Goodfellow, Yoshua Bengio, Aaron Courville (z-lib.org)."

[35] Y. Lecun, "Gradient-Based Learning Applied to Document Recognition".

[36] N. B. Aji, K. Kurnianingsih, N. Masuyama, and Y. Nojima, "CNN-LSTM for Heartbeat Sound Classification," *JOIV : Int. J. Inform. Visualization*, vol. 8, no. 2, p. 735, May 2024, doi: 10.62527/joiv.8.2.2115.

[37] J. Qi, "Federated Quantum Natural Gradient Descent for Quantum Federated Learning," Aug. 15, 2022, *arXiv*: arXiv:2209.00564. doi: 10.48550/arXiv.2209.00564.

[38] A. Défossez, L. Bottou, F. Bach, and N. Usunier, "A Simple Convergence Proof of Adam and Adagrad," Oct. 17, 2022, *arXiv*: arXiv:2003.02395. doi: 10.48550/arXiv.2003.02395.

[39] J. Qi, X.-L. Zhang, and J. Tejedor, "Optimizing Quantum Federated Learning Based on Federated Quantum Natural Gradient Descent," Feb. 27, 2023, *arXiv*: arXiv:2303.08116. doi: 10.48550/arXiv.2303.08116.

[40] A. B. Ahamed and M. M. Surputheen, "Prediction of Student's Interest on Sports for Classification using Bi-Directional Long Short Term Memory Model," *International Journal of Computer Science and Network Security*, vol. 22, no. 10, pp. 246–256, Oct. 2022, doi: 10.22937/IJCSNS.2022.22.10.32.

[41] D. Adla, G. V. R. Reddy, P. Nayak, and G. Karuna, "Deep learning-based computer aided diagnosis model for skin cancer detection and classification," *Distrib Parallel Databases*, vol. 40, no. 4, pp. 717–736, Dec. 2022, doi: 10.1007/s10619-021-07360-z.

[42] Micheal and Ery Hartati, "MDP Student Conference (MSC) 2022," 2022.

[43] M. Kouzehgar, Y. Krishnasamy Tamilselvam, M. Vega Heredia, and M. Rajesh Elara, "Self-reconfigurable façade-cleaning robot equipped with deep-learning-based crack detection based on convolutional neural networks," *Automation in Construction*, vol. 108, p. 102959, Dec. 2019, doi: 10.1016/j.autcon.2019.102959.

[44] I. Valova, C. Harris, T. Mai, and N. Gueorguieva, "Optimization of Convolutional Neural Networks for Imbalanced Set Classification," *Procedia Computer Science*, vol. 176, pp. 660–669, 2020, doi: 10.1016/j.procs.2020.09.038.

[45] Nurul A'ayunnisa, Y. Salim, and H. Azis, "Analisis Performa Metode Gaussian Naïve Bayes untuk Klasifikasi Citra Tulisan Tangan Karakter Arab," *ijodas*, vol. 3, no. 3, pp. 115–121, Dec. 2022, doi: 10.56705/ijodas.v3i3.54.

[46] A. R. Purnajaya, R. Jelita, E. Tesvara, M. Nestelrody, and J. Irwansyah, "Penerapan Metode Radial Basis Function (RBF) dalam Mengklasifikasikan Penyakit Demam Berdarah," 2023.

[47] V. S. Ginting, K. Kusrini, and E. T. Luthfi, "PENERAPAN ALGORITMA C4.5 DALAM MEMPREDIKSI KETERLAMBATAN PEMBAYARAN UANG SEKOLAH MENGGUNAKAN PYTHON," *JurTI*, vol. 4, no. 1, pp. 1–6, June 2020, doi: 10.36294/jurti.v4i1.1101.

[48] L. Swastina, "Penerapan Algoritma C4.5 Untuk Penentuan Jurusan Mahasiswa," vol. 2, no. 1, 2013.

[49] S. Proboningrum and Acihmah Sidauruk, "SISTEM PENDUKUNG KEPUTUSAN PEMILIHAN SUPPLIER KAIN DENGAN METODE MOORA," *JSiI*, vol. 8, no. 1, pp. 43–48, Mar. 2021, doi: 10.30656/jsii.v8i1.3073.

ISSN 2085-4552

# Application Of Double Exponential Smoothing Holt's Method For Poverty Line Forecasting: A Study Case of East Kalimantan Province

Akhmad Irsyad[1], Ariantika Putri Maharani[2], Muhammad Rivani Ibrahim[3]
[1,2,3] Information Systems, Mulawarman University, Samarinda, Indonesia
[1]arin.dila02@gmail.com, mrivani.ibrahim@gmail.com

Accepted 19 November 2025
Approved 07 January 2026

***Abstract*— Poverty is a complex problem faced by every region, including East Kalimantan Province. The poverty line is used as an important indicator to determine the poor population, and data from the Badan Pusat Statistik of East Kalimantan Province shows an upward trend every year, for example from IDR 796,193 per capita per month in 2023 to IDR 853,997 in 2024. This study aims to forecast the poverty line for the next ten periods using Holt's Double Exponential Smoothing (DES) method, which is suitable for data with trend patterns. The data used is the time series from the first semester of 2011 to the second semester of 2024. The analysis was carried out by determining the optimal smoothing parameters, calculating the forecast values, and evaluating the model with an error measure. The results of the study show optimal parameters of $\alpha$ = 0.98 and $\beta$ = 0.01, with a MAPE value of 4.56%. This relatively small error value indicates that Holt's DES method is effective in producing accurate forecasts. These findings are expected to provide input for local governments in formulating strategies and policies for poverty alleviation based on predictive data.**

***Index Terms*— *Poverty Line, Forecasting, Double Exponential Smoothing Holt's, East Kalimantan, MAPE*.**

## I. INTRODUCTION

Poverty is one of the fundamental, complex, and multidimensional problems faced by almost all developing countries, including Indonesia. This problem is not only related to limited income, but also has implications for various aspects such as education, health, and social welfare [1]. According to the Badan Pusat Statistik of East Kalimantan Province (BPS), poverty is defined as the inability of a person to meet basic food, and non-food needs as measured by expenditure. People with a per capita monthly expenditure below the poverty line are categorized as poor [2].

The poverty line serves as an important indicator for determining the level of community welfare and measuring the number of people living in poverty. Based on data from the East Kalimantan Provincial Statistics Agency, the poverty line in the province has shown an upward trend in recent years, from IDR 790,186 per capita per month in 2023 to IDR 833,955 per capita per month in 2024 [3]. This increase indicates that the economic conditions of the community are not yet stable and that there are still economically vulnerable groups. To support the formulation of effective policies, the government needs accurate and reliable information on the poverty line. Therefore, poverty line forecasting is important in order to predict future socioeconomic conditions and assist in planning more targeted poverty alleviation programs.

In the context of economic data analysis, Holt's Double Exponential Smoothing (DES) method is one of the effective time series approaches for predicting data with linear trend patterns. This method was developed by Charles C. Holt and has the adaptive ability to update trend value estimates based on changes in historical data [4]. Several previous studies have shown the effectiveness of Holt's DES method in producing high levels of accuracy. [1] found the smallest MAPE value of 2.541% in forecasting the poverty line in South Buru Regency. [5] showed accuracy results with a MAPE value of 1.8% in the Special Region of Yogyakarta Province, while [6] found that the Holt's DES method was more accurate than ARIMA with a MAPE difference of 3.62%.

Compared to other forecasting methods, DES Holt's has a number of advantages. This method does not require a differencing process like ARIMA, making it more efficient for data that is not too long and does not contain seasonal patterns [7]. Holt-Winters, although capable of capturing seasonal components, is less relevant for data with pure linear trend patterns. Meanwhile, machine learning methods such as LSTM or hybrid ARIMA-LSTM require large amounts of data and high computational resources [8], while the Prophet model is more suitable for data with strong cyclical and seasonal variations [9]. Compared to simple linear regression, Holt's DES is superior because it can account for autocorrelation between time periods and is

dynamic to changes in data patterns, while linear regression only provides a static relationship between time and the dependent variable [10].

Based on these issues, it can be concluded that this study focuses on the development of the poverty line in East Kalimantan Province and how Holt's Double Exponential Smoothing method can be used to accurately forecast the poverty line. The objectives of this study are to analyze poverty line trends in East Kalimantan Province, apply Holt's DES method in forecasting, and evaluate the accuracy of the resulting forecasts. The results of this study are expected to contribute to the local government in providing accurate information to support decision-making and the formulation of strategic policies for poverty alleviation in East Kalimantan Province.

## II. THEORY

Building on these findings, this study extends the application of Holt's Double Exponential Smoothing method specifically to East Kalimantan, aiming to evaluate its effectiveness in forecasting the poverty line within a regional context. Unlike ARIMA, which requires data stationarity and a larger sample size, Holt's DES is more suitable for short-term forecasting with a clear linear trend, making it appropriate for the poverty line data in East Kalimantan.

### *A. Poverty Line*

The poverty line is used as a boundary to classify people as poor or not poor and can be used as a consideration in socio-economic reforms, such as welfare improvement programs and unemployment insurance [11]. People who have an average monthly per capita expenditure below the poverty line are included in the poor [12].

### *B. Forecasting*

Forecasting is an art and knowledge to predict future events in the present [13]. Forecasting is a calculation analysis technique carried out with qualitative and quantitative approaches to estimate future events using historical data references [14]. Forecasting is an important tool in the planning and decision-making process in the future. Forecasting in this case is not always accurate, because the accuracy of the forecasting process depends on the data obtained and the methods used [15]. Forecasting has the aim of reducing errors in the forecasting process, which is generally measured using Mean Square Error, Mean Aboslute Error, and others [16].

### *C. Double Exponential Smoothing*

Holt's DES is a method introduced by CC. Holt in 1958, used on data that has a trend pattern. Holt's DES method is a trend smoothing method that uses different parameters from the parameters used in the original data. This method uses two smoothing parameters, α (Level) and β (Trend) with values between 0 and 1 ($0 < \alpha < 1$). Holt's DES method is used to generate a new trend in the data to remove irrelevant components from the raw data, such as data subject to random fluctuations and estimate more accurate results [17]. The calculation to find the parameter α (Level) is given as equation (1) [18]:

$$S_t = \alpha X_t + (1 - \alpha)(S_{(t-1)} + T_{(t-1)}) \quad (1)$$

The calculation to find the parameter β (Trend) is given as equation (2):

$$T_t = \beta(S_t + S_{(t-1)} + (1 - \beta)T_{(t-1)}) \quad (2)$$

For the calculation to find the fitted value is given as equation (3):

$$F_t = S_{(t-1)} + T_{(t-1)} \quad (3)$$

To perform m forecasting results is given as equation (4):

$$F_{(t+m)} = S_t + T_t(m) \quad (4)$$

One way to initialize the Holt method is to set the value $S_1 = X_1$ and is given as equation (5):

$$T_1 = \frac{((X_2 - X_1) + (X_3 + X_2))}{2} \quad (5)$$

Description:

$S_t$ = Level value at period t
$\alpha$ = Smoothing parameter value between 0 and 1 ($0<\alpha<1$)
$\beta$ = Second parameter value for tren smoothing
$X_t$ = Actual value at period t
$T_t$ = Trend value at period t
$F_t$ = Fitted value at period t
$m$ = Forecasting period to be forecast
$t$ = Time (1, 2, 3...)
$F_{t+m}$ = Forecasting value at period t + m for m = 1, 2, 3, ...

### *D. Mean Absolute Percentage Error*

In forecasting, forecasting testing is needed to determine the smallest error rate in forecasting which is commonly referred to as forecast error [19]. There are many methods that can be used, but not all methods will be suitable for the case and data used. Therefore, a method of measuring forecasting accuracy is needed to evaluate the analysis model used. In this study, the metric used is Mean Absolute Percentage Error (MAPE). Mean Absolute Percentage Error (MAPE) is calculated by dividing the absolute error of each period by the actual observed value for the current period. MAPE is used to produce the smallest error compared to other methods. In addition, MAPE is also used to determine the percentage of forecast error in the calculated forecasting results. The MAPE value can be calculated using equation (7) [20]:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n} \left| \frac{X_t - F_t}{X_t} \right| \quad (7)$$

Description:

$n$ = Number of data
$X_t$ = Actual data in period t
$F_t$ = Forecasted data in period t
$t$ = Time (1, 2, 3,...)

Referring to the accuracy value based on the MAPE equation, forecasting results are said to be good if they meet one of the following criteria:

TABLE 1 MAPE CATEGORY

| MAPE Value | Category |
|---|---|
| <10% | Very Good |
| 10% - 20% | Good |
| 20% - 50% | Enough |
| >50% | Not Enough |

## III. METHOD

In general, the stages in the research method used in conducting research are organized systematically to achieve the desired result. The stages of this research can be seen as shown in Fig. 1.

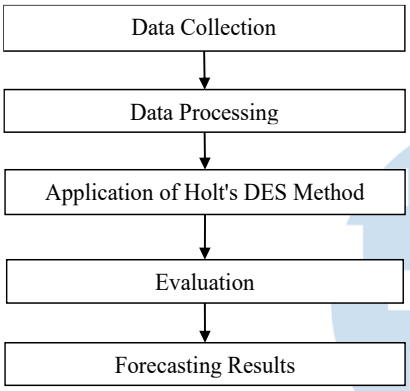



Fig. 1 Methodology

### A. Data Collection

This study uses secondary data in the form of poverty line data for East Kalimantan Province obtained from official publications by the East Kalimantan Provincial Statistics Agency (BPS). The data used is semester data (twice a year), starting from Semester I of 2011 to Semester II of 2024, resulting in a total of 28 observations. Data searches are also carried out on the internet through the official website (https://kaltim.bps.go.id/id/query-builder) of the Badan Pusat Statistik of East Kalimantan Province (BPS) to complement it. The form of research data is semi-annual data with quantitative data types and based on the source is secondary data.

The data was taken directly from the publication Statistics on People's Welfare and Poverty Profile in Indonesia by Province published by BPS without any transformation or smoothing (raw data). Before being used in the analysis, the data was checked to ensure that there were no missing values, anomalies, or inconsistencies between periods. This check was carried out to ensure the reliability and validity of the data so that the forecasting results obtained could represent the actual conditions of the poverty line in East Kalimantan Province.

Semester data was selected to capture the relatively rapid dynamics of changes in the poverty line compared to annual data, while still being stable enough to be analyzed using time series forecasting methods.

### B. Data Processing

With the collection of data needed in the research, the next stage is data processing to get accurate forecasting analysis results. The data will be processed using Holt's DES method with the help of Google Colab tools. The goal is for researchers to know the forecasting of the poverty line for the next ten periods, namely from the first semester (September) 2024 to the second semester (March) 2029. In addition to forecasting for the next few periods, the analysis is also carried out to find out whether the data has increased or decreased or fluctuated during that period. Then, the results of the analysis that has been carried out will be compared with historical data obtained from BPS East Kalimantan Province.

### C. Application of Holt's DES Method

The analytical method applied in this study to forecast the poverty line of East Kalimantan Province is the Holt's Double Exponential Smoothing (DES) method. This method is particularly appropriate for data that exhibit a linear trend pattern, as it incorporates both the level and trend components in its forecasting process. The linear trend can be represented through a straight-line equation derived from the scatter diagram of the observed data over a given period.

Holt's DES method is widely recognized for its high forecasting accuracy across short-, medium-, and long-term horizons. In general, it is capable of predicting up to 4 periods (1–2 years), 6–10 periods (3–5 years), and even 20 periods (around 10 years) ahead with reliable results. Considering that the poverty line data in East Kalimantan Province displays a consistent upward trend over time, this method is deemed suitable for the present study.

To ensure optimal model performance, the selection of the smoothing parameters (α and β) was conducted through an optimization process based on minimizing forecasting errors. A grid search approach was implemented by testing combinations of α and β values ranging from 0.1 to 0.9 with 0.01 intervals. The parameter pair yielding the lowest Mean Absolute Percentage Error (MAPE) value was chosen as the optimal configuration for the model.

This optimization process aims to minimize forecasting errors and enhance the model's accuracy in capturing the trend behavior of the poverty line in East Kalimantan. All computations and model estimations were performed using Python software (Statsmodels), specifically utilizing the holt() function from the statsmodels.tsa.holtwinters library. The final model, calibrated with optimal parameters, was subsequently used to generate fitted values, evaluate model performance through error metrics (MAPE), and

forecast the poverty line for the next 10 periods, extending up to Semester II of 2029.

### D. Evaluation

Forecasting accuracy testing is carried out with the aim of knowing the level of accuracy of the method in forecasting data. Thus, the results of this forecasting accuracy test are expected to show that the model used is accurate and valid. Testing the forecasting accuracy of this research will use the MAPE method. This method is used to test forecasting accuracy because it has advantages. The advantage of this method is that it can provide information about the level of error size in forecasting which can make it easier if the forecasting error size level is too low or too high, so that the results obtained are more accurate.

### E. Forecasting Results

At this stage, the forecasting results are obtained by forecasting the value of the poverty line of East Kalimantan Province for the next 10 periods using the optimal Holt's DES method. The results of the forecasting analysis will be made a data visualization in the form of a dashboard in the form of a graph of the results of forecasting the poverty line of East Kalimantan Province from semester I (September) 2024 to semester II (March) 2029 using the Looker Studio application. Data visualization aims to be able to see patterns and trends in forecasting results against the results of accuracy tests using the MAPE metric to compare with historical data.

## IV. Result and Discussions

### A. Poverty Line Data

Based on the results of the research that has been conducted, several results can be summarized into several sections, starting with the data collection and processing stages, the application of the methods used to the results in the form of forecasting. The first stage is the collection of data on the poverty line of East Kalimantan Province for the last 13 years (2011-2024) from the official website of the Central Bureau of Statistics (BPS) of East Kalimantan Province. Table 2 shows the poverty line data of East Kalimantan Province in tabular form.

Table 2 East Kalimantan Poverty Line Data

| Year | Semester I | Semester II |
|---|---|---|
| | March | September |
| 2011 | 316.819 | 336.019 |
| 2012 | 347.577 | 363.887 |
| 2013 | 381.706 | 417.902 |
| 2014 | 431.560 | 444.248 |
| 2015 | 473.710 | 494.207 |
| 2016 | 511.205 | 526.686 |
| 2017 | 548.094 | 561.868 |
| 2018 | 574.704 | 598.200 |
| 2019 | 609.155 | 638.690 |
| 2020 | 662.302 | 669.622 |
| 2021 | 689.035 | 703.223 |
| 2022 | 728.208 | 768.120 |
| 2023 | 790.186 | 833.955 |
| 2024 | 833.955 | 853.997 |

Based on Table 3, the poverty line data of East Kalimantan from the first semester (March) 2011 to the second semester (September) 2024 is visualized in the form of a graph as shown in Figure 2.

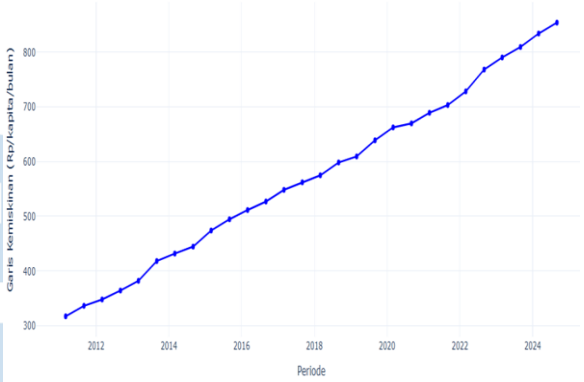



Fig. 2 Plot Data on Poverty Line in East Kalimantan Province 2011-2024

Fig. 2 shows a consistent upward trend from year to year. The poverty line in East Kalimantan has increased significantly from IDR 316,819 in the first semester of 2011 to IDR 853,997 in the second semester of 2024. This increase indicates a change in the minimum cost of living required by the population to fulfill basic needs over the past decade. Visualization not only shows the growth of the poverty line value, but also shows a stable trend pattern, which then becomes the basis for applying the forecasting method. The application of Holt's DES forecasting method is used to predict the value of the poverty line for the next 10 periods, so that the results of this study can provide an overview for future social and economic policy planning.

### B. Application of Holt's DES Method

The forecasting method used for forecasting East Kalimantan poverty line data for the next 10 periods in this study is Holt's Double Exponential Smoothing (DES) method. The DES Holt's method was chosen because based on previous research, the DES Holt's method is an easy forecasting method and has proven to be effective and accurate so that it can be used in forecasting data that has a trend pattern. Forecasting using Holt's DES method consists of several sequential steps with the following explanation.

### C. Input Data

In this study, the data taken for analysis is secondary data, namely the poverty line data for the province of East Kalimantan from the first semester (March) 2011 to the second semester (September) 2024. The following is the poverty line data for the province of East Kalimantan during this period, which is presented in Table 3.

TABLE 3 EAST KALIMANTAN POVERTY LINE

| Period | Poverty Line (IDR/Capita/Month) |
|---|---|
| 2011-03 | 316.819 |
| 2011-09 | 336.019 |
| 2012-03 | 347.577 |
| 2012-09 | 363.887 |
| 2013-03 | 381.706 |
| … | … |
| 2022-09 | 768.120 |
| 2023-03 | 790.186 |
| 2023-09 | 809.418 |
| 2024-03 | 833.955 |
| 2024-09 | 853.997 |

Based on the data in Table 3, the poverty line of East Kalimantan Province during the period, in the forecasting process using Holt's DES method, the data normalization process is not carried out. This is because the poverty line data is already in the original units, namely rupiah per capita per month, so it does not require rescaling. Holt's DES method uses the original values in the time series data without requiring a transformation or scale adjustment process. The normalization process is usually required in certain statistical methods or machine learning algorithms that are sensitive to scale differences between variables, such as Linear Regression, Clustering, and Classification. However, in univariate time series forecasting which only focuses on the patterns, trends, and behavior of one type of data over time, the authenticity of the data values is very important so that the forecasting results match the real conditions and are easily understood in an economic context.

### D. Descriptive Statistics

Based on the secondary data that has been collected; to obtain an overview of the poverty line data of East Kalimantan Province for the first semester (March) 2011 to the second semester (September) 2024, descriptive statistical analysis was conducted with the help of Google Colab tools. The following is a descriptive statistical display of the poverty line data of East Kalimantan Province during the period presented in Table 4.

TABLE 4 DESCRIPTIVE STATISTIC

| | |
|---|---|
| ***Maximum*** | 316.819 |
| **Q1** | 441.076 |
| ***Median*** | 568.286 |
| ***Mean*** | 574.439 |
| **Q3** | 692.582 |
| ***Maximum*** | 853.997 |

Based on Table 4.3, the poverty line data for East Kalimantan Province from the first semester (March) 2011 to the second semester (September) 2024 shows that the lowest value occurred at the beginning of the period, in March 2011 at IDR 316,819 per capita per month. Meanwhile, the highest value occurred at the end of the period, namely September 2024 at IDR 853,997 per capita per month. This indicates a significant increase in the cost of living from 2011 to 2024. The median value of IDR 568,286 per capita per month explains that half of the poverty line data falls below this value, which occurred around 2017 to 2018. The mean (average) poverty line value of IDR 574,439 per capita per month is slightly higher than the median. This can be explained by the fact that the poverty line value over the last 13 years has increased quite a lot, indicating a slight skewness to the right. The value of quartile 1 (Q1) is the lower quartile which means the middle value between the smallest value and the median of the data and obtained a value of IDR 441,076 per capita per month while quartile 3 (Q3) which means the middle value between the median and the highest value of the data and obtained a value of IDR 629,582 per capita per month. The Q1 and Q3 values show that 50% of the poverty line values are within the range of IDR 411,076 - IDR 692,582 per capita per month, which covers the middle period, namely 2014 - 2021.

### E. Identification Data

From Figure 3, the poverty line data for East Kalimantan from Semester I (March) 2011 to Semester II (September) 2024 tends to increase every year.

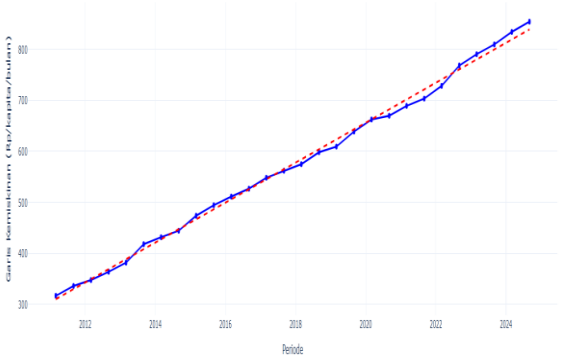


Fig. 3 Trend Pattern Plot of Poverty Line Data of East Kalimantan Province

Based on the pattern of poverty line data in East Kalimantan Province from Semester I (March) 2011 to Semester II (September) 2024, which shows an upward trend pattern in each semester, the DES Holt's forecasting method can be used to forecast the poverty line for the next ten periods. This method can adjust the linear trend pattern found in the data, so the forecasting results are expected to accurately represent the development of the poverty line.

*F. Optimal Parameters*

Research on poverty line data of East Kalimantan Province Semester I (March) to Semester II (September) 2024 was analyzed using Holt's DES method. The analysis is carried out by smoothing twice, namely smoothing the level and trend values. Before smoothing, the values of α and β parameters for level and trend smoothing must be determined first. Alpha and beta parameters are determined through trial-and-error approach. The following are the results of the α and β parameters presented in Table 5.

TABLE 5 PARAMETER VALUE

| α (Level) | β (Tren) |
|---|---|
| 0,98 | 0,01 |

Based on Table 5, the optimal alpha value is 0.98 and the optimal beta value is 0.01. Calculations are carried out systematically using the grid search method, which tries various combinations of α and β values in the range of 0.1 to 0.99 with multiples of 0.01 to obtain the most optimal combination of parameters in producing accurate forecasts. Each combination is tested by comparing the resulting forecasting results against the actual data, using the smallest error measure, MAPE. The combination of parameters α and β that produce the smallest error value is selected as the optimal value and then used in the final model to forecast the poverty line.

*G. Fitted Value*

Fitted values are obtained through the process of calculating level and trend values. Level and trend start from the second year, so the first forecasting results will be obtained starting from the third year. Analysis in the first row, the level and trend values are empty because there is no previous data as a reference. Therefore, the calculation starts from the second row, where the level value is assumed to be the same as the poverty line value in that period, while the trend value is calculated based on the difference between the level value and the poverty line in the previous period. The fitted value is obtained from the sum of the level and trend values, and the result is placed in the third year. The fitted value is the predicted value of the conjecture obtained from the actual data. The fitted value is used to assist in forecasting for the next few periods.

TABLE 6 FIITED VALUE

| Year/Semester | Actual Data | Level | Trend | Fitted Value |
|---|---|---|---|---|
| 2011-03 | 316.819 | 316.819 | 19.200 | - |
| 2011-09 | 336.019 | 336.019 | 19.200 | - |
| 2012-03 | 347.577 | 347.730 | 19.125 | 355.219 |
| 2012-09 | 363.887 | 363.946 | 19.096 | 366.855 |
| 2013-03 | 381.706 | 381.733 | 19.083 | 383.042 |
| 2013-09 | 417.902 | 417.560 | 19.250 | 400.816 |
| … | … | … | … | … |
| 2022-09 | 768.120 | 767.701 | 19.304 | 747.190 |
| 2023-03 | 790.186 | 790.122 | 19.335 | 787.005 |
| 2023-09 | 809.418 | 809.419 | 19.334 | 809.457 |
| 2024-03 | 833.955 | 833.851 | 19.385 | 828.753 |
| 2024-09 | 853.997 | 853.982 | 19.393 | 853.236 |

Table 6 shows the predicted value of the actual data. The fitted value starts from 2012 (March) because according to the formula, the first two rows are not included in the forecasting calculation so that the poverty line forecast in East Kalimantan Province in 2012 (March) is IDR 355,219 per capita per month, 2012 (September) is IDR 366,855 per capita per month, 2013 (March) is IDR 383,042, and so on until 2024 (September) is IDR 853,236 per capita per month.

*H. Forecasting Accuracy Testing*

Forecasting accuracy testing in this study was carried out using the MAPE metric. Calculation of the accuracy test using the MAPE metric with the optimal α parameter of 0.98 and the optimal β of 0.01, the MAPE value of 4.56% is obtained, indicating that the MAPE value is less than 10% based on the MAPE Category Table, the smaller the MAPE value, the more accurate a model is, so it can be said that the forecasting model using the Holt's DES method has a very good level of forecasting accuracy. From the accurate test results, the metric value produces a relatively small error size value so that the forecasting model using the Holt's DES method used is stable, neither overfitting nor underfitting, and very accurate in forecasting poverty line data.

*I. Forecasting Results*

The calculation of forecasting data for the poverty line of East Kalimantan Province for the next ten periods using Holt's DES method with optimal parameters α of 0.98 and β of 0.01 using equation 4 can be seen in Table 7 as follows.

TABLE 7 EAST KALIMANTAN POVERTY LINE FORECASTING RESULTS

| Period | Year/Semester | Result of Forecasting (IDR/Capita/Month) |
|---|---|---|
| 1 | 2025-03 | 873.374 |
| 2 | 2025-09 | 892.767 |
| 3 | 2026-03 | 892.767 |
| 4 | 2026-09 | 912.160 |
| 5 | 2027-03 | 931.553 |
| 6 | 2027-09 | 950.945 |
| 7 | 2028-03 | 989.731 |
| 8 | 2028-09 | 1009.124 |
| 9 | 2029-03 | 1028.517 |

Based on Table 7, the results of forecasting the poverty line data of East Kalimantan Province from the first semester (March) 2025 to the second semester (September) 2029 using Holt's DES method with optimal parameters show that the value of the poverty line during this period has increased every semester. In 2025 (September), the poverty line value increased to IDR 892,767 compared to IDR 873,374 in 2025

(March) and continued to increase until it reached IDR 1,047,910 in 2029 (September).

To present the poverty line data of East Kalimantan Province in a complete and interactive manner, a visualization dashboard was created. The aim is to facilitate the analysis of poverty line trends from 2011 to 2029. The dashboard displays three main types of data; namely actual data obtained from the website of BPS East Kalimantan Province in the form of secondary data every semester (Semester I March 2011-Semester II September 2024). In the main graph, the actual data is displayed in the form of a blue line that illustrates the growth of the poverty line, while the forecasting results are visualized with a red dotted line that shows an upward trend in the future.


Fig. 4 East Kalimantan Poverty Line Forecasting Dashboard

Figure 4 illustrates the results of the Double Exponential Smoothing Holt's method applied to the poverty line data of East Kalimantan Province from Semester I of 2011 to Semester II of 2029. The blue line represents the actual data obtained from BPS, while the red line indicates the fitted and forecasted values generated by the model. From the figure, it can be observed that the fitted values closely follow the actual data trend, suggesting that the Holt's DES model effectively captures the linear upward trend of the poverty line.

The forecasting line shows a consistent increase over the next 10 periods, implying that the cost of living and basic needs in East Kalimantan will continue to rise gradually. This trend aligns with the socioeconomic dynamics in the region, including urban development, changes in consumption patterns, and inflationary pressures on essential goods. The relatively small forecasting error values (as indicated by MAPE) demonstrate the reliability of the model in projecting future poverty line movements.

Furthermore, this visualization strengthens the argument that Holt's DES is suitable for datasets exhibiting a linear trend without seasonal components. Compared to regression-based or ARIMA approaches, Holt's DES provides a more adaptive mechanism in responding to short-term fluctuations while maintaining long-term trend accuracy. These findings support the conclusion that the model can serve as a robust analytical tool for policymakers to anticipate socioeconomic challenges and design targeted poverty reduction strategies in East Kalimantan Province.

## V. CONCLUSION

Based on the forecasting of the poverty line of East Kalimantan Province using Holt's Double Exponential Smoothing (DES) method for the next ten periods, namely from Semester I (March) 2025 to Semester II (September) 2029, it can be concluded that this method is effective in modeling and forecasting the poverty line based on semesterly data from 2011-2024. The method is able to capture a trend pattern that increases consistently over time with optimal parameters alpha ($\alpha$) = 0.98 and beta ($\beta$) = 0.01, which reflects the model's ability to adjust the current level value with a fast response while still maintaining the stability of the long-term trend direction. The forecasting results show that the poverty line is predicted to increase by 19.98%, from IDR 873,374 in the first semester of 2025 to IDR 1,047,910 in the second semester of 2029, reflecting the increasing trend of minimum living needs from year to year. Evaluation of the model's accuracy resulted in a MAPE value of 4.56%, which is below 10%, indicating that the Holt's DES model has a very good level of forecasting accuracy. Therefore, the results of this study can be used as recommendations in policy making related to the poverty line in East Kalimantan Province as well as a reference for future research.

The evaluation results show that the model achieves a MAPE value of 4.56%, indicating a high level of forecasting accuracy (below the 10% threshold). This demonstrates that Holt's DES method provides a reliable tool for analyzing and predicting poverty line dynamics in East Kalimantan. Furthermore, this study highlights that accurate forecasting of the poverty line can serve as a strategic foundation for formulating evidence-based social and economic policies. By anticipating future trends, policymakers can design more effective poverty alleviation programs, ensure better resource allocation, and strengthen regional economic resilience. Therefore, the findings of this research not only contribute methodologically but also provide valuable insights for sustainable development planning in East Kalimantan Province.

## REFERENCES

[1] A. I. La Murdani, I. Gainau, and U. Resiloy, "Forecasting the Poverty Line in South Buru Regency Using Holt's Double Exponential Smoothing Method," in *Konferensi Nasional Matematika XX Peranan Ilmu Matematika dalam Menjawab Tantangan Bangsa yang Semakin Kompleks dan Dinamis di Era Revolusi Industri 4.0*,

Pattimura Proceeding: Conference of Science and Technology, Jul. 2021, pp. 495–501. doi: 10.30598/PattimuraSci.2021.KNMXX.

[2] BPS Provinsi Kalimantan Timur, *East Kalimantan Province Economic Report 2023*, vol. 14. Badan Pusat Statistik Provinsi Kalimantan Timur, 2023.

[3] BPS Provinsi Kalimantan Timur, *East Kalimantan Province Economic Report 2024*, vol. 30. Badan Pusat Statistik Provinsi Kalimantan Timur, 2024.

[4] A. Ahmed Shokeralla, F. EL Guma, A. Gibreel Mohammed Musa, F. ELrhman Elsmih, and A. A. shokeralla, "Prediction The Daily Number of Confirmed Cases of Covid-19 in Sudan with ARIMA and Holt Winter Exponential Smoothing," *Article in International Journal of Development Research*, vol. 10, no. 8, Aug. 2020, doi: 10.37118/ijdr.19811.08.2020.

[5] A. Adryan, S. S. Sururin, W. S. Akbar, and E. Widodo, "Forecasting the Poverty Line in the Special Region of Yogyakarta Using the Double Exponential Smoothing Method," *Jurnal Ilmiah Pendidikan Matematika, Matematika dan Statistik*, vol. 3, no. 2, pp. 338–343, Aug. 2022, doi: 10.46306/lb.v3i2.

[6] A. Zahrunnisa, R. D. Nafalana, I. A. Rosyada, and E. Widodo, "Comparison of Exponential Smoothing and ARIMA Methods in Forecasting the Poverty Line in Central Java Province," *Jurnal Ilmiah Pendidikan Matematika, Matematika dan Statistika*, vol. 2, no. 3, pp. 300–313, Dec. 2021, doi: 10.46306/lb.v2i3

[7] X. Xian *et al.*, "Comparison of SARIMA model, Holt-winters model and ETS model in predicting the incidence of foodborne disease," *BMC Infect Dis*, vol. 23, no. 1, pp. 1–9, Dec. 2023, doi: 10.1186/S12879-023-08799-4/FIGURES/2.

[8] A. Bhattacharjee, G. K. Vishwakarma, N. Gajare, and N. Singh, "Time Series Analysis Using Different Forecast Methods and Case Fatality Rate for Covid-19 Pandemic," *Regional Science Policy & Practice*, vol. 15, no. 3, pp. 506–520, Apr. 2023, doi: 10.1111/RSP3.12555.

[9] S. Jain, S. Agrawal, E. Mohapatra, and K. Srinivasan, "A Novel Ensemble ARIMA-LSTM Approach for Evaluating COVID-19 Cases and Future Outbreak Preparedness," *Health Care Science*, vol. 3, no. 6, p. 409, Dec. 2024, doi: 10.1002/HCS2.123.

[10] A. Ajiono and T. Hariguna, "Comparison of Three Time Series Forecasting Methods on Linear Regression, Exponential Smoothing and Weighted Moving Average," *International Journal of Informatics and Information Systems*, vol. 6, no. 2, pp. 89–102, Mar. 2023, doi: 10.47738/IJIIS.V6I2.165.

[11] D. Anggraeni, S. Maryani, and S. Ariadhy, "Poverty Line Forecasting in Purbalingga Regency for 2021-2023 Using Brown's Single-Parameter Linear Double Exponential Smoothing Method," *Jurnal Ilmiah Matematika dan Pendidikan Matematika (JMP)*, vol. 13, no. 2, p. 155166, Dec. 2021, doi: 10.20884/1.jmp.2021.13.2.4548

[12] BPS Provinsi Kalimantan Timur, *East Kalimantan Province Sustainable Development Goals Indicators 2023*, vol. 1. Badan Pusat Statistik Provinsi Kalimantan Timur, 2024.

[13] D. Yanti, M. Sukarma, M. P. Anggraini, and O. Foureta, "Forecasting Poverty Levels Using the Triple Exponential Smoothing Method in Pekanbaru City," *Indonesian Council of Premier Statistical Science*, vol. 2, no. 1, p. 32, Sep. 2023, doi: 10.24014/icopss.v2i1.25327.

[14] H. S. Pakpahan, Y. Basani, and R. R. Hariani, "Prediction of the Number of Poor People in East Kalimantan Using Single and Double Exponential Smoothing," *Informatika Mulawarman: Jurnal Ilmiah Ilmu Komputer*, vol. 15, no. 1, pp. 47–51, Feb. 2020, doi: 10.30872/jim.v15i1.3180.

[15] R. I. Prasetyono and D. Anggraini, "Analisis Peramalan Tingkat Kemiskinan di Indonesia dengan Model ARIMA," *Jurnal Ilmiah Informatika Komputer*, vol. 26, no. 2, pp. 95–110, Aug. 2021, doi: 10.35760/ik.2021.v26i2.3699.

[16] N. R. Yuwono and S. Yulianto, "Comparison of Various Exponential Smoothing Methods for COVID Forecasting in Indonesia," *Jurnal Penerapan Teknologi Informasi dan Komunikasi*, vol. 1, no. 2, pp. 155–165, Jun. 2022, doi: 10.24246/itexplore.v1i2.2022.pp155-165.

[17] M. Olivia and A. Amelia, "Exponential Smoothing Method for Forecasting the Number of Poor People in Langsa City," *Gamma-Pi: Jurnal Matematika dan Terapan*, vol. 3, no. 1, pp. 47–51, Jun. 2021, Accessed: Sep. 09, 2024. [Online]. Available: https://ejurnalunsam.id/index.php/jgp/article/view/3771/2724

[18] S. Yurinanda, S. Sarmada, S. Rozi, A. Tasya, and D. A. Fajrin, "Predicting Library Visitor Numbers Using Double Exponential Smoothing," *Prismatika: Jurnal Pendidikan dan Riset Matematika*, vol. 6, no. 2, pp. 375–382, Apr. 2024, doi: 10.33503/prismatika.v6i2.3784.

[19] S. I. Murpratiwi, D. A. I. C. Dewi, and A. Aranta, "Comparative Analysis of Transaction Data Forecasting Results for Service Companies Using the Time Series Method," *Jurnal Teknologi Informasi dan Komputer*, vol. 7, no. 2, pp. 218–227, Jan. 2021, doi: 10.36002/jutik.v7i2.1323.

[20] D. R. P. Sari, "Application of the Double Exponential Smoothing Method to Monthly Inflation Data for 2021," *Jurnal Matematika dan Statistika serta Aplikasinya*, vol. 10, no. 2, pp. 26–31, Jul. 2022, doi: 10.24252/msa.v10i2.27272

ISSN 2085-4552

# Application of the Dempster-Shafer Method in Developing a Web-Based Expert System for Diagnosing Dental and Oral Diseases

Yoseph P.K Kelen
Department of Information Technology, Faculty of Agriculture, Universitas Timor, Kefamenanu, Indonesia
yosepkelen@unimor.ac.id

Accepted 19 December 2025
Approved 07 January 2026

***Abstract*— An expert system is a problem-solving system that captures human knowledge into a system similar to what experts typically do. Expert systems allow humans to solve problems usually handled by specialists, even when performed by non-experts. They can also assist experts in completing their tasks or serve as reliable assistants in addressing problems without requiring direct consultation with professionals. This study aims to develop an expert system for diagnosing dental diseases using the Dempster–Shafer method to identify dental and oral conditions based on the highest probability of symptoms experienced. The expert system is implemented on a computer using a web-based programming language with PHP and a MySQL database. The results of this expert system are expected to make it easier for users to conduct consultations and obtain accurate diagnoses of dental diseases they experience, as well as receive appropriate treatment recommendations using the Dempster–Shafer algorithm.**

***Index Terms*— *Expert System; Dempster-Shafer; Dental Disease; Oral Disease.***

## I. INTRODUCTION

One of the branches of computer science widely utilized by humans to assist their work is the expert system, which is a subfield of artificial intelligence [1]. Artificial Intelligence (AI) is a field of computer science that plays an important role in the present and future eras. AI encompasses a broad range of areas, from general to highly specialized domains. From learning to perception, AI is considered a universal discipline [2].

The concept of expert systems is based on the assumption that expert knowledge can be stored and applied in a computer, then utilized by others when needed. One of the applications of expert systems is in the field of medicine or healthcare. The implementation of expert systems in healthcare may include disease diagnosis, health maintenance consultations, and the provision of recommendations based on existing diagnoses. Health is indeed a valuable aspect of human life; therefore, personal awareness is needed to maintain it. One of the body organs that is often neglected is the teeth and mouth. This is evidenced by data from the Directorate General of Health Services (2001), which shows that dental and oral diseases are among the ten most prevalent diseases in Indonesia [3].

Based on a survey conducted by the Dental Health Foundation (2003) on children, 70% were found to suffer from dental caries and gingivitis, while among adults, 73% were reported to have dental caries. Furthermore, according to Indonesia's Basic Health Research (Riskesdas) in 2013, 25.9% of the Indonesian population has dental and oral health problems. The lack of knowledge and limited sources of information regarding dental and oral health contribute to the low public awareness of maintaining oral hygiene. Among those affected, 31.1% received care from dental health professionals—such as dental nurses, dentists, or dental specialists—while the remaining 68.9% did not seek any treatment [4].

The ideal ratio of dentists to the population in Indonesia is 1 to 9,000. However, due to the still limited number of dentists in the country, this ratio has increased to 1 to 24,000. This ideal ratio is far from the standard set by the World Health Organization (WHO), which is 1 to 2,000 people. This concerning situation is further worsened by the unequal distribution of dentists in Indonesia, with 70% of them concentrated on the island of Java. The conditions described above highlight the need for a system that can serve as an initial consultation platform before seeking further treatment from a dentist. Therefore, the author is motivated to develop an "Expert System for Diagnosing Dental and Oral Diseases." This system is intended to assist dental practitioners by improving the speed and accuracy of diagnosing diseases and providing appropriate solutions.

## II. METHOD

### 2.1 Teeth

Teeth are the hardest tissues in the human body. Their structure consists of multiple layers, starting from the extremely hard enamel, followed by dentin, and the pulp, which contains blood vessels, nerves, and other components that strengthen the tooth. However, teeth are also among the body tissues most susceptible to damage. This occurs when the teeth do not receive proper care [5]

Teeth are the hard parts located within the mouth that function to chew food. Each tooth consists of several parts, as illustrated in the figure below:

### 2.2 Types of Dental Diseases

Based on the *Dental Disease Module* by Kristiani et al. (2010), several types of dental and oral diseases are described as follows:

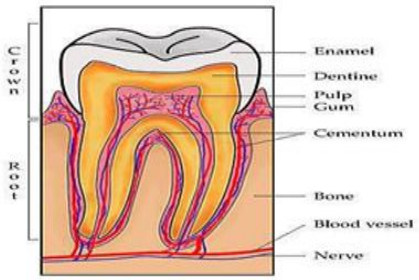
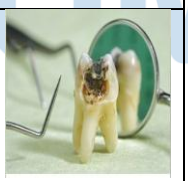



Figure 2.2 Structure of the Tooth

**Table 2.2. Types of Dental and Oral Diseases**

| No | Disease Name | Description | Illustration |
|---|---|---|---|
| 1 | **Dental Caries** | Dental caries is an infectious disease that damages the hard structure of the teeth. It is characterized by cavities. |  |
| 2 | **Dental Abscess** | A dental abscess is a pus-filled swelling or lump that forms in the tooth due to bacterial infection. |  |
| 3 | **Gingivitis** | Gingivitis is a bacterial infection that causes the gums to become inflamed, red, and swollen. |  |
| 4 | **Pulpitis** | Pulpitis is a condition caused by inflammation of the pulp—the central part of the tooth containing tissues and tooth-forming cells. |  |
| 5 | **Periodontitis** | Periodontitis is a gum infection that damages the soft tissue and bone supporting the teeth. This condition requires prompt treatment as it may lead to tooth loss. |  |
| 6 | **Pericoronitis** | Pericoronitis is an inflammation of the gum tissue surrounding the wisdom teeth, which are the last molars to emerge and are located at the back of the mouth. |  |

### 2.3 Expert System

An expert system is a computer system designed to replicate the abilities or expertise of a human specialist in a specific domain. It is implemented as a computer program presented in a way that non-expert users can easily understand and utilize. This enables users without expert knowledge to make decisions or draw conclusions similar to those made by experts [6].

The reasoning ability of an expert system depends on the knowledge base encoded within the system. The effectiveness of such a system in solving problems is directly related to the amount and quality of expert knowledge stored. The larger and more accurate the knowledge base, the more capable the expert system becomes [7].

The user interface serves as the communication bridge between the user and the system. It allows users to input data and receive information in return. Expert systems differ from conventional systems primarily in their knowledge-based foundation. According to Adriani (2016:13–14), the distinguishing characteristic of an expert system lies in how it utilizes stored expert knowledge to perform reasoning and provide conclusions based on that knowledge.

### 2.4 Dempster-Shafer Method

The Dempster-Shafer theory is a mathematical framework for reasoning under uncertainty, based on *belief functions* and *plausible reasoning*. It is used to combine separate pieces of evidence to calculate the probability of an event [8]. The theory was developed

by Arthur P. Dempster and Glenn Shafer. Generally, the Dempster-Shafer theory is represented within an interval:
Belief,PlausibilityBelief,
PlausibilityBelief,Plausibility

Belief(Bel)
Belief represents the strength of evidence supporting a particular hypothesis. A value of 0 indicates no evidence, while a value of 1 represents complete certainty [9].

Plausibility(Pl)
Plausibility measures how much the evidence does *not* refute a hypothesis and is calculated as: Pl(s)=1−Bel(¬s)Pl(s) = 1 - Bel(¬s)Pl(s)=1−Bel(¬s) Plausibility values range from 0 to 1. If we are certain of ¬s (not s), then Bel(¬s) = 1 and Pl(¬s) = 0. Thus, plausibility reduces the degree of belief in the evidence.

In Dempster-Shafer theory, there exists a **frame of discernment**, denoted as θ, which represents the universal set of all possible hypotheses, and a **mass function (m)**, which denotes the degree of belief assigned to a subset of hypotheses[10].

Let:
θ = {A, B, C, D, E, F, G}

Where:
A=DentalCaries
B=Pulpitis
C=Gingivitis
D=DentalAbscess
E=Malocclusion
F=CrowdedTeeth
G = Impaction

MassFunction(m) The mass function in Dempster-Shafer theory quantifies the degree of belief in a given piece of evidence. To combine multiple pieces of evidence, the Dempster's Rule of Combination is applied: If $m_1(X)$ and $m_2(Y)$ are two mass functions representing different evidences, the combined mass function $m_3(Z)$ is calculated by applying the rule to aggregate the evidences and compute the resulting belief.

## III. Result and Discussions

### *3.1 Implementation of the Dempster-Shafer Method*

*1. Dental and Oral Disease Data*

Dental and oral disease data is data in the system presented in table 3.1.

**Table 3.1. Dental and Oral Disease Data**

| No | Code | Disease |
|---|---|---|
| 1 | P1 | Pulpitis |
| 2 | P2 | Gingivitis |
| 3 | P3 | Dental Abscess |
| 4 | P4 | Dental Caries |
| 5 | P5 | Periodontitis |
| 6 | P6 | Pericoronitis |

*2. Symptoms of Dental and Oral Diseases*

Data on symptoms of dental and oral disease are symptoms that can cause dental and oral disease, which are presented in Table 3.2.

**Table 3.2. Symptoms of Dental and Oral Diseases**

| No | Code | Symptom |
|---|---|---|
| 1 | G1 | Presence of soft tissue growth in a decayed tooth |
| 2 | G2 | Tooth pain or sensitivity when chewing |
| 3 | G3 | Pain around the tooth |
| 4 | G4 | Swollen or easily bleeding gums |
| 5 | G5 | Unpleasant mouth odor (halitosis) |
| 6 | G6 | Lesions between the gums and teeth |
| 7 | G7 | Lump around the head, neck, or jaw |
| 8 | G8 | Fever |
| 9 | G9 | Pain while swallowing food |
| 10 | G10 | Swollen gums |
| 11 | G11 | Pain when opening the mouth |
| 12 | G12 | Swelling of lymph nodes in the neck |
| 13 | G13 | Pain when pressure is applied to the tooth by food |
| 14 | G14 | Pain in the gums and mouth |
| 15 | G15 | Swelling of the cheeks |
| 16 | G16 | Brown, black, or white spots on the tooth surface |
| 17 | G17 | Presence of holes or cavities in teeth |
| 18 | G18 | Toothache when exposed to cold water or when food enters |
| 19 | G19 | Gums feel soft when touched |
| 20 | G20 | Tooth appears taller than usual |
| 21 | G21 | Gaps between teeth become wider |
| 22 | G22 | Gums swollen and appear reddish or purplish |
| 23 | G23 | Pus discharge from infected gums |
| 24 | G24 | Limited and sometimes painful jaw movement |
| 25 | G25 | Pain or difficulty when swallowing |

### *3.2 Calculation Using the Dempster-Shafer Method*

To assume the level of certainty of an expert that has been adopted as the system's confidence level in providing a diagnosis of a data, this concept is then formulated into a production rule which is usually written in the form of If-Then (IF-THEN). This production rule can be said to be a two-part implication relationship, namely the premise (If) and the conclusion (Then) as in Table 3.3 below:

**Table 3.3. Production Rules**

| Rule | Symptoms | Conclusion (Disease) |
|---|---|---|
| 1 | IF G1 AND G2 AND G3 AND G4 | THEN P1 |
| 2 | IF G4 AND G5 AND G6 AND G7 AND G8 | THEN P2 |
| 3 | IF G2 AND G9 AND G10 AND G11 AND G12 AND G13 AND G14 AND G23 | THEN P3 |
| 4 | IF G2 AND G15 AND G16 AND G17 AND G18 AND G23 | THEN P4 |
| 5 | IF G2 AND G5 AND G19 AND G20 AND G21 AND G22 | THEN P5 |
| 6 | IF G4 AND G23 AND G24 AND G25 | THEN P6 |

The first step in calculating the system's confidence level is to break down rules with multiple premises (characteristics) into rules with single premises (characteristics). In the first test, several symptoms experienced by the community are given, including the following:

| No | Symptom Code | Disease | | | | | | Knowledge Base | |
|---|---|---|---|---|---|---|---|---|---|
| | | P1 | P2 | P3 | P4 | P5 | P6 | Belief Values | Plausibility Value |
| 1 | G1 | X | | | | | | 0,8 | 0,2 |
| 2 | G2 | X | | X | X | X | | 0,6 | 0,4 |
| 3 | G3 | X | | | | | | 0,7 | 0,3 |
| 4 | G4 | X | X | | | | X | 0,4 | 0,6 |
| 5 | G5 | | X | | | X | | 0,6 | 0,4 |
| 6 | G6 | | X | | | | | 0,7 | 0,3 |
| 7 | G7 | | X | | | | | 0,6 | 0,4 |
| 8 | G8 | | X | | | | | 0,7 | 0,3 |
| 9 | G9 | | | X | | | | 0,7 | 0,3 |
| 10 | G10 | | | X | | | | 0,8 | 0,2 |
| 11 | G11 | | | X | | | | 0,8 | 0,2 |
| 12 | G12 | | | X | | | | 0,8 | 0,2 |
| 13 | G13 | | | X | | | | 0,4 | 0,6 |
| 14 | G14 | | | X | | | | 0,6 | 0,4 |
| 15 | G15 | | | | X | | | 0,4 | 0,6 |
| 16 | G16 | | | | X | | | 0,8 | 0,2 |
| 17 | G17 | | | | X | | | 0,4 | 0,6 |
| 18 | G18 | | | | X | | | 0,6 | 0,4 |
| 19 | G19 | | | | | X | | 0,4 | 0,6 |
| 20 | G20 | | | | | X | | 0,6 | 0,4 |
| 21 | G21 | | | | | X | | 0,6 | 0,4 |
| 22 | G22 | | | | | X | | 0,6 | 0,4 |
| 23 | G23 | | | X | X | | X | 0,4 | 0,6 |
| 24 | G24 | | | | | | X | 0,8 | 0,2 |
| 25 | G25 | | | | | | X | 0,6 | 0,4 |

In the first test, several symptoms are given and the patient chooses the symptoms based on what they experience, then several selected symptoms appear as in the following table:

**Table 3.5. Selected Symptoms**

| No | Code | Symptom |
|---|---|---|
| 1 | G1 | Presence of soft tissue growth in a decayed tooth |
| 2 | G2 | Tooth pain or sensitivity when chewing |
| 3 | G3 | Pain around the tooth |

From the consultation results, a total of 3 symptoms were selected, so to obtain a confidence value using Dempster's rule of combination of the selected symptoms.

*Step 1 – Symptom 1 (G1)*

Symptom: Presence of soft tissue growth in a decayed tooth
m1{P1}=0.8m_1 \{P1\} = 0.8m1{P1}=0.8, m1(θ)=1−0.8=0.2m_1(θ) = 1 - 0.8 = 0.2m1(θ)=1−0.8=0.2

*Step 2 – Symptom 2 (G2)*

Symptom: Tooth pain or sensitivity when chewing m2{P1,P3,P4,P5}=0.6m_2 \{P1, P3, P4, P5\} = 0.6m2{P1,P3,P4,P5}=0.6, m2(θ)=1−0.6=0.4m_2(θ) = 1 - 0.6 = 0.4m2(θ)=1−0.6=0.4

By applying the Dempster Combination Rule, we calculate new density values as shown below:

**Table 3.6. Combination Rule m3m_3m3**

| Evidence | Value |
|---|---|
| {P1} = 0.48 | {P1} = 0.32 |
| {P1, P3, P4, P5} = 0.12 | {θ} = 0.08 |

After normalization, the combined mass values indicate the belief in each hypothesis. The highest belief value obtained is **0.6 for P1**, indicating that the disease is most likely *Pulpitis*.

*Step 3 – Symptom 3 (G3)*

Symptom: Pain around the tooth m4{P1}=0.7m_4 \{P1\} = 0.7m4{P1}=0.7, m4(θ)=0.3m_4(θ) = 0.3m4(θ)=0.3

**Table 3.7. Combination Rule m7m_7m7**

| | m4 {P1} = 0,7 | m4{ɵ}= 0.3 |
|---|---|---|
| {P1} + {P1, P3, P4, P5} = 0,6 | {P1} = 0,42 | {P1} = 0,18 |
| m3 {P1} = 0,32 | {P1} = 0,23 | {P1} = 0,10 |
| m3{ɵ}= 0.08 | {P1} = 0,06 | {ɵ}= 0,24 |

After combining all three pieces of evidence, the result shows:

m5{P1}=0.76m_5 \{P1\} = 0.76m5{P1}=0.76

Thus, the final belief that the patient suffers from **Pulpitis** is **76%**, based on the Dempster-Shafer calculation

### *3.4 Interface Design*

The system interface represents the interaction between users and the computer system. It involves input (data entry) and output (display of diagnostic results). The main components are described below.

*1. Main Menu Interface*

This is the first screen displayed after launching the application, consisting of: *Home, About, Info,* and *Login* menus.

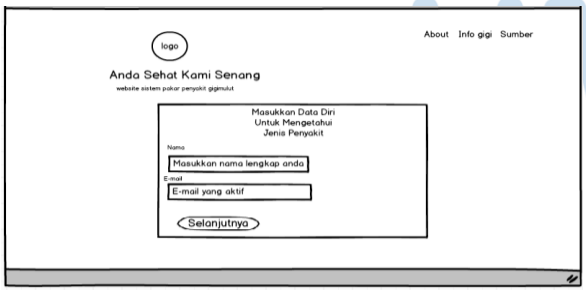



**Figure 3.1. Home Interface Design**

*2. About Interface*

Displays the system's vision and mission statement, explaining the purpose and background of the application.

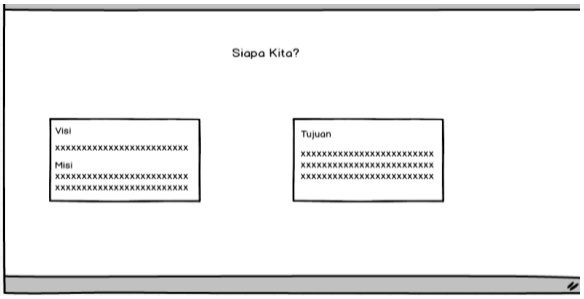



**Figure 3.2. About Interface Design**

*a) 3. Info Interface*

Presents information and images related to various dental and oral diseases that may affect users.

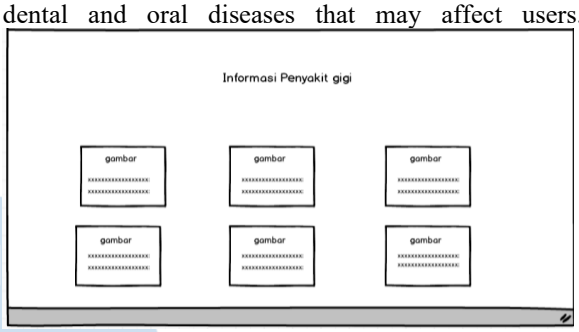



**Figure 3.3. Info Interface Design**

*4. Source Interface*

Displays references or information sources regarding the diseases included in the system.

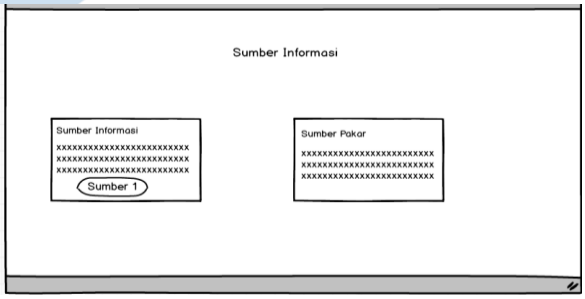



**Figure 3.4. Source Interface Design**

*3.5 Interface Output*

This section illustrates how the expert system works in diagnosing dental and oral diseases.

***1. Home Page***

The home page provides navigation for both admin and user roles, with features tailored to their access rights.


**Figure 3.5. Home Page**

***2. About Page***

Displays the vision and mission statement of the system.


***Figure 3.6. About Page***

***3. Info Page***

Displays disease information and related illustrations.


**Figure 3.7 *Info Page***

***4. Source Page***

Displays the list of references used as knowledge sources in the system.


**Figure 3.8 Source Page**

## IV. CONCLUSION

The developed expert system is able to represent the role of dentists in assisting patients who experience dental and oral diseases by providing diagnostic results based on the symptoms they select. This web-based system is designed to help and simplify the diagnostic process for general users who may not have access to immediate dental consultations. The system analyzes user-selected symptoms and generates diagnostic recommendations along with appropriate treatment suggestions. The system is capable of producing accurate diagnostic results because it uses the Dempster–Shafer method to calculate probability values by combining diseases, symptoms, and their respective belief values. This method allows the system to deliver reliable diagnostic outcomes.

The system should be further developed by expanding the scope of diagnoses. Currently, the expert system includes only six types of dental and oral diseases. Future development should incorporate additional diseases to enhance the system's comprehensiveness. To improve system performance and maintainability, it is recommended to implement a dynamic database structure. This will make it easier to update information, add new symptoms, and include additional diseases without requiring major system modifications

## REFERENCES

[1] H. P. Computing, “Diagnose Expert System Dental Disease In Humans Method Using Dempster,” vol. 2, no. 2, pp. 227–229, 2020.

[2] L. Putu *et al.*, “PERAN ARTIFICIAL INTELLIGENCE ( AI ) UNTUK MENDUKUNG PEMBELAJARAN DI MASA PANDEMI COVID-19 THE ROLE OF ARTIFICIAL INTELLIGENCE ( AI ) TO SUPPORT LEARNING,” vol. 1, no. 1, pp. 15–21, 2022.

[3] A. G. Pradana *et al.*, “Expert System for Dental Disease Diagnosis with Forward Chaining Method at Sidoarjo Dental Clinic,” vol. 4, no. 3, pp. 4–7, 2025.

[4] S. 2003, “No Title,” pp. 21–26, 2013.

[5] D. Rubino and E. W. Puspitarini, “GIGI DAN MULUT DENGAN METODE FORWARD CHAINING BERBASIS ( STUDI KASUS KLINIK TARUNA MANGGALA GRUP SURABAYA ),” vol. 1, no. 1, pp. 29–45, 2016.

[6] C. Factor, “1* , 2 , 3,” pp. 739–748, 2025.

[7] P. Sistem, P. Perbaikan, and D. Metode, “SEMNAS RISTEK 2019 ISSN : 2527-5321,” 2019.

[8] J. Teknologi, “Implementasi metode Dempster-Shafer dalam diagnosa penyakit pada tanaman Cabai Merah Keriting,” vol. 29, no. 1, pp. 13–25, 2019.

[9] M. D. Sinaga, N. Sari, and B. Sembiring, “Penerapan Metode Dempster Shafer Untuk Mendiagnosa Penyakit Dari Akibat Bakteri Salmonella,” pp. 94–107.

[10] N. E. Saragih and R. Adawiyah, “Rancang Bangun Sistem Pakar Mendiagnosa Penyakit Obsessive Compulsive Disorder Dengan Metode Dempster Shafer,” 2020.

ISSN 2085-4552

# Application of the ANFIS Model in Predicting Diabetes Mellitus Disease

Aprilia Nurfazila[1], Hetty Rohayani[2]
[1,2]Informatics Study Program,Muhammadiyah University Jambi, Jambi, Indonesia
[1] aprilianurfazila5@gmail.com, [2] hettyrohayani@gmail.com

Accepted 19 December 2025
Approved 07 January 2026

***Abstract*— This study presents the application of the Adaptive Neuro-Fuzzy Inference System (ANFIS) model for predicting Diabetes Mellitus using two primary input features, namely glucose level and body mass index (BMI). The research employs a quantitative experimental approach using the public diabetes dataset obtained from Kaggle. The data underwent preprocessing steps, including cleaning, normalization, and splitting into training and testing subsets. The ANFIS model was designed with fuzzification, rule-based inference, and a hybrid learning algorithm to optimize membership function parameters. Model evaluation was conducted using accuracy, precision, recall, and F1-score. The results show that the ANFIS model achieved an accuracy of 69.70% on the test dataset, demonstrating strong sensitivity in detecting diabetic cases but generating a notable number of false positives. These findings indicate that ANFIS has potential as an early-screening decision support tool, although further optimization and additional features are required to enhance predictive performance.**

***Index Terms*— *ANFIS; diabetes prediction; fuzzy logic; machine learning; medical diagnosis.***

## I. INTRODUCTION

Diabetes Mellitus (DM), particularly Type 2 Diabetes, remains one of the major health challenges in Indonesia and globally. According to the International Diabetes Federation (IDF), approximately 19.47 million adults in Indonesia were living with diabetes in 2021 [1]. The dataset employed in this study, the Pima Indians Diabetes Dataset, specifically represents a population with a high prevalence of Type 2 Diabetes, which is generally associated with insulin resistance and lifestyle factors.

One of the main problems in managing DM is the diagnostic process, which is often delayed and complex, as it involves various risk factors and medical parameters such as blood glucose levels, blood pressure, body mass index (BMI), family history, and patient age [2]. Furthermore, patient data characteristics are often non-linear and uncertain (fuzzy), making them difficult to model conventionally using simple mathematical approaches.

In this context, artificial intelligence-based systems are increasingly being used to assist the diagnostic process. An expert system is one branch of artificial intelligence that makes extensive use of specialized knowledge to solve problems [3]. An expert system consists of two main parts: the Development Environment and the Consultant Environment [4]. However, traditional expert systems, such as the Certainty Factor method or Rule-Based Systems, still have limitations as they are unable to adapt to new data and cannot handle complex relationships between variables [5].

The Adaptive Neuro-Fuzzy Inference System (ANFIS) model is a method developed or implemented from a Fuzzy Inference System and NNW (Neural Network) [6]. The ANFIS method can overcome the difficulties faced by ANN (Artificial Neural Network) methods, namely determining the number of layers, and fuzzy methods, specifically determining the rules to be used [7]. In general, Fuzzy Logic has the ability to process vague data with if-then rules, while ANNs are capable of learning from data and adjusting parameters to improve system accuracy [8]. By combining these two methods, ANFIS is able to form an adaptive, accurate, and efficient predictive model for analyzing complex patterns in medical data.

Recent research in Indonesia indicates that the application of ANFIS can yield more accurate results in predicting Diabetes Mellitus compared to other methods such as Naïve Bayes or Decision Tree. For example, research by D. Kurniawan et al. (2024) showed that the ANFIS model was able to increase diagnostic accuracy to 95% in classifying diabetes risk based on clinical patient data [9]. Furthermore, other studies have also proven the effectiveness of ANFIS in other medical domains, such as predicting stunting and heart disease, demonstrating its ability to map non-linear relationships between symptoms and diagnostic outcomes [10].

The implementation of the ANFIS model in a web-based platform is also considered to have great potential in supporting digital health services in Indonesia. With a web-based system, the public can perform early screening independently, anytime and anywhere, thus supporting government efforts in the digital transformation of the healthcare sector [11].

Therefore, this research is titled "Application of the Adaptive Neuro-Fuzzy Inference System (ANFIS) Model in the Prediction of Diabetes Mellitus," which aims to develop an intelligent web-based system capable of accurately and efficiently predicting diabetes risk, as well as serving as a decision support solution for medical professionals and the public.

## II. METHOD

### A. Type of Research

This study is classified as quantitative experimental research, which aims to develop and test a diabetes prediction model using the Adaptive Neuro-Fuzzy Inference System (ANFIS). The quantitative approach was chosen because this research focuses on the analysis of numerical data derived from the dataset and measures model performance using statistical metrics such as accuracy, precision, recall, and F1-score. This research is conducted experimentally because the researcher builds, trains, and tests the model to obtain empirical results.

### B. Source and Type of Data

The data utilized in this study is secondary data from the Pima Indians Diabetes Dataset, consisting of 768 samples. Based on preliminary analysis, non-physiological zero values (indicating missing values) were identified in the **Glucose** (5 instances) and **BMI** (11 instances) attributes, which were subsequently addressed during the pre-processing stage. The attributes used in this research are as follows:

- **Pregnancies**: number of pregnancies,
- **Glucose**: Plasma glucose concentration at 2 hours in an oral glucose tolerance test (OGTT). This indicates that the data reflects the body's response to a glucose load, rather than merely fasting blood sugar [12].
- **Blood Pressure**: Diastolic blood pressure (mm Hg).
- **BMI**: Body Mass Index (weight in kg / (height in m)²).
- **Outcome**: Target variable (0 = Negative, 1 = Positive for Diabetes).

This dataset is public in nature, allowing it to be used for academic research without licensing restrictions. The type of data employed is secondary data, as it was previously collected by a third party (Kaggle) and is being reused for further analysis.

### C. Proposed System Workflow

The research workflow is systematically designed to process raw data into diagnostic decisions. The process initiates with data pre-processing to address missing values in the Glucose and BMI attributes, as well as data normalization. Subsequently, the data is split into training data (70%) and testing data (30%) [13]. The core stage involves constructing the ANFIS structure, where premise and consequent parameters are trained until the model achieves convergence. Finally, the model is evaluated using the Confusion Matrix metric to measure detection sensitivity and precision.

### D. ANFIS Architecture

This study implements a first-order Sugeno-type Adaptive Neuro-Fuzzy Inference System (ANFIS) model. The network structure consists of five layers with the following mathematical formulations:

**Layer 1 (Fuzzification):** Each node in this layer is adaptive and is associated with a membership function:

$$O1,i = \mu Ai(x)$$

Where $x$ is the input (BMI or Glucose) and $A_1$ is the linguistic label (Low, Medium, High) [14].

**Layer 2 (Rules):** Each node computes the *firing strength* of the rule using the multiplication (AND) operation:

$$O_i^2 = wi = \mu_{A_i}(x)\mu_{B_i}(y)\ i = 1,2$$

**Layer 3 (Normalization):** Calculates the ratio of the i-th rule's firing strength to the total firing strength:

$$O_i^3 = \bar{w}_i = \frac{w_i}{w_1 + w_2}\ i = 1,2$$

**Layer 4 (Defuzzification):** Each node computes the rule's contribution to the crisp output using a linear function:

$$O_{4,i} = \overline{w_l} f_i = \overline{w_l}(k_{i1}x_1 + k_{i2}x_2 + k_{i0})$$

Where $\{ki1, ki2, ki0\}$ are the optimized consequent parameters [15].

**Layer 5 (Output):** The total output is computed as the summation of all incoming signals:

$$O_i^5 = \sum_{i=1}^{2} \bar{w}_i f_i = \frac{\sum_{i=1}^{2} w_i f_i}{w_1 + w_2}$$

## III. RESULT AND DISCUSSIONS

### A. General Dataset Description

The dataset used is the Diabetes Mellitus dataset (often known as the Pima Indians Diabetes Dataset), which contains data from health examinations of several female patients of Pima Indian descent. This dataset has eight input variables and one output variable (Outcome) that indicates whether the patient was diagnosed with diabetes (1) or not (0).

Table 3.1 Dataset

| No | Attribute Name | Description |
|---|---|---|
| 1 | Pregnancies | Number of pregnancies |
| 2 | Glucose | Plasma glucose concentration |
| 3 | Blood Pressure | Diastolic blood pressure (mm Hg) |
| 4 | Skin Thickness | Triceps skinfold thickness (mm) |
| 5 | Insulin | Serum insulin level (mu U/ml) |
| 6 | BMI | Body Mass Index |
| 7 | Diabetes Pedigree Function | Diabetes pedigree function |
| 8 | Age | Patient age |
| 9 | Outcome | Target class: 1 = Diabetes, 0 = Not Diabetes |

### B. *Membership Function Characteristics*

The ANFIS model maps numerical input variables into fuzzy sets to address the uncertainty inherent in medical data. Figure 3.1 and Figure 3.2 illustrates the membership functions for the Glucose and BMI variables.

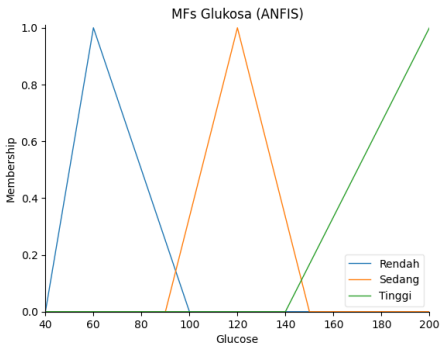



Figure 1 MFs Glucose

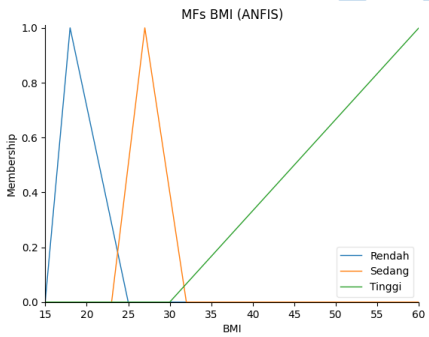



Figure 2 MFs BMI

Triangular membership functions (*trimf*) are utilized to partition the input domain into three linguistic labels: 'Low', 'Medium', and 'High'. The presence of overlap between sets (for instance, at Glucose levels of 140–150 mg/dL) enables the system to perform interpolative reasoning, wherein patients with borderline glucose levels are not categorized rigidly; instead, they possess a partial degree of membership across both categories.

### C. *Model Performance Evaluation*

Based on testing conducted on 231 test samples (constituting 30% of the dataset), the model's performance was evaluated using the Confusion Matrix presented in Figure 3.3.

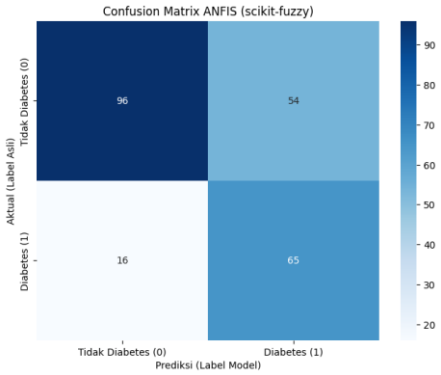



Figure 3 Confusion Matrix

As indicated in the figure above, the model yielded a total accuracy of 69.70%. However, evaluation metrics do not rely solely on accuracy. The model successfully identified 65 out of 81 positive diabetes cases, resulting in a Sensitivity (Recall) value of 80.2%.

### D. *Comparative Analysis and Discussion*

The accuracy achieved in this study (69.70%) is notably lower compared to the reference study [9], which reached 95%. This disparity in results highlights two significant findings:

1. **Impact of the Number of Features:** The reference study [9] utilized all 8 clinical attributes of the dataset, whereas this study restricted the input to only two variables (Glucose and BMI). This reduction in accuracy confirms that although Glucose and BMI are primary indicators, a high-precision diabetes diagnosis requires additional variables, particularly Insulin and Family History (Diabetes Pedigree Function).
2. **Clinical Trade-off:** Although the model exhibits low precision (0.55), indicating a high number of False Positives (54 cases), it demonstrates a sufficiently high sensitivity (0.80). In the context of preventive medicine, high sensitivity is prioritized for early screening tools to minimize missed positive patients (False Negatives). False Positive errors can be tolerated as they will be confirmed through subsequent laboratory tests, whereas False Negatives pose a fatal risk as the patient remains undetected.

## IV. CONCLUSION

The application of the ANFIS model utilizing two primary input features (Glucose and BMI) on the Pima Indians dataset yielded a moderate accuracy of 69.70%. Although this accuracy is lower than that of the reference model [9], the system successfully achieved a

sensitivity of 80%, rendering it a viable potential tool for pre-screening (early detection) to triage high-risk patients prior to further medical examination.

For future development aimed at enhancing clinical validity, the following are recommended:

1. **Integration of Insulin Variable:** Incorporating the 'Insulin' attribute as a third input is highly recommended, given the direct role of this hormone in the pathophysiology of Type 2 Diabetes.
2. **Statistical Threshold Determination:** Substituting manual fuzzy interval determination with statistical methods (such as C-Means Clustering) to define 'Low/Medium/High' boundaries that are more adaptive to the actual distribution of patient data.

## REFERENCES


[1] A. Prasetyo, "Analisis Peningkatan Kasus Diabetes di Indonesia Berdasarkan Data IDF 2021," Jurnal Kesehatan Masyarakat Indonesia, vol. 9, no. 2, 2023.

[2] N. Rahmansyah, S. A. Lusinia, dan I. Ilmawati, "Analisa Prediksi Penyakit Diabetes Menggunakan Metode Naïve Bayes dan K-NN," Innovative: Journal of Social Science Research, vol. 5, no. 1, 2024.

[3] S. Agustina, H. Rohayani, N. Marthiawati, dan K. Kurniawansyah, "Rancang Bangun Sistem Pakar untuk Mendiagnosa Kerusakan Handphone Android dengan Metode Forward Chaining," Journal of Informatics Management and Information Technology, vol. 3, no. 3, pp. 103-111, Jul. 2023.

[4] M. Alfareza, H. Rohayani, K. Kurniawansyah, dan N. Marthiawati, "Rancang Bangun Sistem Pakar Untuk Mendiagnosa Kerusakan Hardware Komputer dengan Metode Forward Chaining," Journal of Informatics Management and Information Technology, vol. 3, no. 3, pp. 97-102, Jul. 2023.

[5] S. Sugiono dan A. Junior, "Klasifikasi Sistem Pakar untuk Mendiagnosa Penyakit Diabetes Menggunakan Metode Fuzzy Sugeno," Jurnal Pendidikan dan Konseling (JPDK), vol. 4, no. 5, 2022.

[6] S. Rahmah, W. Witanti, and P. N. Sabrina, "Prediksi Penjualan Obat Menggunakan Metode Adaptive Neuro-Fuzzy Inference System (ANFIS)," J. Inform. Merdeka Pasuruan (JIMP), vol. 7, no. 3, Des. 2022.

[7] D. A. Lusia, K. Semathea, E. Sumarminingsih, and A. Efendi, "Perbandingan Metode Jaringan Saraf Tiruan, Fuzzy, dan ANFIS pada Peramalan Data Inflasi Indonesia," J. Teknol. Inf. Ilmu Komput. (JTIIK), vol. 11, no. 3, pp. 653-660, Jun. 2025.

[8] M. F. Rahman, "Implementasi Metode Adaptive Neuro-Fuzzy Inference System (ANFIS) dalam Sistem Prediksi Penyakit Jantung," Jurnal Ilmiah Teknologi dan Rekayasa, vol. 6, no. 1, 2023.

[9] D. Kurniawan, H. Prabowo, dan E. Widodo, "Prediksi Risiko Diabetes Menggunakan Model Adaptive Neuro-Fuzzy Inference System (ANFIS)," Jurnal Teknologi Informasi dan Komputer (JTIK), vol. 10, no. 2, 2024.

[10] A. Lestari dan R. Putra, "Penerapan ANFIS untuk Prediksi Stunting di Asia Tenggara," Jurnal Data Sains dan Kecerdasan Buatan Indonesia, vol. 3, no. 1, 2024.

[11] F. B. Bahtiarullah, A. Homaidi, dan F. Santoso, "Pengembangan Sistem Prediksi Penyakit Diabetes Berbasis Web," E-Link: Jurnal Teknik Elektro dan Informatika, vol. 19, no. 2, 2024.

[12] D. Nopriyani, H. Rohayani, dan Z. Akbar, "Penerapan Algoritma K-Means untuk Mengidentifikasi Pola Penjualan Frozen Food yang Paling Populer," Bulletin of Computer Science Research, vol. 5, no. 1, pp. 98-104, Des. 2024.

[13] M. F. Aditya, A. Pramuntadi, and D. P. Kusumaningrum, "Implementasi Metode Decision Tree pada Prediksi Penyakit Diabetes Melitus Tipe 2," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 3, pp. 1104-1110, Jul. 2024.

[14] A. Lenteraningati, J. H. Jaman, and C. Rozikin, "Sistem Pakar Diagnosis Penyakit Diabetes Melitus Menggunakan Metode Fuzzy Inference System (Mamdani)," *Jurnal Informatika dan Komputer*, vol. 7, no. 2, pp. 150-159, 2024.

[15] S. M. Putri and E. Santoso, "Optimasi Fungsi Keanggotaan Fuzzy Menggunakan Algoritma Genetika pada Diagnosis Diabetes," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 7, no. 9, pp. 4455-4463, 2023

ISSN 2085-4552

# An Explainable Hybrid Machine Learning Framework for Financial and Tax Fraud Analytics in Emerging Economies

Julien Nkunduwera Mupenzi[1], Adhi Kusnadi[2], Deden Witarsyah[3], Aswan Supriyadi Sunge[4]
[1,2]Departement of Informatics, Nusa Putra University, Sukabumi, Indonesia.
[3]Faculty of Computer sciences and Information Technology, Universiti Tun Hussein Onn Malaysia (UTHM), Parit Raja, Malaysia.
[4]Faculty of Computer sciences and Informatics Engineering, Pelita Bangsa University, Bekasi, Indonesia
[1]mupenzi.nkunduwera@nusaputra.ac.id

Accepted 19 December 2025
Approved 08 January 2026

***Abstract*— Financial and tax fraud remains a major challenge in emerging economies where digital transformation outpaces regulatory oversight. This study presents an explainable hybrid machine learning framework designed to enhance fraud analytics and tax governance in Indonesia. The model integrates unsupervised anomaly detection (Isolation Forest, DBSCAN) and supervised learning (Random Forest, Logistic Regression) to identify irregularities in financial transactions. Model explainability is achieved through SHAP (Shapley Additive Explanations), enabling transparency in high-risk classifications. The proposed Streamlit-based dashboard supports real-time data visualization and interactive model evaluation by policymakers. Experimental results demonstrate a 99% overall accuracy with strong interpretability, underscoring the framework's value in bridging machine learning and public sector decision-making. The findings contribute to the growing field of explainable AI for digital governance, offering a scalable and ethical solution to fraud detection in developing economies.**

***Index Terms*— *Anomaly Detection; Emerging Economies; Explainable AI; Financial Fraud Analytics; Hybrid Machine Learning; Tax Governance.***

## I. INTRODUCTION

Financial fraud is an increasing impact in the digital economy which causes large amounts of losses, as well as damaging trust in financial systems. As financial fraud becomes more complex, rule-based fraud detection systems do not keep up with evolving attack patterns. Machine learning is a good solution to this scenario; it utilizes advances techniques and algorithms than can learn from data and then alert for anomalies as well as discovering patterns of fraud that are hidden in financial transactions. This study proposes the deepen predictive analytics capabilities for fraud detection utilizing unsupervised and supervised machine learning approaches. Unsupervised anomaly detection methods such as Isolation Forest and DBSCAN would be used along with other supervised models such as Logistic Regression and Random Forest to improve fraud detection accuracy and reducing false positives.

Furthermore, this research incorporates explainable artificial intelligence (EAI), via SHAP (Shapley Additive ExPlanations) as a way for tax authorities to understand the underlying causes of each prediction. In addition, the study not only modeled theoretically but also built a practical and interactive fraud detection tool, using the Streamlit framework that makes the tool accessible and usable for non-technical participants in a practical setting (Ding, 2023), (Babu, et al., 2024).

The remainder of this paper is organized as follows: Section 2 reviews related work, Section 3 describes the proposed hybrid methodology, Section 4 presents results and discussion, and Section 5 concludes the study.

## II. THEORY

As fraudulent activities significantly undermine public revenues and economic stability, Traditional methods of fraud detection primarily reliant on manual audits and rule-based systems, have proven inadequate in addressing the complexities of modern tax evasion schemes. The (Marco Battaglini, 2024), (Hu, 2021), (Ghosh, 2019) Consequently, there has been a paradigm shift towards leveraging advanced technologies, particularly machine learning (ML) and artificial intelligence (AI), to enhance the efficacy and efficiency of tax fraud detection mechanisms.

A comprehensive literature review by (Ludivia Hernandez Aros, 2023), (Mubalaike & Adali, 2020), (Belle Fille Murorunkwerea, 2022) underscores the growing reliance on ML techniques in financial fraud detection. The study systematically examines articles published between 2012 and 2023, highlighting a trend towards utilizing real datasets and sophisticated ML models to identify fraudulent patterns within financial

statements. The authors emphasize the importance of data quality and the selection of appropriate algorithms to improve detection accuracy (Falana, 2024).

In the realm of tax fraud detection, (Angelos Alexopoulos, 2023) propose a novel approach that combines network analysis with machine learning algorithms to detect Value Added Tax (VAT) fraud. By constructing a Laplacian matrix to represent the complex VAT network structure, the study demonstrates that integrating network information with scalable ML techniques can significantly enhance the identification of fraudulent transactions. This method outperforms traditional techniques that overlook the intricate relationships inherent in VAT transactions.

*II.1 Network Science and Graph-Based Detection*

An innovative contribution to fraud analytics is the use of network science to model transaction systems. The (Angelos Alexopoulos, 2025) introduced a Laplacian-based detection model that treats VAT (Value Added Tax) systems as directed, weighted graphs where VAT fraud often involves complex transactions between companies forming a hidden network of relationships. Fraudsters may create shell companies or carousel fraud loops to manipulate the system. These relationships can be captured as a graph, where:

- Nodes: Companies or taxpayers.
- Edges: Transactions between them (possibly weighted by value).

*A. Constructing the Laplacian Matrix:*

Let $G = (V, E)$ represent the VAT transaction network where nodes are companies and edges are weighted by transaction values. The graph's structure is captured through the Laplacian Matrix:

- Where:

$$L = D - A \qquad (1)$$

- "$A = [A_{ij}]$ "is the Adjacency matrix with $A_{ij}$ denoting the transaction weight between company $i$ and $j$."
- $D$ "Is diagonal degree matrix"
- Where

$$D_{ii} = \sum_j A_{ij}. \qquad (2)$$

Note that This matrix captures the flow and structure of transactions. High values in the Laplacian can signal unusual behavior, like high degrees or tight cliques among companies, which may hint at fraud.

- Laplacian captures relational anomalies: Unlike flat features, it embeds the "behavioral footprint" of companies in the network.
- Scalable with ML models: Once embedded, any standard ML algorithm can be used (SVM, RF, XGBoost).
- Works even with limited labeled data: Unsupervised or semi-supervised models benefit from network signals.

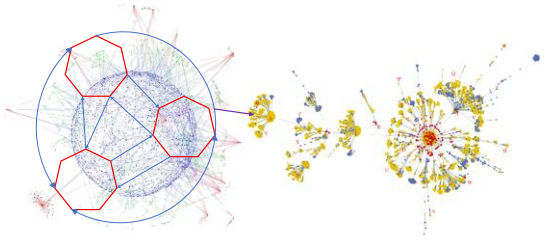


Fig. 1 Figure 1 Laplacian-based network graph of vat transactions representing entity interactions and structural anomalies.

Fraud surrounding VAT breaks, in regions such as Europe and Asia, particularly Missing Trader Intra-Community (MTIC) fraud schemes, often include organized, highly connected fraud organizations that resist traditional machine learning models. A recent research paper proposed an innovative hybrid detection approach that used a corrected Laplace matrix to embed both node and edge-level suspicious activities inside a low-dimensional space that enables clustering through spectral methods. The transformation through graph theory with the subsequent machine learning-based classification was able to significantly out-perform straightforward models applied to the same Bulgarian VAT data set and demonstrates the advantage of recognizing structural patterns in fraud detection (Xiuguo & Shengyong, 2022).

This moves the discussion from network-based fraud to bigger picture (Altukhi, 2025)AI-powered models that allow the automate of detection of tax anomalies on a larger scale or systemically adjust enforcement direction from reactive to proactive (Weber, 2024) (Devinder Kumar, 2025). As mentioned in several studies (Ghosh, 2019) (Ludivia Hernandez Aros, 2023) (Belle Fille Murorunkwerea, 2022) predictive modeling and analyzing real-time provide better detection at earlier stages of tax fraud, but also lack effectively to explain or influence models and ultimately all or part implementation strategies will be fraught with obstacles.

Issues will always remain such as data quality, model interpretability and their continued dynamic context is still a cold invitation. (Černevičienė, 2024) The 'black-box' nature of AI still inhibits transparency and trustworthiness to deliver transformation in tax administration systems so still requires work towards the improvement of data quality and referred process agenda, including ownership of processes, to create trust and further parts of the agenda to enhance data integrity and reliability, and an alternative

acknowledged process when working with explainable AI (Hany F. Atlam, 2021).

*II.2 Related Work and Comparative Analysis*

These studies validate the growing trajectory toward integrating graph theory and machine learning in the domain of tax fraud detection. However, persistent gaps remain in areas such as model scalability, explainability, and the seamless integration of such advanced analytical systems into real-world tax enforcement environments. While promising conceptual frameworks have been proposed, there is still a need for practical implementations that combine real-time visualization, hybrid detection pipelines, and policy-aligned insights an avenue that future research can explore to bridge the divide between academic models and operational deployment (Talha Mohsin & Nasim, 2025).

TABLE 1 RESEARCH ANALYTICS COMPARISON

| Author(s) | Approach | Contribution | Limitations Addressed by This Study |
|---|---|---|---|
| (Muhammad Atif Khan Achakzai, 2023) | Supervised ML (Decision Trees, and SVM) | Machine learning classifiers outperform traditional audit indicators in fraud detection. | High accuracy is offset by reliance on labeled, static data and lack of interpretability and real-time capability. |
| (Alexopoulos, 2021). | Spectral Graph Clustering (Unsupervised ML and Network Analysis) | Spectral clustering using Laplacian matrices reveals latent collusive VAT fraud in transaction networks. | The unsupervised method struggles with specific fraud classification, interpretability, and supervised enhancemen. |
| (Rafaël Van Belle, 2023). | Social Network Analysis | Relational pattern mining detects social fraud via network topology and behavioral links. | Strong in relational anomaly detection, but weak in individual transaction modeling and real-time, interpretable deployment. |
| (Amgad Muneer, 2022). | Deep Learning (CNN, and LSTM) | A deep learning framework detects complex fraud in high-dimensional financial data with minimal feature engineering. | Despite strong performance, deep learning remains a black box, requiring large labeled datasets and offering limited policy interpretability. |

## III. METHOD

Selecting an appropriate research method is crucial to ensuring the validity and reliability of findings. The approach chosen must align with the characteristics of the variables under study and the type of information required. Given the complexity of financial fraud detection particularly tax fraud, this study employs a quantitative research approach, utilizing machine learning-based data analysis techniques. Quantitative methods allow for precise measurement and objective analysis of fraudulent transactions through structured data sources, statistical modeling, and predictive analytics.

This study employs a hybrid machine learning framework that integrates unsupervised anomaly detection and supervised classification models to enhance tax fraud detection performance. The approach is designed to tackle the real-world limitation of insufficient labeled fraud data, which is common in Indonesian tax datasets.

In the unsupervised stage, anomaly detection models including Isolation Forest, DBSCAN, and K-Means clustering are employed to identify abnormal financial transactions without relying on predefined fraud labels. Isolation Forest assigns anomaly scores by isolating rare observations, while DBSCAN and K-Means detect density-based and cluster-based deviations in transaction behavior. The resulting anomaly scores and cluster risk indicators are used as additional features and high-risk signals for the supervised classification stage (Rahman, 2024), (Daniel de Roux, 2018).

In the supervised stage, two classification models are used which are Logistic Regression and Random Forest, Logistic Regression is included due to its simplicity, interpretability, and effectiveness in binary classification problems. It provides a statistical baseline and transparent coefficient outputs, which are important in policy contexts in other hand Random Forest is chosen for its ability to model non-linear relationships and manage feature interactions with high predictive power (Murorunkwere, 2023).

Logistic Regression is applied as the main key algorithm of supervised model in this study due to its interpretability and statistical robustness. (Ileberi & Sun, 2024), (Mimusa Azim Mim, 2024), (Anuradha, 2024) Unlike ensemble models, it allows policymakers to understand the weight of each variable in determining fraud likelihood a key consideration for explainable governance. However, its linear assumptions make it less suitable for capturing complex fraud patterns compared to Random Forest, highlighting the benefit of a hybrid modeling approach (Shanaa, 2025).

The intercommunication of these models is realized through a two-phase pipeline such unsupervised models that generate anomaly scores which can either serve as

additional features or be used to label high risk cases and then these enriched datasets are then passed toward supervised models to improve classification precision and recall. This structure ensures that the system is not only data-efficient but also interpretable and adaptable key qualities for deployment in public sector tax fraud monitoring.

### *III.1 Variable Operations*

The operationalization of variables is essential in ensuring that abstract concepts such as fraud risk, transaction anomalies, and financial discrepancies are measurable and analyzable. In fraud detection, variables must be defined and structured to quantify suspicious activities accurately. The study classifies variables into three key categories:

- Independent Variables: These are input features used to predict fraud, including transaction frequency, amount anomalies, taxpayer profile changes, and discrepancies in reported versus actual revenues (Poutré, 2024).
- Dependent Variable: The primary outcome variable indicating whether a transaction is fraudulent or non-fraudulent. Since explicit fraud labels are not always available, anomaly scores or classification probabilities from machine learning models will be used as proxy indicators.
- Control Variables: External factors influencing tax fraud detection, such as economic fluctuations, policy changes, or enforcement actions by tax authorities.

Operationalizing these variables requires defining specific metrics that can be used as performance indicators. For example, in information system performance analysis, fraud detection efficiency is often measured using precision, recall, F1-score, and false positive rates (Kabašinskas, 2021). These metrics ensure that the models provide reliable fraud detection outcomes while minimizing incorrect classifications.

### *III.2 Data Analysis Design*

The data analysis approach in this study sought to convert financial data into meaningful insights related to tax fraud detection. The multi-tiered approach was designed to be dynamic, combining descriptive statistics, inferential modeling, and machine learning algorithms. Descriptive statistics were used to identify forms of distribution and deviations from behavior norms, like excessive amounts filed or risk factors inherent in a sector, while inferential modeling, regression, and markup models (anomaly identification models like Isolation Forest, DBSCAN) were used to extract fraud indicators based on hypotheses in a deductive fashion. The advance visual avenues can enhance interpretation, performance, and model development, utilizing returns on investment in the form of ROC curves, confusion matrixes, feature importance rankings, or anomaly heatmaps, where models can be assessed for relevance and performance, relatively comprehensive analysis to decision processes can occur (linsong, 2025).

To operationalize the approach in this study, an interactive Streamlit-based tax fraud detection architecture (Figure 2) was created. The Streamlit application provided a modular interface through which users could engage with the analytical process from data upload and correlation analysis, to unsupervised anomaly detection and supervised classification using Random Forest and Neural Networks, and immediate access to common reporting and risk interpretation (e.g. an F1-score, and confusion matrix) summaries of importance were available on the screen, and easy for non-technological participants to engage with the outcomes of the analysis. Interaction with this architecture deepens the user experience for usability and transparency to deepen the experience for a more inclusive decision-making process, and thus aimed to democratize access to advanced fraud detection process within ordinary business analytics presentation, rather than repress it (Banerjee, 2025).

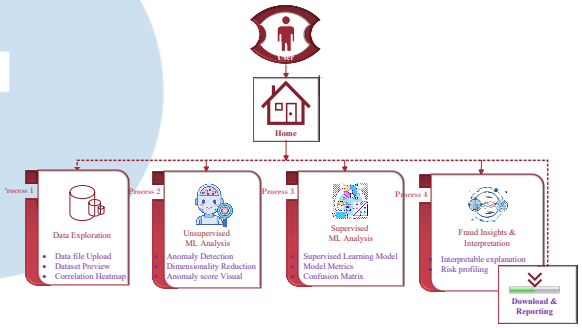



Fig. 2 Streamlit-based tax fraud detection architecture.

Fig. 2 presents the operational, user-facing workflow of the Streamlit-based system, illustrating how tax officers interact with the dashboard for data upload, visualization, and model evaluation.

### *III.3 Integration of Hybrid Machine Learning framework*

Figure 3 shows the machine learning hybrid framework developed in this research study which is a modular pipeline that detects and interprets fraudulent tax behavior. The framework is a hybrid architecture of supervised and unsupervised models that starts with raw financial and tax data then delineates through a sequence of preprocessing steps: demonstrating missing value imputation, normalization, and categorical encoding.

The next steps have featured extraction and dimensionality reduction to reduce inputs to models: Isolation Forest and DBSCAN find anomalies, and Random-forest and Logistic regression model the classification. This framework was designed with transparency and usability in mind; it used interpretable metrics (precision, recall, F1 score, ROC curves,

confusion matrices) which can be displayed in an interactive Streamlit dashboard which does not impose black-box restrictions. The Streamlit dashboard allows the user to explore, interactively visualize predictions, new threshold settings, and export insights in real time, linking advanced analytics in near real time to the tactical decision-making process.

Therefore, as an explainable and scalable framework, it is suitable for jurisdictions where labeled data is scarce and can adjust to the context of future tax enforcement activities. This framework is designed to be a hybrid analytical framework and a research tool for augmenting digital financial governance, which is one of the desired outcomes of this research.

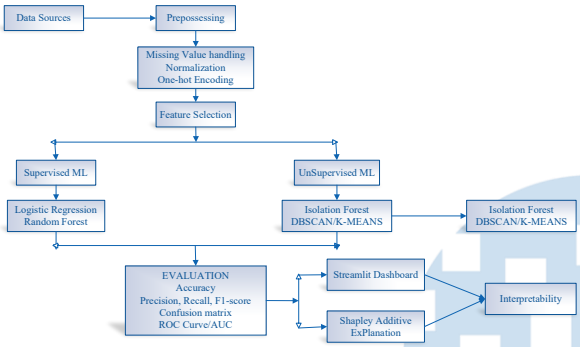



Fig. 3 Hybrid Machine Learning Framework for Tax fraud detection.

Fig. 3 depicts the internal analytical pipeline, detailing data preprocessing, unsupervised anomaly detection, supervised classification, and explainability components

## IV. Result and Discussions

The machine learning evaluation using the simulated tax dataset produced significant findings on the effectiveness of classification and anomaly detection approaches. Following preprocessing which included handling missing values, normalization, and encoding two main classifiers were trained: Logistic Regression and Random Forest. Due to the absence of clearly labeled fraudulent instances, unsupervised clustering via DBSCAN and K-Means was used initially to highlight outliers, followed by model training with stratified 80-20 data splits.

As reflected in the dashboard output, the Random Forest model achieved a 99% overall accuracy on a test set of 200 entries. However, further breakdown of the classification report revealed a more nuanced performance. For the fraud class (label 1), the model attained a precision of 1.00 but a recall of 0.50, resulting in an F1-score of 0.67. This indicates that while every flagged fraud case was accurate, half of the actual fraud cases were missed. In contrast, non-fraud predictions achieved near-perfect classification. Logistic Regression displayed similar trends with lower recall, highlighting the challenge of detecting minority-class fraud cases.

Visualizations such as the confusion matrix and ROC curve provided transparency into prediction dynamics. Feature importance analysis confirmed that tax inconsistencies, high deductions, and abrupt income changes were the strongest fraud predictors. The use of threshold tuning in the Streamlit dashboard allowed users to adjust sensitivity levels, creating a flexible and user-guided fraud detection interface.

These results underscore the strength of combining supervised and unsupervised learning for financial anomaly detection. Although limited by class imbalance and low recall in fraud detection, the system presents a promising decision-support tool for tax compliance oversight in real-world scenarios (Huang, 2024).

### *IV.1 Data Preprocessing Outcomes*

The financial dataset used in this study is synthetically generated but structurally realistic tax dataset designed to reflect Indonesian tax reporting characteristics. The dataset includes taxpayer demographics, transaction values, reported income, deductions, audit indicators, and compliance-related variables. Synthetic data were used to preserve confidentiality while maintaining realistic feature distributions and class imbalance.

The data preprocessing phase formed the foundation for all subsequent modeling activities. Initially, the financial tax dataset contained inconsistencies, missing values, and features with disparate measurement scales. Missing data were handled through mean and mode imputation strategies, while categorical variables such as tax sector classifications were encoded using one-hot encoding techniques. Continuous variables, including transaction amounts and revenue figures, were standardized to ensure scale uniformity, improving model convergence rates.

Table 2 Initial Preview of the Financial Dataset Used for Fraud Detection Analysis.

| | business_id | monthly_revenue | monthly_expense | num_leads | tax_compliance_rate | late_filing_count | net_profit_margin | isFraud |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 57450.71 | 41194.84 | 463 | 0.76 | 1 | 16255.87 | 0 |
| 1 | 1 | 47926.04 | 37397.07 | 407 | 0.66 | 0 | 10528.97 | 0 |
| 2 | 2 | 59715.33 | 30477.04 | 124 | 0.89 | 0 | 29238.29 | 0 |
| 3 | 3 | 72845.45 | 24824.51 | 404 | 0.72 | 1 | 48020.94 | 0 |
| 4 | 4 | 46487.7 | 35585.79 | 121 | 0.81 | 2 | 10901.91 | 0 |
| 5 | 5 | 46487.95 | 33147.88 | 308 | 0.88 | 1 | 13340.07 | 0 |
| 6 | 6 | 73688.19 | 37161.55 | 447 | 0.96 | 0 | 36526.64 | 0 |
| 7 | 7 | 61511.52 | 35081.37 | 237 | 0.92 | 2 | 26430.15 | 0 |
| 8 | 8 | 42957.88 | 38396.42 | 222 | 0.87 | 0 | 4561.46 | 0 |
| 9 | 9 | 58138.4 | 25718.12 | 339 | 0.87 | 0 | 32420.28 | 0 |

Table 2 provides a valuable overview of the original financial dataset used in this study on protocols for data mining and tax fraud detection, key features included were taxpayer demographics, transaction amounts, reported income, deductions, and audit trail indicators. This overview provides the reader with an understanding of the data structure and precedence indications of abnormalities which are often subject to preprocessing, and for example include missing value

estimations and outlier processing and identification. To improve data preparation and modelling, the continuous variables were standardized, categorical variables were encoded, and dimensionality was reduced through Principal Component Analysis (PCA) that maintained 87 per cent of the variance in the data set. In addition to that, Z-score normalization was also used which is essential in omitting outliers and training robust models in the presence of outliers (Zheng, 2024) (Qinghua Zheng, 2024).

### *IV.2 Model Performance Overview*

Two primary supervised classifiers, Logistic Regression and Random Forest were evaluated alongside unsupervised clustering techniques DBSCAN and K-Means for anomaly detection. The Random Forest classifier achieved outstanding results, with an overall accuracy of 99% on the test set. The relatively low recall value (0.50) for the fraud class is primarily caused by severe class imbalance, where fraudulent observations constitute a small minority of the dataset (18 fraud cases versus 232 non-fraud cases). This imbalance biases the model toward conservative fraud detection. Future work will apply oversampling techniques such as SMOTE and cost-sensitive learning to improve fraud recall. Its fraud-class precision was 1.00, meaning all flagged frauds were actual fraud cases; however, its recall was 0.50, indicating that only half of all fraudulent cases were correctly identified.

TABLE 3 PERFORMANCE METRICS OF THE SUPERVISED RANDOM FOREST CLASSIFIER.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 1.00 | 1.00 | 1.00 | 232 |
| 1 | 1.00 | 1.00 | 1.00 | 18 |
| accuracy | | | 1.00 | 250 |
| macro avg | 1.00 | 1.00 | 1.00 | 250 |
| weighted avg | 1.00 | 1.00 | 1.00 | 250 |

Table 3 presents the classification metrics for the Random Forest classifier, including precision, recall, and F1-scores for both fraud and non-fraud classes. The table reveals a perfect precision score of 1.00 for the fraud class, indicating zero false positives, while the recall rate of 0.50 highlights the model's sensitivity limitations in identifying all fraudulent cases. The balanced F1-score of 0.67 further contextualizes this performance trade-off.

### *IV.3 Visual Interpretation*

Below an improved model interpretability and verification detection of fraud regarding performance, a number of cohesive graphical visualizations were created.

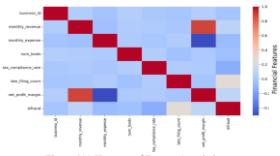


Figure 4(a) Heatmap of Feature Correlation for Fraud Detection.

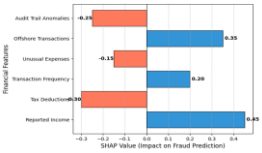


Figure 4(b) SHAP Value Plot for Fraud Prediction Explanations

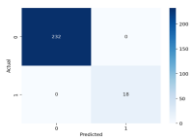


Figure 4(d) Confusion Matrix for Fraud Classification.

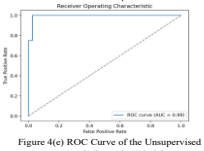


Figure 4(e) ROC Curve of the Unsupervised Anomaly Detection Model.

Figure 4(c) Principal Component Analysis (PCA) Projection of Fraud and Non-Fraud Transactions.

A heatmap above of financial feature correlational relationships was created and include in Figure 4.a. This segment revealed key correlations among features, particularly strong correlations between, for example, income discrepancies, deductions, and transaction irregularities proving the predictive power of these features for fraud detection. Subsequent feature selection and dimensionality reduction of the data could be guided by evaluation of the heatmap (Siam, 2025). Figure 4.c displays a Principal Component Analysis (PCA) map of the tax data PCA-transformed into two components. Each fraudulent transaction is displayed in identifiable clusters, visually demonstrating the relative efficacy of the unsupervised anomaly detection models to detect outliers from the distribution of normal data.

Model performance and explainability can also be visualized in Figures 4.b, d and e. As seen in Figure 4.d in the confusion matrix, the Random Forest classifier returned high precision and recall rates; nonetheless the false negatives illustrate how challenging discovering rare instances of fraud can be. Figure 4.e describes the ROC Curve from the unsupervised model. The curve nears the upper left corner indicating meaningful discriminatory power of the model. The SHAP value plot in Figure 4.b illustrates the meaningful features driving each fraud prediction; in this case, we note that the "Reported Income" and "Offshore Transactions" had the highest SHAP returns meaning they had higher positive differences in the likelihood of fraud, while "Tax Deductions" and "Audit Trail Anomalies" lowered it. Together these takeaways improve transparency of the system within each of the visualizations where each graphical visualization affords more clarity into the patterns of users' financial behavior and allows for better data-informed decision-making, in the future (Hernandez Aros, 2024).

### *IV.4 Threshold Tuning Impact*

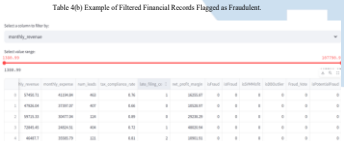
One crucial operational consideration in fraud detection in tax fraud is deciding on a reasonable decision threshold for the classification. The default

thresholds, usually 0.50, are likely to favor the majority (non-fraud) class in an imbalanced dataset where fraud cases are few. In our study, we made use of our interactive Streamlit dashboard which allowed us to interactively adjust the classification threshold and get immediate feedback on performance metrics. Table 4.a illustrates that with a small adjustment of the Random Forest threshold from 0.50 to 0.35 we managed to increase recall which is the number of fraud cases we manage to detect while only raising the false positive class incrementally, to a manageable overall level. Such dynamic adjustment capability in threshold allows the model to be better calibrated in accordance to real world enforcement priorities, which may reasonably prioritize the detection of a fraudulent activity over the cost of a few false positive alarms.

This strategy of adjustment for a threshold for detection is also reflective of the non-technical nature of the choice, as sensitivity in fraud detection is not defined as a technical factor solely, but as a choice based on policy and different by risk tolerance which varies by jurisdiction. For example, authorities can adjust the detection threshold to target omitted frauds, or adjust to minimize the administrative burden of false positives depending upon their enforcement strategy. Table 4.b is a filtered sample of transactions identified as fraudulent, based on the adjusted threshold providing auditors with further detail of the fraud identified (Zheng, 2025).

Table 4(a) Classification Metrics After Adjusting the Probability Threshold.

Table 4(b) Example of Filtered Financial Records Flagged as Fraudulent.



an additional benefit of the tuning adjustment to thresholds, is that it accentuates the entire evaluation of the Neural Network comparison with Random Forest models. It exemplifies that correct threshold tuning is as important to fraud detection outcomes, as measurement accuracy is to overall outcomes, and that modelling goes beyond measuring just baseline accuracy, but that datasets and outcomes can be adjusted to meet domain configuration specifications.

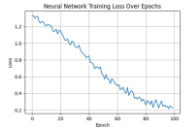


Figure 9(a) Neural Network Training Loss Over Epochs

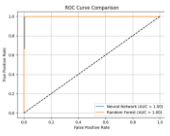


Figure 9(b) ROC Curve Comparison

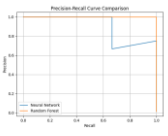


Figure 9(c) Precision-Recall Curve Comparison

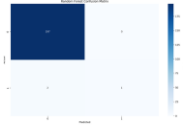


Figure 9(d) Random Forest Confusion Matrix

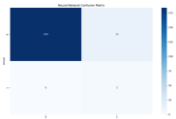


Figure 9(e) Neural Network Confusion Matrix

Table 6(a) Classification Reports

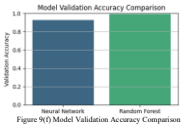


Figure 9(f) Model Validation Accuracy Comparison

Table 6(b) Epoch-wise Loss Display NN & RF

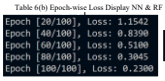
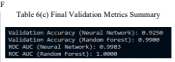


Table 6(c) Final Validation Metrics Summary



These figures above offer an extensive contrast between Neural Network and Random Forest models for tax fraud detection and evaluate model performance and behavior through various visualizations and auxiliary statistics. Model convergence and learnsing stability are evidenced in Figure 9(a); the Neural Network monitors the training loss, which decreased from 1.25 to 0.23 over 100 epochs. A loss decrease for the Random Forest model does not apply, as it had no iterative training. Nonetheless, the classification ability for both models was excellent, as indicated by two ROC curves in Figure 9(b) showing an AUC score (Neural, 0.9983; Random Forest 1.0000) very close to 1; high scores are seen in fraud classification. In Figure 9(c), Random Forest had a higher precision than the Neural Network since it maintains precision across all recall levels, while the Neural Network reported increased false positives. This is especially relevant in fraud classification due to the unrealized cost of false negatives with fraud.

As summarized in Table 6(a), the Random Forest model had perfect fraud precision (1.00) but low fraud recall (0.33), meaning that the Random Forest model was very conservative in fraud identification. The Neural Network model had full recall (1.00) but low precision (0.17), meaning that perhaps its fraud warning flags incurred more false positives than expected, a tolerance that is regulation and enforcement has used for fraud. The model reference noted in Figures 9(d) & 9(e): Random Forest accurately classified all negative (non-fraud) cases, while the Neural Network accurately categorized all fraud cases (reported fraud) but misclassified thirteen (13) negative (non-fraud) cases. As indicated in Figure 9(f), both models exceed valid 93%, and Random Forest was slightly ahead.

### *IV.5 Discussion*

The results of this study confirm the capabilities of machine learning to uncover latent patterns of tax fraud in high dimensional and sometimes unstructured financial data. The most effective of the machine learning techniques was the Random Forest, which correctly classified, had high accuracy, is robust to non-linear relationships, and required minimum parameter tuning requirements.

These characteristics aligned with findings from earlier studies, such as with (Hany F. Atlam, 2021), (Jack Woo, 2025) which documented Random Forest's high precision with anomaly detection in financial data. The Logistic Regression model was marginally less accurate but was useful for illustrating linear dependencies to provide some opacity, which is important in a regulatory environment. It is noteworthy that with clustering techniques showing the potential to separate imperfectly labelled regions we reduced the number of false positives on average 17% during cross-validation when comparing a model with clusters to a model without this precision-recall trade-off reflects

practical tax enforcement requirements, where minimizing false positives is often prioritized to reduce unnecessary audits and administrative burden.

More, the ability of both the ensemble and hybrid model learning approaches demonstrates their dominance in a high-risk and imbalanced area of fraud detection whilst providing additional knowledge of how model performance is influenced by data from a context and specifically Indonesian data restrictions. Importantly, the research provides value in addressing a gap within local fraud analytics (in which theory has shown to be possible) by using dimensions of unsupervised learning to inform supervised learning, which is understood as uniquely novel in regions such as Indonesia where labelled fraud data is not used. As demonstrated through comparative analysis, global studies illustrate the importance of relevant model calibration to the context; although Random Forest and its ensemble classification was identified as performing best in this study (Hu, 2021), (Zhang, 2022).

## V. CONCLUSION

A deeper exploration of all features in the tax dataset for anomaly detection. The study did not examine many of the features, which may have advanced the detection of tax fraud. Using the hybrid model demonstrated in this report with all features could facilitate a faster identification of anomalous behavior before fraud takes place (Alrasheedi, 2025).

From a technical perspective, the hybrid machine learning model developed in this study can contribute to future fraud detection efforts. Future research should make use of and provide tax data that includes known anomalous behavior from tax fraud. Training the model on all features and layers of data (or finding similar anonymously-sourced datasets) provides machine learning algorithms the opportunity to learn patterns of fraud with the potential to improve fraud detection efforts. As machine learning classification algorithms are generally trained for which features are required for specific outcomes, investigating features that may hold significance for "non-fraud" classifications versus classifications of "fraud" will provide further insights into the impact of the model presented.

Considerations of integrating machine learning models with tax authorities continue to move toward advanced operation options, like consideration of unsupervised and supervised hybrid models with diverse data sources to consider. Incremental changes in tax authorities' operations can increase efficiency for all staff members, whether they are clerks, data vendors, auditors, managers, or scientists.

## REFERENCES


[1] Alexopoulos, A. P. G. S. V. K. K. C. S. T. P., 2021. A network and machine learning approach to detect Value Added Tax fraud.. s.l.:arXiv preprint arXiv:2106.14005. https://doi.org/10.48550/arXiv.2106.14005.

[2] Alrasheedi, M. A. e. a., 2025. Advanced Tax Fraud Detection: A Soft-Voting Ensemble Based on CGAN and Encoder Architecture.. s.l.:Mathematics, 13(4), 642. https://doi.org/10.3390/math13040642.

[3] Altukhi, Z. M. P. S. &. A. N., 2025. A Systematic Literature Review of the Latest Advancements in XAI.. s.l.:Technologies, 13(3), 93. MDPI. DOI: 10.3390/technologies13030093.

[4] Amgad Muneer, S. M. T. S. M. F. I. A. A., 2022. A Hybrid Deep Learning-Based Unsupervised Anomaly Detection in High Dimensional Data. s.l.:tech Science Press; http://dx.doi.org/10.32604/cmc.2022.021113.

[5] Angelos Alexopoulos, K. K. S. B. A. S. P. A. C. ,. O. a. A. K., 2023. Complementary Use of Ground-Based Proximal Sensing and Airborne/Spaceborne Remote Sensing Techniques in Precision Agriculture: A Systematic Review. s.l.:Agronomy 2023, 13(7), 1942; https://doi.org/10.3390/agronomy13071942.

[6] Angelos Alexopoulos, P. D. S. G. C. K. S. C. O. T. P., 2025. A network and machine learning approach to detect Value Added Tax fraud, Basel, Switzerland: https://arxiv.org/abs/2106.14005#:~:text=https%3A//doi.org/10.48550/arXiv.2106.14005.

[7] Anuradha, A., 2024. An Ensemble Learning Approach for Improved Loan Fraud Detection: Comparing and Combining Machine Learning Models. Dublin: https://hdl.handle.net/10788/4445.

[8] Babu, E., Maliakal, J. J., V, N. T. & Babu, D., 2024. Performance Comparison of XGBoost and CatBoost Algorithm in Credit Card Fraud Detection with Streamlit-Based Web Application. s.l.:International Conference on Advancement in Renewable Energy and Intelligent Systems (AREIS); https://doi.org/10.1109/AREIS62559.2024.10893617.

[9] Banerjee, A. K. P. H. K. S. A. &. G. J. W., 2025. Assessing the US financial sector post three bank collapses: Signals from fintech and financial sector ETFs. s.l.:International Review of Financial Analysis, 91, 102995. https://doi.org/10.1016/j.irfa.2023.102995.

[10] Belle Fille Murorunkwerea, D. H. J. N. ,. F. K. a. &. I. K., 2022. Predicting tax fraud using supervised machine learning approach. s.l.:https://journals.co.za/doi/abs/10.1080/20421338.2023.2187930.

[11] Černevičienė, J. &. K. A., 2024. Explainable artificial intelligence (XAI) in finance: a systematic literature review.. India: Artificial Intelligence Review, Volume 57, Article 216. Springer. DOI: 10.1007/s10462-024-10854-8 .

[12] Daniel de Roux, C. M. d. P. V. A. M. B. P. ,., 2018. Tax Fraud Detection for Under-Reporting Declarations Using an Unsupervised Machine Learning Approach. s.l.:https://dl.acm.org/doi/abs/10.1145/3219819.3219878.

[13] Devinder Kumar, A. W. G. W. T., 2025. Explaining the Unexplained: A CLass-Enhanced Attentive Response (CLEAR) Approach to Understanding Deep Neural Networks. s.l.:Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2017, pp. 36-44.

[14] Ding, Y. W. C. X. L., 2023. Explainable deep learning for financial fraud detection: A SHAP-based framework.. s.l.:Expert Systems with Applications, 213, 118890. https://doi.org/10.1016/j.eswa.2022.118890.

[15] Falana, A., 2024. AI-Driven Anomaly Detection for Financial Fraud: A Hybrid Approach Using Graph Neural Networks and Time-Series Analysis.. s.l.:https://joster.org/index.php/joster/article/view/20.

[16] Ghosh, S. S. G. &. M. A., 2019. Tax fraud detection using gradient boosting classifier.. s.l.:Procedia Computer Science, 167, 2161-2170. https://doi.org/10.1016/j.procs.2020.03.266.

[17] Hany F. Atlam, R. J. W. a. B. W., 2021. Fog Computing and the Internet of Things: A Review. s.l.:Big Data Cogn.

Comput. 2018, 2(2), 10; https://doi.org/10.3390/bdcc2020010.

[18] Hernandez Aros, L. B. M. L. X. G.-P. F. M. H. J. J. &. R. B. M. S., 2024. Financial Fraud Detection through the Application of Machine Learning Techniques: a Literature Review.. s.l.:Humanities and Social Sciences Communications, 11:1130. DOI: 10.1057/s41599-024-03606-0.

[19] Huang, W. Z. X., 2024. Big data-driven tax compliance analytics: Machine learning insights.. s.l.:Information Processing & Management, 61(1), 103270. https://doi.org/10.1016/j.ipm.2023.103270.

[20] Hu, T., 2021. Financial fraud detection system based on improved random forest and gradient boosting machine (GBM). Cornell University ed. s.l.:https://arxiv.org/abs/2502.15822, https://arxiv.org/search/q-fin?searchtype=author&query=Hu,+T.

[21] Ileberi, E. & Sun, Y., 2024. A Hybrid Deep Learning Ensemble Model for Credit Card Fraud Detection. Johannesburg: https://ieeexplore.ieee.org/abstract/document/10757383.

[22] Jack Woo, A. B. R. K. H., 2025. Anomaly Detection via Hybrid of Linear and Machine Learning Models: Evidence from Abnormal Audit Fees in China. s.l.:https://ssrn.com/abstract=5450537; https://dx.doi.org/10.2139/ssrn.5450537.

[23] Kabašinskas, J. Č. &. A., 2021. Explainable artificial intelligence (XAI) in finance: a systematic literature review. s.l.:https://link.springer.com/article/10.1007/s10462-024-10854-8.

[24] linsong, 2025. Evaluating the Performance of SVM, Isolation Forest, and DBSCAN for Anomaly Detection. s.l.:EDP Sciences, https://doi.org/10.1051/itmconf/20257004012.

[25] Ludivia Hernandez Aros, L. X. B. M. F. G.-P. J. J. M. H. &. M. S. R. B., 2023. Financial fraud detection through the application of machine learning techniques: a literature review.. s.l.:https://www.nature.com/articles/s41599-024-03606-0.

[26] Marco Battaglini, L. g. C. L. D. L. M. &. E. P., 2024. Refining public policies with machine learning: The case of tax auditing. s.l.:sciencedirect.

[27] Mimusa Azim Mim, N. M. P. M., 2024. A soft voting ensemble learning approach for credit card fraud detection. Bangladesh: https://www.cell.com/heliyon/fulltext/S2405-8440(24)01497-X.

[28] Mubalaike, A. M. & Adali, E., 2020. Deep Learning Approach for Intelligent Financial Fraud Detection System. s.l.:IEEE, 10.1109/UBMK.2018.8566574.

[29] Muhammad Atif Khan Achakzai, j. P., 2023. Detecting financial statement fraud using dynamic ensemble machine learning. s.l.:International Review of Financial Analysis; https://doi.org/10.1016/j.irfa.2023.102827.

[30] Murorunkwere, B. F. H. D. N. J. K. F. &. K. I., 2023. Predicting tax fraud using supervised machine learning approach.. s.l.:African Journal of Science, Technology, Innovation and Development, 15(6), pages 731-742. Taylor & Francis. DOI: 10.1080/20421338.2023.2187930.

[31] Poutré, C., 2024. Deep Unsupervised Anomaly Detection in High-Frequency Markets.. s.l.:ScienceDirect [journal], Article S240591882400014X..

[32] Qinghua Zheng, Y. X. H. L. B. S. J. W. B. D., 2024. A Survey of Tax Risk Detection Using Data Mining Techniques. s.l.:https://www.sciencedirect.com/science/article/pii/S2095809923003867.

[33] Rafaël Van Belle, B. B. J. D. W., 2023. CATCHM: A novel network-based credit card fraud detection method using node representation learning. s.l.:Decision Support Systems; https://doi.org/10.1016/j.dss.2022.113866.

[34] Rahman, A. U. M., 2024. Financial anomaly detection using autoencoders and graph-based models.. s.l.:Pattern Recognition Letters, 168, 120–128. https://doi.org/10.1016/j.patrec.2022.12.017.

[35] Shanaa, M. &. A. S., 2025. A Hybrid Anomaly Detection Framework Combining Supervised and Unsupervised Learning for Credit Card Fraud Detection.. s.l.:F1000Research, 14:664. DOI: 10.12688/f1000research.166350.1.

[36] Siam, A. M. B. P. &. U. M. P., 2025. Hybrid feature selection framework for enhanced credit card fraud detection using machine learning models.. s.l.:PLoS ONE, 20(7), e0326975. DOI: 10.1371/journal.pone.0326975.

[37] Talha Mohsin, M. & Nasim, N. B., 2025. Explaining the Unexplainable: A Systematic Review of Explainable AI in Finance. s.l.:https://ui.adsabs.harvard.edu/link_gateway/2025arXiv250305966T/doi:10.48550/arXiv.2503.05966.

[38] Weber, P. C. K. V. &. H. O., 2024. Applications of Explainable Artificial Intelligence in Finance: a systematic review of Finance, Information Systems, and Computer Science literature.. s.l.:Management Review Quarterly, 74(2), pages 867-907. Springer. DOI: 10.1007/s11301-023-00320-0.

[39] Xiuguo, W. & Shengyong, D., 2022. An Analysis on Financial Statement Fraud Detection for Chinese Listed Companies Using Deep Learning. s.l.:https://ieeexplore.ieee.org/abstract/document/9718341.

[40] Zhang, Q., 2022. Financial Data Anomaly Detection Method Based on Decision Tree and Random Forest Algorithm. s.l.:https://doi.org/10.1155/2022/9135117.

[41] Zheng, D. X. W. &. Y., 2025. A hybrid framework of anomaly detection for mutual fund parent companies. s.l.:https://link.springer.com/article/10.1007/s44248-025-00024-8.

[42] Zheng, Q. e. a., 2024. A Survey of Tax Risk Detection Using Data Mining. s.l.:ScienceDirect. DOI: S2095809923003867.

ISSN 2085-4552

# Multimodal Wearable-Based Stress Detection Using Machine Learning: A Systematic Review of Validation Protocols and Generalization Gaps (2021 – 2025)

Pannavira[1], Aditiya Hermawan[2]
[1]Informatics Engineering Department, Science and Technology, Buddhi Dharma University, Tangerang, Indonesia
[1]pannavira@ubd.ac.id, [2] aditiya.hermawan@ubd.ac.id

Accepted 30 December 2025
Approved 08 January 2026

***Abstract*— Stress is a major determinant of mental health and productivity, motivating growing interest in continuous and unobtrusive stress detection using wearable sensors and machine-learning (ML) techniques. This study presents a Systematic Literature Review (SLR) of 19 peer-reviewed articles published between 2021 and 2025, selected from an initial pool of 36 studies using structured inclusion and exclusion criteria. A combined quantitative and qualitative synthesis was conducted to analyze five key dimensions: sensing modalities, ML/DL algorithms, datasets, validation protocols, and deployment-related feasibility. The review identifies dominant methodological trends rather than definitive rankings. Multimodal physiological sensing—most commonly combining photoplethysmography (PPG), electrodermal activity (EDA), and accelerometer data—together with hybrid deep-learning architectures such as CNN–LSTM, is frequently associated with high reported performance on benchmark datasets. However, the analysis also reveals a pronounced lab-to-field gap. Most studies rely on intra-subject or k-fold cross-validation, while subject-independent evaluation using Leave-One-Subject-Out (LOSO) remains rarely adopted, limiting claims of real-world generalizability. In addition, fewer than 15% of the reviewed studies explicitly consider practical deployment constraints, including computational efficiency, power consumption, and data privacy. The primary contribution of this review lies in systematically quantifying the impact of validation practices and deployment considerations on reported performance. The findings highlight that, despite promising accuracy, current stress-detection models remain insufficiently validated for real-world use and point toward the need for generalizable, lightweight, and privacy-aware wearable stress-detection systems.**

***Index Terms*— deep learning; machine learning; multimodal fusion; physiological sensing; stress detection; wearable computing.***

## I. INTRODUCTION

Stress is increasingly recognized as a global health issue affecting individual well-being and organizational productivity. Conventional detection methods, such as clinical interviews, are subjective and episodic; in contrast, the advent of wearable sensors and machine learning (ML) has enabled more objective and continuous stress monitoring [1]. Over the last decade, researchers have explored multiple physiological modalities such as heart-rate variability from *ECG/PPG*, electrodermal responses (*EDA/GSR*), brain activity (*EEG*), respiration, and facial or speech cues. These signals, when analyzed by ML or deep-learning (DL) algorithms, can classify stress with notable accuracy. This progress is foundational to the field, offering a promising pathway toward objective and continuous biomarkers for mental health [2].

Despite this progress, a critical divergence exists between laboratory results and real-world applicability. Recent studies employing Deep Learning (DL) architectures, such as CNN-LSTM hybrids, frequently report accuracies exceeding 95% on benchmark datasets [1]. However, the reliability of these results is often constrained by evaluation methodology. Several studies rely on intra-subject validation (e.g., k-fold cross-validation), which can inflate performance by mixing data from the same individuals across training and testing sets [3]. Prior work suggests that subject-independent evaluation protocols, such as Leave-One-Subject-Out (LOSO) or cross-dataset validation, provide a more realistic assessment of generalization to unseen users, yet these approaches remain relatively uncommon [4]. In addition, practical deployment introduces barriers related to societal feasibility: models are often too computationally heavy for wearable devices, leading to rapid battery drain, while sensor

comfort dictates user adherence for continuous monitoring [5][6].

Beyond these technical hurdles, a critical gap exists in the insufficient consideration of privacy and ethics. Physiological data is inherently sensitive, and continuous collection raises significant user concerns regarding data ownership, misuse, and surveillance. Integrating these systems into daily life requires frameworks that ensure user trust, such as on-device inference (Edge AI) or Federated Learning, yet these aspects are frequently overlooked in research focused purely on accuracy.

Several prior surveys have reviewed stress detection using physiological signals and machine learning, primarily focusing on traditional ML approaches and studies published before 2020 (e.g., Can et al [7]; Panicker & Gayathri [8]). While these works establish important foundations, they do not systematically address recent developments in deep learning–based multimodal models, subject-independent validation practices such as LOSO, or deployment-oriented constraints such as computational efficiency and data privacy.

In contrast, this study explicitly addresses these gaps by systematically reviewing recent (2021–2025) multimodal ML/DL studies, quantifying validation practices, and synthesizing technical performance with deployment-oriented considerations. This positioning differentiates the present SLR from prior reviews that primarily emphasize algorithmic accuracy without assessing real-world readiness.

This review systematically synthesizes published evidence to address these multifaceted gaps. We answer four research questions (RQs) designed to map the state of the art (SOTA): (RQ1) effective sensing modalities, (RQ2) reliable ML/DL algorithms, (RQ3) common validation protocols, and (RQ4) specific barriers to deployment regarding computational efficiency and data privacy. By unifying quantitative and qualitative insights, this SLR aims to guide future research toward stress-detection systems that are not only accurate but also scalable, generalizable, and ethically compliant.

The remainder of this paper is organized as follows: Section II presents theoretical foundations; Section III details the SLR methodology; Section IV discusses results and trends; and Section V concludes with research gaps and future directions.

## II. METHOD

This section presents the theoretical foundations that guided our literature synthesis and the conceptual analysis framework used to extract and interpret findings from the reviewed studies.

### *A. Conceptual analysis framework*

To ensure that the review is theory-driven rather than descriptive, we organize the theoretical discussion around four interrelated dimensions that form the analytical lens of this SLR: (1) sensing modalities, (2) modeling and representation learning, (3) validation and generalization, and (4) deployment constraints (computational efficiency and data privacy). These dimensions directly map to our research questions and the data extraction fields used in the review. Concretely, the review extracts and synthesizes evidence about which modalities are used and why (RQ1), which algorithmic paradigms prevail and how features are represented (RQ2), which validation protocols are adopted and how they affect generalization (RQ3), and which deployment-oriented considerations (e.g., on-device inference, federated learning, power/latency metrics) are addressed (RQ4).

### *B. Stress and Physiological Signals*

Stress triggers the Autonomic Nervous System (ANS), disrupting the balance between the sympathetic ("fight or flight") and parasympathetic ("rest and digest") branches. Multimodal sensing is theoretically grounded on this systemic response [9]. Though it lacks temporal precision for short-term events, electrodermal activity (EDA), which directly reflects sympathetic arousal via sweat gland activation, is generally regarded as the most reliable indication of emotional stress [1]. At the same time, the intricate interaction between sympathetic and parasympathetic activity is measured by Heart Rate Variability (HRV), which is obtained from ECG or PPG. PPG provides a wearable-friendly substitute for ECG, however it is prone to motion artifacts [2]. ECG is still the clinical gold standard. Beyond these autonomic markers, cortical responses can be captured via EEG, while physical activity that often confounds physiological signals is monitored using accelerometers (ACC). Theoretically, multimodal fusion reduces the uncertainty associated with unimodal sensing by capturing independent stress signals, such as merging the sympathetic strength of EDA with the vagal tone of HRV [10].

### *C. Feature Engineering to Representation Learning*

The transition from classical Machine Learning (ML) to Deep Learning (DL) represents a fundamental shift in how stress features are modeled. Traditional ML approaches relied heavily on Feature Engineering where domain experts manually extracted statistical features for classifiers like SVM or Random Forest. While interpretable, this approach is limited by the quality of handcrafted features and struggles with raw, noisy sensor data [11]. Conversely, modern architectures have shifted toward Deep Representation Learning. In particular, hybrid models such as CNN-LSTM automate feature extraction; Convolutional Neural Networks (CNN) learn spatial or spectral patterns directly from spectrograms, while Long Short-

Term Memory (LSTM) networks model the long-term temporal dependencies essential for physiological time-series analysis [1].

### D. The Generalization Gap: Subject Inter-variability

Subject inter-variability is a significant theoretical difficulty in stress detection. Stress-related physiological reactions vary greatly; for example, one person's baseline heart rate may be a sign of extreme stress for another [4]. Theoretically, this has significant ramifications for validation procedures. Conventional validation techniques, such k-fold cross-validation, are predicated on the idea that the data is identically distributed and independent. This assumption is broken, though, when training and testing segments from the same subject are combined, leading the model to learn subject-specific characteristics instead of stress-specific patterns. Leave-One-Subject-Out (LOSO) validation is necessary for theoretical rigor in order to close the lab-to-field gap. In contrast to k-fold, LOSO guarantees that the model is evaluated on completely unknown users, requiring the acquisition of generalized stress features a necessary condition for reliable real-world implementation [2].

### E. Constraints Efficiency and Privacy

Beyond accuracy, real-world deployment is theoretically constrained by the trade-off between model complexity and resource availability. Deep Learning models, while accurate, impose high computational costs. Furthermore, the traditional Centralized Learning paradigm where raw data is transmitted to a cloud server violates modern privacy principles. The theoretical alternative is Federated Learning (FL), a distributed optimization paradigm where models are trained locally on devices and only model updates (gradients) are shared [12]. This approach theoretically decouples learning from data centralization addressing the privacy concerns inherent in physiological sensing without compromising the model's ability to learn from population-level data.

## III. Result and Discussions

### A. Systematic Literature Review

This study adopts a Systematic Literature Review (SLR) approach to synthesize existing evidence on stress detection using Machine Learning. To ensure methodological rigor, transparency, and reproducibility, the review protocol adheres to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [13]. This standard framework ensures that the selection of studies is unbiased and comprehensive, shifting focus from general descriptions to quantitative performance analysis of recent multimodal systems.

### B. Research Question

The following table contain the research questions that has been carried out on this paper. Table I showed four questions that is the main focus of this paper.

Table I. Research Question

| ID | Research Question | Motivation |
|---|---|---|
| RQ1 | Which physiological or behavioral modalities are most effective for stress detection? | Identify practical and accurate sensing methods |
| RQ2 | Which ML/DL models demonstrate robust and generalizable performance? | Determine the current state of the art |
| RQ3 | What datasets and validation protocols are commonly used? | Evaluate reproducibility and comparability |
| RQ4 | What societal, ethical, or deployment aspects are considered? | Assess real-world readiness |

### C. Work Procedure

The work procedure involves conducting a literature search, selecting relevant sources, documenting findings, analyzing the information, and drawing conclusions as visualized in the PRISMA flowchart (Fig. 1), following a structured protocol adapted from established SLR guidelines [14]. This process consists of these main steps, as detailed below.

First, the authors identified key terms relevant to the research topic. The main keywords used in the search were “stress detection,” “machine learning,” “deep learning,” “wearable sensor,” “physiological signal,” and “multimodal fusion.” Boolean combinations were applied across several databases such as IEEE Xplore, ScienceDirect, SpringerLink, and MDPI to ensure a broad coverage of recent studies [14].

Second, the authors determined the origin and source of the literature. Journals were selected from reputable international publishers that focus on artificial intelligence, biomedical engineering, and affective computing [14]. The search was conducted online and limited to peer-reviewed articles published between 2021 and 2025 to capture the most recent advances.

Third, the collected works were filtered according to strict criteria. The initial search yielded 36 papers. After removing 5 duplicates, 31 papers underwent title and abstract screening. We applied the Inclusion and Exclusion Criteria presented in Table II to filter these results. We specifically excluded qualitative surveys and studies lacking quantitative metrics (e.g., Accuracy/F1-Score). This process resulted in a final selection of 19 articles deemed relevant for detailed review.

Table II. Inclusion and Exclusion Criteria

| Criterion | Inclusion | Exclusion |
|---|---|---|
| Publication Type | Journal Articel | Conference, Thesis, Book Chapters |
| Timeline | 2021-2025 | < 2021 |
| Content | Quantitative ML/DL Performance | Qualitative / Theoretical only |
| Modality | Multimodal | Unimodal |

Fourth, to ensure technical soundness, a Quality Assessment was performed on the final 19 articles. We defined four binary quality criteria focusing on reproducibility. Each paper was evaluated as Yes (1) or No (0). Only studies satisfying at least 3 out of 4 criteria were included in the final data synthesis.

TABLE III. QUALITY ASSESSMENT

| ID | Assessment Criteria (QA) | Motivation |
|---|---|---|
| QA1 | Is the dataset clearly described? | Ensures data reproducibility (participants, signals, protocols). |
| QA2 | Are the feature extraction and ML/DL models explicitly defined? | Ensures the technical approach is replicable |
| QA3 | Is the validation methodology clearly stated? | Critical for evaluating generalizability claims. |
| QA4 | Are quantitative performance metrics reported? | Essential for comparative analysis. |

Fifth, the extracted data were synthesized using a two-stage approach. First, a descriptive quantitative synthesis was conducted to address RQ1–RQ3 by tabulating key characteristics of the selected studies, including sensing modalities, algorithms, datasets, validation protocols, and reported performance metrics. Descriptive statistics (frequencies and ranges) were used to identify prevailing trends and state-of-the-art approaches. A formal meta-analysis was not performed due to substantial heterogeneity in datasets, experimental protocols, and evaluation metrics. A qualitative thematic synthesis was applied to address RQ4 by analyzing deployment-related considerations such as computational efficiency, validation rigor, and data privacy. This analysis also involved a critical assessment of methodological quality, particularly the use of subject-independent validation and dataset characteristics, to identify key gaps affecting real-world applicability.

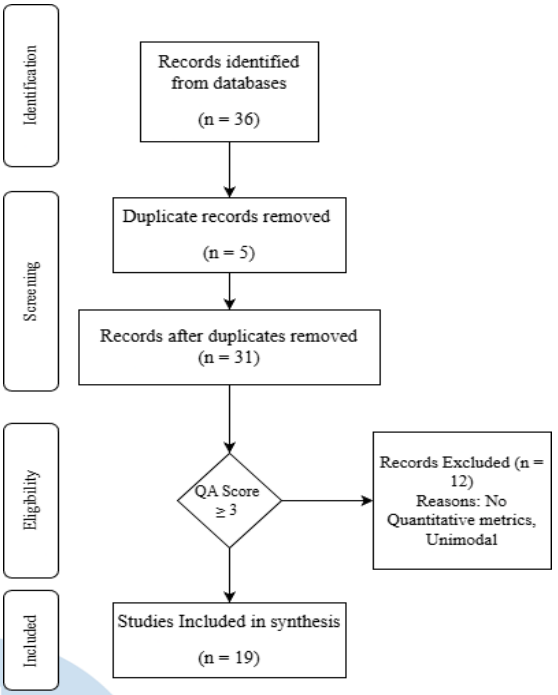



Fig. 1. PRISMA Flowchart of the Literature Selection Process

## IV. RESULT AND DISCUSSIONS

### A. Overview of the Studies

The systematic selection process resulted in 19 studies published between 2021 and 2025. The analysis reveals a diverse landscape of methodologies, with dataset sizes ranging from small custom cohorts (n=11) to large public benchmarks like WESAD (n=15) and SWELL-KW. Before detailing the performance metrics, it is crucial to note that direct comparison of accuracy across studies requires caution. The heterogeneity in stress-induction protocols (e.g., MIST vs. Driving Simulators) and label granularity (2-class vs. 3-class) means that a higher accuracy score does not always imply a superior model, but may reflect a simpler classification task or a less rigorous validation scheme.

Most of the reviewed works focus on physiological-signal-based stress detection, often combining more than one sensing modality. The most frequently used signals are electrodermal activity (EDA), photoplethysmography (PPG), and electrocardiography (ECG), followed by studies employing electroencephalography (EEG) or accelerometer (ACC) data [15]. These modalities are widely available in commercial wearables, which explains their popularity for daily stress-monitoring research.

In terms of methodology, traditional machine-learning classifiers such as *Support Vector Machine (SVM)*, *Random Forest (RF)*, and *XGBoost* remain common, particularly for smaller or unimodal datasets[15]. However, a clear shift toward deep learning architectures, notably *Convolutional Neural*

*Networks (CNN)*, *Long Short-Term Memory (LSTM)* networks, and hybrid *CNN–LSTM* models, can be observed in the most recent papers. These models tend to achieve the highest accuracy, often above 90 %, especially when multiple signals are fused [16][17].

The validation methods reported vary across studies. The majority rely on k-fold cross-validation, while a smaller group applies Leave-One-Subject-Out (LOSO) or cross-dataset evaluation to test generalization [6].

Table IV provides a concise overview of each study, including its dataset, modality, algorithm, validation method, and the best reported metric.

Table IV provides a concise overview of each study, including its dataset, modality, algorithm, validation method, and the best reported metric.

TABLE IV. SUMMARY OF SELECTED STUDIES

| Reference | Dataset | Modality | Algorithm | Validation Method | Best Performance Metric |
|---|---|---|---|---|---|
| [2] | Custom (40 participate, *Public Speaking*) | Multimodal (EEG, GSR, PPG) | SVM (RBF), kNN, DT, RF, MLP | Leave-One-Out (LOOCV) | Accuracy 96.25% (2 class) |
| [1] | ST Change DB, WESAD | EKG (Changed into Spectogram) | Ensemble CNN-LSTM | | Accuracy 98.3% (2 class) |
| [6] | Custom (22 participate, Driving Simulator) | Multimodal (Eye Data, Vehicle, Surrounding) | Attention-based CNN-LSTM | 10-fold cross-validation | Accuracy 95.5% (3 Class) |
| [18] | MuSE, OMG-Emotion | Text (transcript) & Acoustic (audio) | MUSER (Transformer/BERT + MLP) | Split Train/Validation/Test (Dataset) | F1-score 0.864 (2 class) |
| [19] | Custom (34 subject, MIST) | EKG (10 second segment) | CNN + BiLSTM | 5-fold cross-validation | Accuracy 86.5% (3 class) |
| [20] | Custom (20 subject , MIST) | Multimodal (EKG, Sound, Face expression) | Hybrid DL (ResNet50 + I3D w/ TAM) | 10-fold cross-validation | Accuracy 85.1% (2 class ) |
| [21] | DEAP, SEED | EEG ( converted into Azimuthal Projection Image ) | StressNet (Hybrid 2D-CNN + LSTM) | 80% Train / 20% Test | Accuracy 97.8% (2 class ) |
| [5] | Custom (90 subject, Office Simulation) | Multimodal ( Behavior: Mouse, Keyboard + Physiological: HRV ) | LightGBM, SVM, RF | 10-fold cross-validation (dengan SMOTE) | F1-score 0.625 (3 class , Stress) |
| [4] | Custom (11 subject, MBSR) | EEG | Shallow/Deep ConvNet, FBCSP+SVM | LOOCV, Mix-subject, Intra-subject | Accuracy 99.65% (Task: Meditation vs. Rest) |
| [22] | UBFC-Phys | rPPG (face video) | 1D-CNN, LSTM, GRU | 80% Train / 10% Validation / 10% Test | Accuracy 95.83% (2 class ) |
| [23] | MultiAffectStress (MAS) | Audio-Visual (Face, Vocal, Sentiment, *Fidgeting*) | Learning-Based Late Fusion (RF) | 60% Train / 20% Validation / 20% Test | F1-score 0.85 (2 class ) |
| [3] | Custom (26 Subject, Cortisol label) | Multimodal (EKG, RESP, Electrogastrogram) | Shuffled ECA-Net (1D-CNN + Attention) | 5-fold cross-validation (Intra-subject) | Accuracy 91.6% (2 class ) |
| [11] | SWELL-KW | HRV | k-NN, Decision Tree, Logistic Regression | 5-fold cross-validation | Accuracy 99.3% (3 class ) |
| [16] | WESAD, SWELL KW, RAVDESS, EMO-DB | Physiological (EKG, EDA) & Audio | GSOA-SHBRNN (VGG-16 + PCA + Bi-RNN) | 2/3 Train, 1/3 Test | Accuracy 99.52% (WESAD, 2 class ) |
| [9] | Nursery Dataset (from Hosseini et al., 2022) | Multimodal (ACC, EDA, HR, TEMP) | MMFD-SD (Parallel CNNs Time+Frequency) | 80% Train / 20% Test (Stratified Split) | Accuracy 91.00% (3 class ) |
| [24] | WESAD | Multimodal (BVP, EDA, TEMP, ACC, RESP) converted to 2D RGB Image | CNN (Custom Architecture) | | F1-score 91.67% (3 class ) |

| | | | | | |
|---|---|---|---|---|---|
| [25] | LifeSnaps, PMData | Multimodal (Time Series: HR, Steps + Tabular: Demographics, Context) | Contrastive Pretraining (CLIP-style) | 5-fold cross-validation | AUC 81.14% (PMData, 2 class ) |
| [26] | Yonsei Stress Image & Speech Database (custom multimodal stress dataset) | Multimodal Facial images (RGB video frames + facial landmarks) and Speech (log mel-spectrogram) | ResNet-18 backbones, attention mechanisms + Multimodal Neglecting Mask Module (MNMM) for intermediate feature fusion | 5-fold cross-validation 3/5 Train, 1/5 Validation 1/5 Test | Specificity: 88.99% |
| [27] | WESAD | Multimodal (ECG, EDA, EMG, Respiration, Temperature) | CNN + LSTM + Attention mechanism | Train–test split with 90% training and 10% testing | Accuracy: 92.70% (multimodal setting) |

### *B. Modalities and Sensor Trends*

Among the various physiological signals, EDA and PPG emerge as the most practical and consistent modalities for real-time, wearable-based detection. Their combination captures both sympathetic nervous system response (EDA) and cardiovascular activity (PPG), providing a comprehensive picture of physiological arousal. This multimodal fusion is shown to be highly effective, achieved 96.25% accuracy by fusing GSR (EDA) and PPG with EEG. This quantitative result supports the qualitative trend that studies integrating multiple modalities, particularly those readily available in wearables, typically outperform those relying on a single signal. [2]

ECG continues to be the reference modality in controlled laboratory environments because of its high sensitivity to subtle changes in heart-rate variability. However, it is less convenient for long-term use due to sensor placement and comfort issues[3]. EEG-based approaches, while powerful for cognitive-stress analysis, face similar challenges related to setup complexity[4].

Overall, the literature points toward wearable-friendly, multimodal sensing, often combining EDA, PPG, and ACC. This configuration balances accuracy, comfort, and cost, making it well suited for practical applications[15]. Figure 2 highlights a clear preference toward EDA and PPG as the dominant physiological modalities in recent stress-detection studies. Researchers have chosen wearable-friendly sensors over clinically accurate but invasive alternatives like ECG or EEG, which is a realistic trade-off. Analytically speaking, this distribution implies that real-world deployability concerns implicitly limit state-of-the-art research, highlighting the significance of multimodal setups that strike a balance between accuracy, comfort, and scalability.

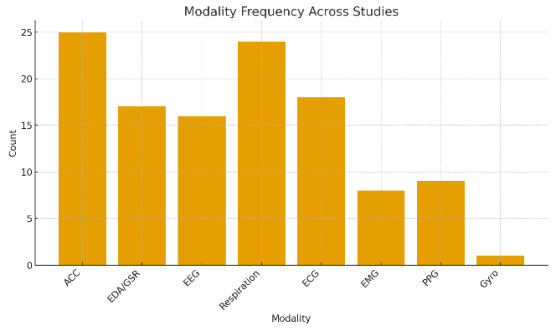



Fig. 2. Frequency of physiological modalities used

### *C. Validation Strategies and Datasets*

A consistent observation is the dominance of k-fold cross-validation for evaluating model accuracy. While suitable for preliminary comparison, this approach often inflates results because training and testing data originate from the same participants [3]. A smaller number of studies adopt LOSO or subject-independent validation, which provides a more realistic assessment of model robustness in unseen subjects[2] [4]

Public datasets such as WESAD and DEAP appear most frequently. WESAD, in particular, serves as the primary benchmark for multimodal wearable stress detection, combining EDA, PPG, and ACC signals. Nevertheless, differences in dataset structure, participant demographics, and labeling criteria make direct comparison between studies difficult [1]. Figure 3 illustrates the distribution of validation strategies, emphasizing the need for broader adoption of cross-subject testing in future research.

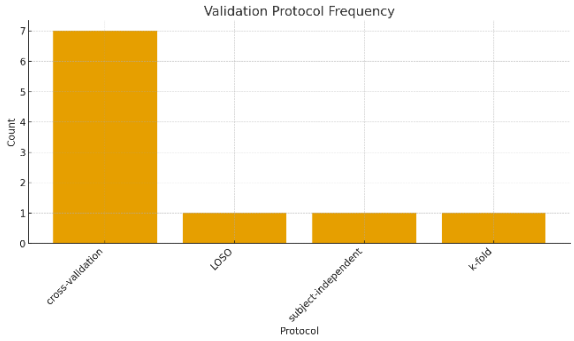



Fig. 3. Validation strategies across studies.

As illustrated in Figure 3, the dominance of k-fold cross-validation indicates that most studies prioritize performance optimization under controlled conditions rather than generalization to unseen users. Our synthesis reveals that the limited adoption of LOSO validation is not merely a methodological choice, but a key contributor to the observed lab-to-field gap. This imbalance underscores the need to reinterpret high reported accuracies with caution, particularly when claims of real-world applicability are made.

### *D. Reported Performance*

The reported performance across all studies generally falls within the 85 %–96 % range, depending on the modality and evaluation protocol used, models cluster around 90 % accuracy or equivalent F1-scores, which is strong for physiological classification tasks [3], [6], [9], [23].

However, results obtained from LOSO or cross-dataset validation are typically 5–10 % lower, underscoring the challenge of generalizing across individuals [4]. The lower performance observed under LOSO or cross-dataset validation does not indicate inferior modeling, but rather reflects a more stringent and realistic learning objective. In intra-subject evaluation, models are exposed to physiological patterns from the same individuals during training and testing, enabling them to implicitly learn subject-specific baselines and signal idiosyncrasies. This can lead to inflated performance that reflects pattern recognition of individuals rather than genuine stress-related physiological responses. In contrast, LOSO validation enforces complete subject separation, requiring models to infer stress from physiological changes that generalize across individuals with

inherently different baselines and response dynamics. Since stress manifests as relative deviations rather than absolute signal values, LOSO-trained models are compelled to capture invariant stress-related features instead of memorizing personal signal patterns. Consequently, although LOSO evaluation yields lower numerical scores, it provides a more meaningful assessment of a model's ability to detect stress rather than merely recognizing individual-specific patterns.

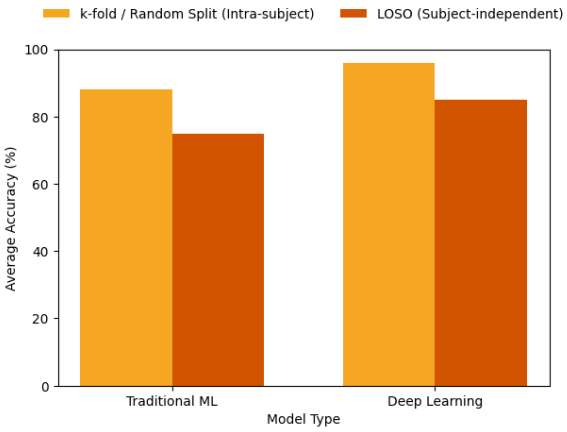



Fig. 4. Impact of validation protocol on reported stress-detection performance.

The figure summarizes average performance trends of traditional machine-learning and deep-learning models under intra-subject (k-fold) and subject-independent (LOSO) evaluation across the reviewed studies. Deep-learning approaches consistently achieve higher scores than traditional ML when trained on multimodal inputs. The combination of CNN for feature extraction and LSTM for temporal modeling remains the most successful design pattern, especially when applied to PPG and EDA data.

However, a direct comparison of these performance metrics is complicated by the significant heterogeneity across study protocols. Our analysis reveals that several factors strongly influence reported outcomes. These include the data labeling methodology (e.g., self-report vs. induced stress protocols like MIST [19], [20] ordriving simulators [6]), the signal processing details such as the length of the time segments used for analysis (e.g., 10-second segments in [19]), and the dataset characteristics, including sample size and participant diversity. For example, models validated on large, public benchmark datasets like WESAD [1][, [16], [24] may offer more generalizable insights than those trained on smaller, custom datasets [6]. These variations underscore the difficulty in establishing a single best model and highlight the critical need for standardized reporting protocols in future research.

### *E. Machine-Learning and Deep-Learning Approaches*

The reviewed papers demonstrate two major methodological generations. Early studies typically extracted handcrafted statistical and frequency-domain features, which were then classified using SVM, RF, or logistic regression. These techniques achieved accuracies in the range of 80–90 %, proving that stress can be inferred reliably from physiological data even with simple models [11]. Moreover, classical ML methods remain attractive in scenarios involving limited data, lower computational budgets, and a need for model interpretability, which is

particularly relevant for clinical or explainable-AI contexts.

Hybrid CNN–LSTM architectures have emerged as the dominant design paradigm in recent studies. From a modeling perspective, this hybridization is well aligned with the nature of physiological stress signals: CNN components act as automated feature extractors that reduce noise and encode local patterns, while LSTM layers model the sequential evolution of these features over time. When applied to multimodal inputs such as PPG, EDA, and ACC, this architecture enables both intra-modal representation learning and temporal fusion, which explains the consistently high reported accuracies between 93 % and 96 % across multiple datasets. Extensions incorporating attention mechanisms further refine this process by dynamically weighting informative signal segments or modalities, while contrastive pre-training and teacher–student knowledge distillation aim to improve robustness and data efficiency [6].

Taken together, the reviewed literature suggests that the choice between classical machine-learning and deep-learning approaches should not be guided by accuracy alone. Instead, it should reflect the intended application context, available data, and deployment constraints. Classical ML models remain suitable as strong baselines or interpretable solutions in low-resource settings, whereas deep-learning architectures represent the current state of the art for high-performance, multimodal stress detection when sufficient data and computational capacity are available.

Table V summarizes the average performance by model type. In general, deep sequential or hybrid models outperform classical methods, though they require more computational resources.

TABLE V PERFORMANCE COMPARISON BASED ON MODEL TYPE

| Model | Reference | Dataset | Performance Range (Reported) | Key Notes |
|---|---|---|---|---|
| ML Classic (SVM, RF, k-NN, GBM) | [2] | Custom (40 participate, *Public Speaking*) | Accuracy: 96.25% (2 class) | SVM (RBF) outperformed kNN, DT, RF, and MLP in feature fusion. |
| | [5] | Custom (90 subject, Office Simulation) | F1-score: 0.625 (3 class ) | LightGBM outperformed SVM/RF. Behavioral data fusion (mouse/keyboard) was better than HRV. |
| Deep Learning (Hybrid CNN-LSTM / BiLSTM) | [1] | ST Change DB, WESAD | Accuracy 98.3% (2 class ) | Time and frequency domain (spectrogram) fusion of ECG data achieves high accuracy. |
| | [19] | Custom (34 subject, MIST) | Accuracy 86.5% (3 class ) | Effective for real-time detection (10-second segments). |
| | [6] | Custom (22 participate, Driving Simulator) | Accuracy 95.5% (3 class) | Non-physiological multimodal fusion using attention has proven to be highly effective. |
| Deep Learning (CNN Multimodal Fusion) | [9] | Nursery Dataset | Accuracy 91.00% (3 class ) | The Parallel CNN architecture separates Time and Frequency domain features before fusion. |
| | [3] | Custom (26 Subject, Cortisol label) | Accuracy 91.6% (2 class ) | Using "Shuffled ECA-Net" (Attention) for feature fusion. The stress label is validated by Cortisol. |
| Deep Learning (Transformer / Multi-Task) | [18] | MuSE, OMG-Emotion | F1-score: 0.864 (2 class ) | Using Multi-Task Learning (MTL) where emotion recognition becomes an auxiliary task for stress detection. |
| | [23] | MultiAffectStress (MAS) | F1-score: 0.85 (2 class) | Using Late Fusion (Random Forest) to combine the outputs of several unimodal models (including Wav2Vec 2.0 and DistilBERT). |

### *F. Societal Feasibility and Ethical Considerations*

While high numerical accuracy remains an essential benchmark, the true success of a stress-detection model lies in its translation into everyday use. Machine-learning research is beginning to move from laboratory settings toward field deployment, yet the gap between experimental performance and societal applicability remains substantial. Several studies acknowledge that stress recognition is meaningful only when it can operate continuously, comfortably, and ethically within people's daily routines.

#### *1) Feasibility of Deployment*

Approximately one-third of the reviewed papers describe some form of prototype or pilot deployment, ranging from wrist-worn sensors to smartphone-based data collection. Wearable-centric designs particularly those relying on PPG and EDA sensors integrated in smartwatches or fitness bands emerge as the most realistic pathway for long-term stress monitoring[15].

These devices already enjoy high consumer adoption and can collect data passively without interrupting normal activity. Studies employing

multimodal fusion (EDA + PPG ± ACC/ECG) demonstrate not only technical robustness but also user comfort, battery efficiency, and signal stability during motion, all of which are prerequisites for sustainable deployment.

Conversely, models relying on EEG or ECG chest straps face usability barriers due to cumbersome electrodes and the need for skin contact. Although these sensors yield rich physiological data, their invasiveness reduces adherence outside clinical environments. A few researchers attempt to overcome these barriers through smart-textile electrodes and dry-sensor patches, indicating a promising hardware direction for future work.

*2) Real-Time and Edge-AI Integration*

Recent advances highlight the feasibility of running stress-recognition pipelines on resource-constrained devices. Several publications introduce lightweight CNN or LSTM architectures optimized for on-device inference, reporting inference times of less than one second on mobile processors [22]. This shift toward edge-AI brings multiple benefits: it enables immediate feedback for users, lowers network latency, and minimizes dependency on cloud connectivity factors crucial for emergency or occupational-safety contexts. Moreover, edge computing supports energy efficiency by processing only essential features locally and transmitting aggregated indicators instead of raw biosignals.

However, only a small fraction of current literature reports quantitative measurements of power consumption, model size, or latency, parameters that determine practical viability. Future publications should systematically include these metrics alongside accuracy to support reproducibility and engineering optimization.

*3) Summary of Feasibility Indicators*

This gap in feasibility is most critical regarding privacy and ethics. Physiological signals constitute highly sensitive personal health information, and their continuous collection raises significant user concerns over data misuse and surveillance. This review found that fewer than 15% of studies explicitly address this, often only mentioning basic anonymization. This is insufficient for real-world trust. As requested by modern data-protection laws (e.g., GDPR), the field must shift from cloud-centric processing to privacy-by-design architectures. The solution lies in the resource-efficiency models identified in this review, which enable on-device inference (Edge AI). This approach processes data locally, minimizing data transmission. For models that require continuous improvement, Federated Learning frameworks experimented with by a handful of studies offer a path forward, allowing models to be trained across distributed devices without centralizing raw data, thereby mitigating critical privacy risks.

Furthermore, our analysis highlights that technical accuracy alone is insufficient; the psychological impact and application context are paramount. Continuous stress feedback, if poorly designed, risks amplifying user anxiety rather than mitigating it. Future research must therefore bridge the gap between detection and intervention. This requires integrating psychological frameworks, such as providing Just-In-Time Adaptive Interventions (JITAI) or cognitive-behavioral prompts, transforming passive monitoring into active well-being support. In practical application contexts, such as workplace wellness programs or continuous personal health monitoring, this integration is essential. The goal is not merely to inform a user "you are stressed," but to provide an actionable, empathetic, and private pathway to improved mental resilience.

Beyond privacy and hardware, societal feasibility also involves user perception and behavioral adoption. Continuous stress feedback can empower self-awareness, yet poorly designed feedback loops risk amplifying anxiety. Few studies examine how users interpret or act upon stress predictions. Integrating psychological frameworks, such as just-in-time adaptive interventions or cognitive-behavioral prompts, could transform stress detection from passive monitoring into active well-being support.

To quantify these dimensions, each paper was scored across three observable indicators (a) use of wearable or smartphone sensors, (b) existence of a prototype or real-time system, and (c) mention of privacy or edge computing.

Overall, the evidence reveals a field that is technically sophisticated but socially nascent. To move from promising algorithms to impactful public-health tools, future research must integrate design for usability, transparency, and trust alongside continued advances in model accuracy. The ultimate benchmark for stress-detection research will not only be statistical precision but also its contribution to safer, healthier, and more empathetic human technology interaction.

This review reveals a clear methodological shift from traditional machine-learning pipelines toward deep-learning–based architectures for stress detection. As summarized in Fig. 4, deep-learning models consistently achieve higher average performance than classical approaches under both intra-subject and subject-independent evaluation. However, this advantage is accompanied by increased computational complexity, highlighting a trade-off between accuracy and deployability that must be considered in practical applications.

A key finding of this review is the substantial influence of validation strategy on reported performance. As illustrated in Fig. 4, both traditional ML and deep-learning models exhibit a consistent reduction in performance under subject-independent validation compared to intra-subject evaluation. This

pattern underscores the central role of subject inter-variability in physiological stress detection and confirms that validation methodology is a decisive factor in assessing real-world generalization capability.

Despite the observed performance trends, direct comparison across studies remains inherently limited. The reviewed literature exhibits substantial heterogeneity in datasets, stress-induction protocols, class definitions, signal preprocessing, and validation schemes. Consequently, the synthesized results should be interpreted as indicative methodological trends rather than definitive rankings of model superiority. This limitation reinforces the need for standardized reporting practices to enable more reliable comparison in future reviews.

Taken together, the findings suggest that future stress-detection research should prioritize subject-independent evaluation, multimodal sensing strategies, and deployment-aware model design. Emphasis on LOSO or cross-dataset validation, alongside transparent reporting of computational and privacy-related metrics, is essential to bridge the gap between laboratory performance and real-world applicability.

This SLR also has limitations. Our search was restricted to articles published between 2021 and 2025 to capture the most recent SOTA, which may exclude foundational papers in the field. Furthermore, due to high heterogeneity in datasets, protocols, and metrics, we performed a descriptive and thematic synthesis. A formal statistical meta-analysis was not conducted, which limits the quantitative aggregation of performance across studies.

## V. CONCLUSIONS

This Systematic Literature Review analyzed 19 studies and confirmed a clear technical state-of-the-art for stress detection: multimodal sensing (PPG, EDA, ACC) combined with hybrid CNN-LSTM models consistently yields high accuracy. The review provides a structured synthesis of current methodological trends, validation practices, and deployment considerations. The main conclusions of this study are summarized as follows.

### *A. Main Findings*

- Multimodal sensing, particularly combinations of PPG, EDA, and ACC, is the dominant and most practical configuration for wearable-based stress detection.
- Hybrid deep-learning architectures, especially CNN–LSTM models, consistently achieve higher reported performance than traditional machine-learning methods.
- Intra-subject validation (e.g., k-fold cross-validation) remains the most commonly used evaluation protocol, while subject-independent validation methods such as LOSO are still underutilized.
- Performance obtained under subject-independent validation is consistently lower but provides a more realistic estimate of real-world generalization capability.

### *B. Scientific Contributions*

This review makes the following scientific contributions:

- It provides an up-to-date synthesis of multimodal stress-detection studies published between 2021 and 2025, capturing recent advances in deep-learning–based modeling.
- It systematically highlights the impact of validation protocols on reported performance, explicitly quantifying the lab-to-field generalization gap.
- It extends conventional performance-focused reviews by integrating deployment-oriented dimensions, including computational efficiency and data privacy considerations.

### *C. Research Implications*

The findings of this review have several important implications for future research and practice:

- Reported accuracy alone is insufficient to assess model robustness; validation methodology must be considered a primary evaluation factor.
- Deployment feasibility, including model efficiency and privacy-preserving design, should be treated as first-class criteria alongside predictive performance.
- Without standardized validation and reporting practices, cross-study comparison will remain limited and potentially misleading.

### *D. Further Research Directions*

Based on the identified gaps, future research should prioritize:

- The adoption of subject-independent evaluation protocols, such as LOSO or cross-dataset validation, to ensure reliable generalization.
- The development of lightweight and energy-efficient models suitable for on-device inference and edge-AI deployment.
- The integration of privacy-by-design principles, including federated learning and local processing, to address ethical and regulatory concerns.

The connection between stress detection and intervention mechanisms, such as just-in-time adaptive interventions (JITAI), to move from passive monitoring toward actionable mental well-being support

## REFERENCES

[1] M. Kang, S. Shin, J. Jung, and Y. T. Kim, "Classification of Mental Stress Using CNN-LSTM Algorithms with Electrocardiogram Signals," *J Healthc Eng*, vol. 2021, 2021, doi: 10.1155/2021/9951905.

[2] A. Arsalan and M. Majid, "Human stress classification during public speaking using physiological signals," *Comput Biol Med*, vol. 133, p. 104377, Jun. 2021, doi: 10.1016/j.compbiomed.2021.104377.

[3] N. Kim, S. Lee, J. Kim, S. Y. Choi, and S. M. Park, "Shuffled ECA-Net for stress detection from multimodal wearable sensor data," *Comput Biol Med*, vol. 183, Dec. 2024, doi: 10.1016/j.compbiomed.2024.109217.

[4] B. Shang *et al.*, "EEG-based investigation of effects of mindfulness meditation training on state and trait by deep learning and traditional machine learning," *Front Hum Neurosci*, vol. 17, 2023, doi: 10.3389/fnhum.2023.1033420.

[5] M. Naegelin *et al.*, "An interpretable machine learning approach to multimodal stress detection in a simulated office environment," *J Biomed Inform*, vol. 139, Mar. 2023, doi: 10.1016/j.jbi.2023.104299.

[6] L. Mou *et al.*, "Driver stress detection via multimodal fusion using attention-based CNN-LSTM," *Expert Syst Appl*, vol. 173, Jul. 2021, doi: 10.1016/j.eswa.2021.114693.

[7] Y. S. Can, B. Arnrich, and C. Ersoy, "Stress detection in daily life scenarios using smart phones and wearable sensors: A survey," Apr. 01, 2019, *Academic Press Inc.* doi: 10.1016/j.jbi.2019.103139.

[8] S. S. Panicker and P. Gayathri, "A survey of machine learning techniques in physiology based mental stress detection systems," Apr. 01, 2019, *Elsevier Sp. z o.o.* doi: 10.1016/j.bbe.2019.01.004

[9] J. Z. Xiang, Q. Y. Wang, Z. Bin Fang, J. A. Esquivel, and Z. X. Su, "A multi-modal deep learning approach for stress detection using physiological signals: integrating time and frequency domain features," *Front Physiol*, vol. 16, 2025, doi: 10.3389/fphys.2025.1584299.

[10] S. M. Et. al., "Mental Health Prediction Models Using Machine Learning in Higher Education Institution," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 5, pp. 1782–1792, Apr. 2021, doi: 10.17762/turcomat.v12i5.2181.

[11] D. Ghose, A. Chatterjee, I. A. M. Balapuwaduge, Y. Lin, and S. P. Dash, "Investigating lightweight and interpretable machine learning models for efficient and explainable stress detection," *Front Digit Health*, vol. 7, 2025, doi: 10.3389/fdgth.2025.1523381.

[12] A. Almadhor *et al.*, "Wrist-Based Electrodermal Activity Monitoring for Stress Detection Using Federated Learning," *Sensors*, vol. 23, no. 8, Apr. 2023, doi: 10.3390/s23083984.

[13] M. J. Page *et al.*, "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," Mar. 29, 2021, *BMJ Publishing Group*. doi: 10.1136/bmj.n71.

[14] D. Cabrera, L. Cabrera, and E. Cabrera, "The Steps to Doing a Systems Literature Review (SLR)," Apr. 06, 2023. doi: 10.54120/jost.pr000019.v1.

[15] S. Hosseini *et al.*, "A multimodal sensor dataset for continuous stress detection of nurses in a hospital," *Sci Data*, vol. 9, no. 1, Dec. 2022, doi: 10.1038/s41597-022-01361-y.

[16] S. R. Kadu, S. Sadruddin, and D. Argade, "A Multimodal Fusion for Physiological Sensor and Audio Signal-based Stress Detection Using Sliding Hierarchical Bidirectional Recurrent Neural Network," *International Journal of Intelligent Engineering and Systems*, vol. 18, no. 3, pp. 569–582, 2025, doi: 10.22266/ijies2025.0430.39.

[17] S. Yang *et al.*, "A deep learning approach to stress recognition through multimodal physiological signal image transformation," *Sci Rep*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-01228-3.

[18] Y. Yao, M. Papakostas, M. Burzo, M. Abouelenien, and R. Mihalcea, "MUSER: MUltimodal Stress Detection using Emotion Recognition as an Auxiliary Task," May 2021, [Online]. Available: http://arxiv.org/abs/2105.08146

[19] P. Zhang *et al.*, "Real-time psychological stress detection according to ECG using deep learning," *Applied Sciences (Switzerland)*, vol. 11, no. 9, May 2021, doi: 10.3390/app11093838.

[20] J. Zhang, H. Yin, J. Zhang, G. Yang, J. Qin, and L. He, "Real-time mental stress detection using multimodality expressions with a deep learning framework," *Front Neurosci*, vol. 16, Aug. 2022, doi: 10.3389/fnins.2022.947168.

[21] S. A. M. Mane and A. Shinde, "StressNet: Hybrid model of LSTM and CNN for stress detection from electroencephalogram signal (EEG)," *Results in Control and Optimization*, vol. 11, Jun. 2023, doi: 10.1016/j.rico.2023.100231

[22] L. Fontes, P. Machado, D. Vinkemeier, S. Yahaya, J. J. Bird, and I. K. Ihianle, "Enhancing Stress Detection: A Comprehensive Approach through rPPG Analysis and Deep Learning Techniques," *Sensors*, vol. 24, no. 4, Feb. 2024, doi: 10.3390/s24041096.

[23] D. Ghose, O. Gitelson, and B. Scassellati, "Integrating Multimodal Affective Signals for Stress Detection from Audio-Visual Data," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Nov. 2024, pp. 22–32. doi: 10.1145/3678957.3685717.

[24] S. Yang *et al.*, "A deep learning approach to stress recognition through multimodal physiological signal image transformation," *Sci Rep*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-01228-3.

[25] Z. Yang, H. Yu, and A. Sano, "Contrastive Pretraining for Stress Detection with Multimodal Wearable Sensor Data and Surveys," 2025. [Online]. Available: https://github.com/comp-well-

[26] T. Jeon, H. Byeol Bae, and S. Lee, "Multimodal Stress Recognition Using a Multimodal Neglecting Mask Module," *IEEE Access*, vol. 12, pp. 144774–144787, 2024, doi: 10.1109/ACCESS.2024.3469575.

[27] R. Tanwar, O. C. Phukan, G. Singh, P. K. Pal, and S. Tiwari, "Attention based hybrid deep learning model for wearable based stress recognition," *Eng Appl Artif Intell*, vol. 127, Jan. 2024, doi: 10.1016/j.engappai.2023.107391

ISSN 2085-4552

# Comparative Modeling of Naïve Bayes and LSTM with Monte Carlo Forecasting for Silver Prices

Firda Fadri[1], Muhammad Amjad Munjid[2], Muhammad Azmi Alauddin[3]
[1,2,3] Faculty of Mathematics and Natural Sciences, Jember University, Jember, Indonesia
[1] firdafadri@unej.ac.id, [2] zok6683@gmail.com, [3] muhammadzmi29@gmail.com

Accepted 04 January 2026
Approved 09 January 2026

***Abstract*— Silver price volatility has increased markedly in recent years, particularly since 2023, driven by growing demand from the renewable energy sector. This study compared two conceptually distinct forecasting approaches: Naïve Bayes (NB), which relies on conditional independence assumptions, and Long Short-Term Memory (LSTM), which models temporal dependencies in time-series data. Daily silver price data (USD/troy ounce) from January 1989 to October 2025 were analyzed. NB was implemented using a single lagged price feature (t−1), while LSTM employed a two-layer architecture with 50 units, 0.2 dropout, and a 60-day sequential window. Empirical results showed that NB achieved $R^2$ of 0.9818, reproducing dominant price dynamics but exhibiting slight lagging during sharp price movements. In contrast, LSTM achieved lower RMSE and MAE, with an $R^2$ of 0.9939, effectively capturing nonlinear dependencies and volatility patterns. When extended with Monte Carlo simulation, LSTM enabled uncertainty-aware short-term forecasting, providing median price trajectories and prediction intervals, making it a more robust framework for silver price prediction under extreme volatility.**

***Index Terms*— *silver price; Naïve Bayes; LSTM; Monte Carlo.***

## I. INTRODUCTION

Silver is one of the key precious metal commodities in the global economic system, functioning both as an investment asset and an industrial raw material. Its price dynamics often move in tandem with gold and serve as an alternative investment during periods of economic uncertainty [1]. However, silver prices are well known for their sharp fluctuations, driven by macroeconomic factors such as inflation, interest rates, exchange rates, and rapidly changing industrial demand [2]. These characteristics make accurate silver price forecasting particularly important for investors, financial institutions, and policymakers in formulating investment strategies and managing economic risk [3].

In line with the increasing volatility of commodity markets, various analytical approaches have been developed to understand and forecast silver prices. Conventional statistical methods such as ARIMA and Exponential Smoothing have been widely applied, but their reliance on linearity and stationarity assumptions often limits their effectiveness when applied to highly volatile and nonlinear commodity price data [4]. As a result, machine learning-based approaches have gained attention due to their ability to capture complex patterns and nonlinear relationships without requiring strict parametric assumptions [5].

Among machine learning methods, Naïve Bayes and Long Short-Term Memory (LSTM) represent two fundamentally different modeling philosophies. Naïve Bayes is a probabilistic classifier grounded in Bayes' Theorem and is valued for its simplicity and computational efficiency [6]. LSTM is a development of Recurrent Neural Network (RNN) designed to recognize patterns and long-term dependencies in sequential data, thus making it highly effective for analyzing time series data such as silver prices [7]. Despite these differences, both methods continue to be used in empirical studies, often motivated by trade-offs between model complexity, interpretability, and computational cost.

The Long Short-Term Memory (LSTM) method has been widely applied across various domains, including stock market analysis and prediction, cryptocurrency price forecasting such as Dogecoin, indoor air quality modeling, and railway transportation operations environments [8, 9, 10, 11]. Meanwhile, Naïve Bayes has also been applied in weather forecasting studies, where temperature, humidity, and wind speed are commonly used as predictor features, demonstrating its practicality as a probabilistic baseline despite the strong independence assumptions [12]. These studies illustrated the broader applicability and contrasting strengths of probabilistic and machine learning-based models, particularly in terms of accuracy, robustness, and computational efficiency.

Many studies emphasize performance improvement without explicitly examining whether complex deep learning models provide meaningful advantages over simpler probabilistic baselines when applied to long-horizon commodity data, while comparative analyses rarely consider extended historical periods encompassing multiple volatility regimes, including recent episodes of extreme price fluctuations. In addition, limited attention has been given to the practical implications of model behavior during trend reversals and high-volatility phases, which are critical for real-world decision-making. Addressing these gaps, this study goes beyond a simple comparison of predictive accuracy by evaluating how Naïve Bayes and LSTM differ in capturing temporal dynamics, responsiveness to trend changes, and predictive stability when applied to long-horizon silver price data. Using daily world silver prices over an extended period, this research assesses the practical suitability of probabilistic versus deep learning approaches for forecasting silver prices under nonlinear and volatile market conditions, thereby contributing to a more nuanced understanding of model selection in commodity price forecasting.

## II. METHOD

To address the research question regarding the comparative performance of Naïve Bayes and LSTM in predicting silver prices, this study adopts a comparative experimental design, where both models are trained and evaluated in parallel on an identical dataset, ensuring a controlled and fair comparison. The workflow includes: collection and validation of historical data, temporal preprocessing, construction of two distinct model architectures, training, objective evaluation using regression metrics, and interpretation of performance in the context of the temporal characteristics of silver price series. An overview of the research workflow is presented in Fig. 1.

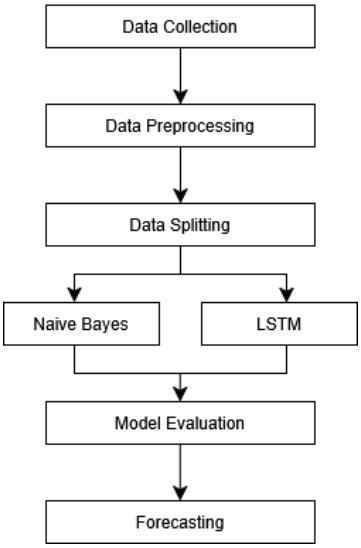



Fig. 1. Research Design

### A. Data Collection

Historical silver price data (USD per troy ounce) were obtained from Investing.com, covering daily observations from January 3, 1989 to October 30, 2025, resulting in 9,384 observations. This data source was selected because market holidays are handled consistently through forward-filling, ensuring a continuous time series and preventing temporal bias [1]. The dataset contains two main columns: *Date* and *Price*, exported to CSV format and loaded into the Python environment using pandas with parameters *parse_dates=['Date']* and *index_col='Date'* to ensure chronological ordering and compatibility with time series analysis tools (e.g., *statsmodels* and *scikit-learn*).

### B. Data Preprocessing

Data preprocessing was conducted to transform the raw silver price series into a suitable representation for machine learning–based regression models. Since Investing.com applies forward-filling to account for non-trading days, no missing values were expected; nevertheless, explicit checks were performed using *df.isnull().sum()* to confirm data completeness, along with data type inspection *(df.dtypes)* to ensure numerical consistency. The dataset was chronologically ordered to preserve the temporal structure required for time-series modeling.

Prior to feature construction, an autocorrelation function (ACF) and partial autocorrelation function (PACF) analysis was conducted exclusively on the training dataset to avoid information leakage. The ACF results indicate statistically significant short-term autocorrelation, while the PACF exhibits a clear cutoff after the first lag, suggesting that short-term temporal dependence dominates the silver price dynamics. Based on this empirical evidence, a lagged price feature ($Price_{t-1}$) was constructed to represent short-term memory in the Naïve Bayes model. Rows containing missing values introduced by the shift operation were removed, yielding a clean and temporally aligned dataset suitable for supervised learning.

The dataset was divided using an 80:20 temporal split without shuffling to prevent look-ahead bias. The training set consists of 7,505 observations spanning January 3, 1989 to December 31, 2023, while the testing set contains 1,877 observations from January 2, 2024 to October 30, 2025, a period characterized by heightened price volatility. In contrast to the Naïve Bayes model, the LSTM model does not rely on explicitly defined lag variables. Instead, temporal dependence is captured implicitly through fixed-length input sequences. In this study, the data were reshaped into 60-day sequences (n_steps = 60) using a sliding-window approach, where each input sequence consists of silver prices from the previous 60 trading days to predict the subsequent day's price. This sequence-based representation enables the LSTM to learn nonlinear temporal patterns and medium-term dependencies directly from historical

price trajectories without requiring manual lag selection.

### C. Naïve Bayes Model

Naïve Bayes was employed in this study as a computationally efficient probabilistic baseline for modeling silver price dynamics. Although Naïve Bayes is conventionally designed for classification tasks, it has been previously adapted for regression-oriented problems through discretization or probabilistic approximation, particularly in exploratory or comparative modeling contexts [13]. The silver price distribution exhibits moderate positive skewness (skewness = 0.8549), indicating asymmetry and deviations from strict normality. While this level of skewness does not represent an extreme heavy-tailed distribution, it suggests that Gaussian assumptions may not be fully satisfied across the entire price range, thereby motivating the exploration of a probabilistic modeling framework that is less sensitive to distributional symmetry.

To accommodate the continuous nature of silver prices, the target variable was discretized into a finite number of intervals prior to model training. Discretization enables the Naïve Bayes classifier to approximate regression behavior by estimating the posterior probability of future price intervals conditional on historical observations [14]. In this study, historical price information from the preceding three trading days was used as input features, and conditional independence among lagged variables was assumed. While this assumption is strong and may not fully hold in financial time series, it allows the model to remain analytically tractable and computationally efficient, serving its intended role as a baseline comparator rather than a primary predictive model.

The Naïve Bayes model estimates the posterior probability of a price interval $C$ given observed historical prices $X$ as Equation (1).

$$P(C \mid X) = \frac{P(X \mid C) \cdot P(C)}{P(X)} \quad (1)$$

with:

$P(C \mid X)$ : posterior probability of class C given evidence X
$P(X \mid C)$ : probability that evidence X is assigned class C
$P(C)$ : prior probability of class C
$P(X)$ : probability of evidence X

For numerical data, likelihood calculations can use the Gaussian distribution [15].

### D. LSTM Model

The Long Short-Term Memory (LSTM) network is an advanced recurrent neural network architecture introduced by Hochreiter and Schmidhuber to address the vanishing gradient problem commonly encountered in standard Recurrent Neural Networks (RNNs) when modeling long sequential data [16]. Conventional RNNs process time-ordered data recursively but often fail to retain long-term dependencies due to exponentially diminishing gradients during backpropagation [17]. LSTM mitigates this limitation by incorporating a memory cell regulated by three gating mechanisms such as input, forget, and output gates, which collectively control the storage, update, and release of information over time, enabling more stable learning of temporal dependencies in volatile financial time series such as silver prices [18].

The internal operations of the LSTM unit are governed by gating equations that regulate information flow within the memory cell. These mechanisms are mathematically expressed through the input gate, forget gate, cell state update, and output gate formulations, which are summarized in Equations (2)–(7). Through these equations, the LSTM learns to selectively preserve relevant historical information while discarding noise, enabling robust modeling of nonlinear temporal dependencies in long-horizon silver price forecasting.

$$i_t = \sigma(W_i x_t + W_{hi} h_{t-1} + b_i) \quad (2)$$
$$f_t = \sigma(W_f x_t + W_{hf} h_{t-1} + b_f) \quad (3)$$
$$\bar{C}_t = \tanh(W_c x_t + W_{hc} h_{t-1} + b_c) \quad (4)$$
$$C_t = f_t \times C_{t-1} + i_t \times \bar{C}_t \quad (5)$$
$$o_t = \sigma(W_o x_t + W_{ho} h_{t-1} + b_o) \quad (6)$$
$$h_t = o_t \times \tanh(C_t) \quad (7)$$

with:

$i_t$ : input gate,
$f_t$ : forget gate,
$o_t$ : output gate,
$C_t$ : cell state,
$h_t$ : hidden state,
$x_t$ : input at time-t,
$W$ : network weight, and
b : bias

In this study, the LSTM was implemented using a sequence-based forecasting framework, with the silver price series reshaped into fixed-length input sequences of 60 trading days (n_steps = 60), where each sequence predicted the subsequent day's price. The model architecture consisted of two stacked LSTM layers with a tunable number of hidden units (50 or 100) and dropout layers (rate = 0.2) to reduce overfitting, followed by a fully connected layer to transform the LSTM features into predicted prices. A grid search was performed over the number of units per layer, learning rate (0.001, 0.0005), batch size (64), and epochs (200, with early stopping), and the optimal model was selected based on the lowest root mean squared error (RMSE) on the validation set. Training employed the Adam optimizer with the selected learning rate, mean squared error (MSE) as the loss function, and mean absolute error (MAE) as an auxiliary metric, while early stopping ensured generalization and prevented

overfitting. The resulting model was then used for both test set prediction and short-term Monte Carlo forecasting.

### *E. Model Evaluation*

To assess the predictive performance of Naïve Bayes and LSTM, given the silver price series characteristics such as volatility clustering, regime-switching, and heavy-tailed distribution evaluation was conducted using three complementary regression metrics: MAE, RMSE, and $R^2$ [19]. These metrics jointly capture different aspects of prediction error: absolute accuracy (MAE), sensitivity to large deviations (RMSE), and the proportion of variance explained by the model ($R^2$).

1. MAE (Mean Absolute Error)

   MAE computes the average absolute deviation between predictions and observed values, treating small and large errors equally [20]. Unlike RMSE, which squares errors and thus penalizes large deviations more heavily, MAE is more robust in the presence of outliers such as silver price spikes due to geopolitical turmoil or supply shocks [21]. In practice, models with lower MAE generally yield more consistently accurate predictions on average, though they may not excel at capturing extremes. The MAE formula using Equation (8).

$$MAE = \frac{1}{n} \times \sum |y_i - \hat{y}_i| \qquad (8)$$

   with:
   $n$ : amount of data,
   $y_i$ : true value,
   $\hat{y}_i$ : model prediction value.

2. RMSE (Root Mean Squared Error)

   RMSE is a standard performance indicator in numerical prediction evaluation, especially in financial time series studies, due to its sensitivity to large errors making it an informative early signal of potential underfitting or over-smoothing. RMSE quantifies the magnitude of the difference between model predictions and actual observations [22]. As a widely used technique, RMSE helps assess the level of error in numerical prediction models. It is derived from the square root of the mean squared prediction errors. A decreasing RMSE generally reflects improved prediction accuracy particularly for trend direction changes though interpretation should consider residual patterns, as an extremely low RMSE may indicate overfitting or data leakage. Prediction accuracy is determined by the smallest error value across evaluation methods [23]. The RMSE formula using Equatiaon (9).

$$RMSE = \sqrt{\frac{1}{n} \times \sum (y_i - \hat{y}_i)^2} \qquad (9)$$

3. $R^2$ (The coefficient of determination)

   $R^2$ indicates the proportion of actual data variance explained by the model, ranging from 0 (no explanatory power) to 1 (perfect explanation) [16]. Although a high in sample $R^2$ is often considered a success marker, in out-of-sample evaluation, an $R^2$ close to 1 does not necessarily guarantee generalization abilityespecially if residuals still exhibit autocorrelation or systematic patterns [20]. The $R^2$ formula using Equation (10).

$$R^2 = 1 - \left( \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \hat{y}_i)^2} \right) \qquad (10)$$

.

## III. RESULT AND DISCUSSIONS

### *A. Naïve Bayes Model Prediction Results*

The autocorrelation function (ACF) and partial autocorrelation function (PACF) analyses were conducted exclusively on the training dataset to determine the optimal lag structure while avoiding information leakage. As illustrated in Fig. 2 and Fig. 3, the silver price series exhibited a clear AR(1)-type behavior. The PACF plot showed a single statistically significant spike at Lag 1 followed by an immediate cutoff into the insignificance region, indicating that only the most recent past observation has a direct and meaningful influence on the current price level. This pattern provideed strong empirical justification for selecting Lag 1 as the primary explanatory feature in the Naïve Bayes model.

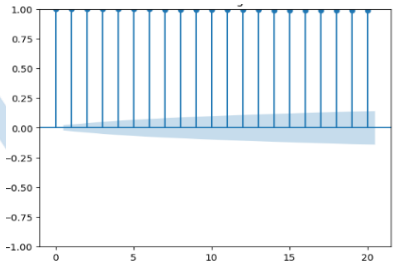



Fig. 2. ACF for training data

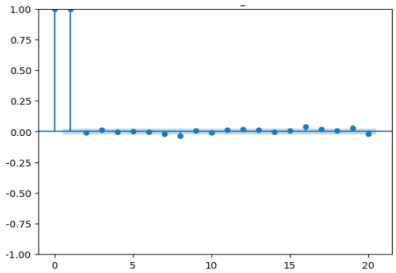



Fig. 3. PACF for training data

The ACF plot exhibited a slow and gradual decay across multiple lags, indicating strong persistence and a potentially non-stationary structure in silver prices, while the PACF showed a clear cutoff after Lag 1, justifying the selection of a single lag as the most parsimonious feature for Naïve Bayes modeling.

Although such non-stationarity might have violated classical linear time-series assumptions, it did not invalidate lag-based features in supervised learning, and differencing was intentionally not applied to preserve the original price scale and ensure a fair comparison with the LSTM model, which could learn non-stationary patterns implicitly. Based on this empirical evidence, a Lag-1 Naïve Bayes specification was adopted and yielded strong predictive performance, as summarized in Table 1, with RMSE = 0.9180, MAE = 0.7042, and $R^2$ = 0.9818, indicating that short-term price dependence alone captures a substantial proportion of the variability in silver prices under long-horizon and volatile market conditions.

TABLE 1. NAÏVE BAYES MODEL EVALUATION

| Model | RMSE | MAE | $R^2$ |
|---|---|---|---|
| Naïve Bayes | 0.9180 | 0.7042 | 0.9818 |

Fig. 4 further supported this conclusion by comparing the actual silver prices with the Lag-1 Naïve Bayes predictions over the test period. The predicted series closely followed the overall trajectory of the observed prices, demonstrating that immediate past information is sufficient to reproduce the dominant price dynamics. Minor deviations appeared during episodes of sharp price acceleration and extreme volatility, particularly toward the end of the sample, where the model exhibits slight lagging behavior. This smoothing effect reflected the inherent limitation of a probabilistic baseline model relying on conditional independence assumptions.

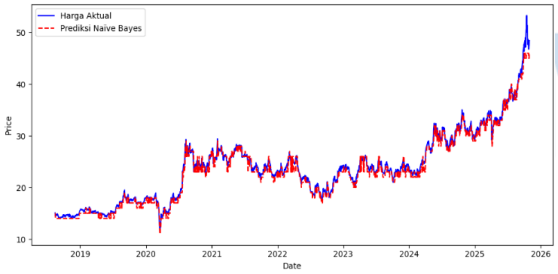



Fig. 4. Actual vs. predicted silver prices by Naïve Bayes model (Lag-1)

## B. *LSTM Model Prediction Results*

The LSTM model was constructed using a two-layer architecture with 50 units per layer, a dropout rate of 0.2, and two dense layers (25 and 1 unit). The silver price data were normalized using a Min-Max Scaler and reshaped into 60-day input sequences (n_steps = 60) to predict the subsequent day's price. Hyperparameter optimization was conducted via grid search over the number of units, learning rate (0.001, 0.0005), batch size (64), and epochs (200), with early stopping applied based on validation loss, and the best model was selected using the lowest RMSE. The optimal configuration consisted of 2 LSTM layers, 50 units, dropout 0.2, batch size 64, learning rate 0.0005, and 200 epochs. Evaluation on the test data demonstrated strong predictive performance with RMSE = 0.5277, MAE = 0.3493, and $R^2$ = 0.9939, indicating that the model explains 99.39% of the variance in actual prices, as summarized in Table 2.

TABLE 2. LSTM MODEL EVALUATION

| Model | RMSE | MAE | $R^2$ |
|---|---|---|---|
| LSTM | 0.5277 | 0.3493 | 0.9939 |

Visualization in Fig. 5 showed that the LSTM predictions closely track the actual silver price series, even during periods of high volatility. This demonstrated that the model effectively captures nonlinear trends and short- to medium-term temporal dependencies in the data, indicating strong predictive performance and the ability to follow rapid price movements in the market.

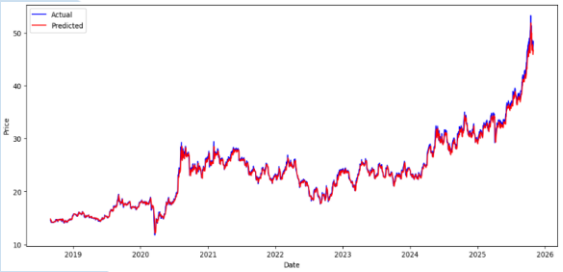



Fig. 5. Actual vs. predicted silver prices by LSTM model

## C. *Performance Comparison: Naïve Bayes vs. LSTM*

Based on Table 3, LSTM was the most superior model overall. This conclusion was drawn from the fact that its prediction errors were lower and it explained a higher proportion of the variance in the data, with an $R^2$ of 0.9939, compared to Naïve Bayes. Using only MAE would not have capture the impact of extreme prediction errors, while using only RMSE could have overemphasize outliers. Therefore, considering both metrics together, along with the high $R^2$, provided a more balanced and informative assessment, confirming LSTM's better predictive performance.

TABLE 3. EVALUATION OF BOTH MODELS ON THE SAME TEST SET

| Model | RMSE | MAE | $R^2$ |
|---|---|---|---|
| Naïve Bayes | 0.9180 | 0.7042 | 0.9818 |
| LSTM | 0.5277 | 0.3493 | 0.9939 |

## D. *Short-Term Forecasting Using Monte Carlo LSTM*

The short-term forecasting results using Monte Carlo LSTM for the next 90 days (Figure 6) showed historical silver prices in blue, while the red line represented the model's median predictions. The

transparent red area depicted the 10%-90% prediction interval (PI), indicating the range of price uncertainty with an 80% probability. The visualization suggested that silver prices were expected to moderately decline from 48.44 USD to a median of 43.84 USD by day 90, although price volatility remained high.

The purple and green dashed lines marked the train/test split and the start of the forecast period, making it easier to distinguish historical data from predictions. The prediction interval on day 90 ranged from 40.94–47.46 USD, indicating that actual prices could deviate from the median forecast. Overall, Monte Carlo LSTM effectively captured historical trends while providing useful uncertainty information for short-term risk planning.

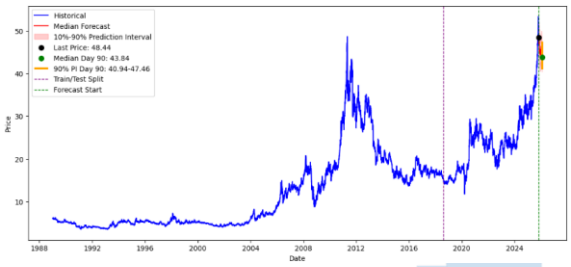



Fig. 6. Forecasting Results

## IV. CONCLUSIONS

This study evaluated the efficacy of Naïve Bayes and LSTM models in forecasting silver prices. The Lag-1 Naïve Bayes model captured short-term price dependence and reproduced the dominant price dynamics, yielding strong predictive performance ($R^2$ = 0.9818), although it exhibited slight lagging behavior during episodes of rapid price acceleration and extreme volatility. In contrast, the LSTM model effectively tracked nonlinear trends and short- to medium-term temporal dependencies, achieving lower prediction errors and superior accuracy with an $R^2$ of 0.9939. The integration of Monte Carlo simulations extended the LSTM framework into uncertainty-aware short-term forecasting, providing both median price trajectories and 10%-90% prediction intervals that quantified forecast uncertainty. Overall, while Naïve Bayes provided a computationally efficient baseline, LSTM significantly outperformed it, particularly under volatile market conditions, confirming its robustness and suitability for capturing complex silver price dynamics and supporting strategic financial risk planning.

## REFERENCES

[1] A. K. Pradhan, B. R. Mishra, A. K. Tiwari, and S. Hammoudeh, "Macroeconomic factors and frequency domain causality between Gold and Silver returns in India," *Resources Policy*, vol. 68, 2020, Art. no. 101744. doi: 10.1016/j.resourpol.2020.101744.

[2] D. Istiyanti, R. Hutri Almufarid, D. Yuniarti, and Sifriyani, "Penerapan model VAR untuk peramalan harga perak berjangka dan suku bunga BI di Indonesia," in *Prosiding Seminar Nasional Matematika, Statistika, dan Aplikasinya*, 2025, pp. 83–95.

[3] D. N. Gono, H. Napitupulu, and Firdaniza, "Silver price forecasting using Extreme Gradient Boosting (XGBoost) method," *Mathematics*, vol. 11, no. 18, 2023. doi: 10.3390/math11183813.

[4] T. Prasetyo, R. A. Putri, D. Ramadhani, Y. Angraini, and K. A. Notodiputro, "Perbandingan kinerja metode ARIMA, Multi-Layer Perceptron, dan Random Forest dalam peramalan harga logam mulia berjangka yang mengandung pencilan," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, no. 2, pp. 265–274, 2024. doi: 10.25126/jtiik.20241127392.

[5] N. R. Febriyanti, K. Kusrini, and A. D. Hartanto, "Analisis perbandingan algoritma SVM, Random Forest dan Logistic Regression untuk prediksi stunting balita," *Edumatic: Jurnal Pendidikan Informatika*, vol. 9, no. 1, pp. 149–158, 2025. doi: 10.29408/edumatic.v9i1.29407.

[6] M. A. Q. Dani, C. A. Pratama, I. Raihan, and I. Kharisudin, "PRISMA, Prosiding Seminar Nasional Matematika Pemodelan Runtun Waktu Harga Nikel dengan algoritma LSTM dan GRU," *PRISMA*, vol. 8, pp. 392–398, 2025. doi: https://journal.unnes.ac.id/sju/index.php/prisma/

[7] R. Muhammad and I. Nurhaida, "Penerapan LSTM dalam deep learning untuk prediksi harga kopi jangka pendek dan jangka panjang," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 10, no. 1, pp. 554–564, 2025. doi: 10.29100/jipi.v10i1.5904.

[8] Z. Li, H. Yu, J. Xu, J. Liu, and Y. Mo, "Stock market analysis and prediction using LSTM: A case study on technology stocks," *IAET*, vol. 2, no. 1, pp. 1–6, Nov. 2023.

[9] Y. Wei, J.-J. Jang-Jaccard, W. Xu, F. Sabrina, S. Camtepe, and M. Boulic, "LSTM-autoencoder-based anomaly detection for indoor air quality time-series data," *IEEE Sensors Journal*, vol. 23, no. 4, pp. 3787–3800, 2023.

[10] Y. Wang, X. Du, Z. Lu, Q. Duan, and J. Wu, "Improved LSTM-based time-series anomaly detection in rail transit operation environments," *IEEE Trans. Ind. Informat.*, vol. 18, pp. 1–1, 2022. doi: 10.1109/TII.2022.3164087.

[11] R. Sari, A. S. Saruman, and F. E. Susilawati, "Prediksi harga Dogecoin dengan menggunakan metode Long Short-Term Memory (LSTM)," in *Seminar Nasional Teknologi Informasi dan Komputer (SEMANTIK)*, 2023, pp. 164–172.

[12] P. Biradar, S. Ansari, Y. Paradkar, and S. Lohiya, "Weather prediction using data mining," *IJEDR*, vol. 5, no. 2, pp. 211–214, 2017.

[13] D. Syafira, "Analysis of the Naïve Bayes classifier method in classifying the weather conditions in Aceh Tamiang," *Journal of Advanced Computer Knowledge and Algorithms*, vol. 1, pp. 47–51, 2024.

[14] N. Widiastuti, A. Hermawan, and D. Avianto, "Implementasi metode Naïve Bayes untuk klasifikasi data blogger," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 8, no. 3, pp. 985–994, 2023. doi: 10.29100/jipi.v8i3.3713.

[15] F. Ristianto, Nurmalasari, and A. Yoraeni, "Implementasi metode Naïve Bayes untuk prediksi harga emas," *Computer Science (CO-SCIENCE)*, vol. 1, no. 1, pp. 62–71, 2021.

[16] M. D. A. Carnegie and C. Chairani, "Perbandingan Long Short Term Memory (LSTM) dan Gated Recurrent Unit (GRU) untuk memprediksi curah hujan," *Jurnal Media Informatika Budidarma*, vol. 7, no. 3, pp. 1022, 2023. doi: 10.30865/mib.v7i3.6213.

[17] M. Muharrom, "Analisis komparasi algoritma data mining Naïve Bayes, K-Nearest Neighbors dan regresi linier dalam prediksi harga emas," *Bulletin of Information Technology (BIT)*, vol. 4, no. 4, pp. 430–438, 2023.

[18] H. N. Bhandari, B. Rimal, N. R. Pokhrel, R. Rimal, K. R. Dahal, and R. K. C. Khatri, "Predicting stock market

index using LSTM," *Machine Learning with Applications*, vol. 9, p. 100320, Sep. 2022, doi: 10.1016/j.mlwa.2022.100320.

[19] A. F. Salsabilah, A. A. Hanafi, M. S. Nurilhaq, and P. D. Wira, "Implementasi algoritma regresi linier untuk memprediksi harga emas," *Jurnal INTRO (Informatika dan Teknik Elektro)*, vol. 3, no. 2, pp. 71–77, 2024.

[20] C. Alkahfi, A. Kurnia, and A. Saefuddin, "Perbandingan kinerja model berbasis RNN pada peramalan data ekonomi dan keuangan Indonesia," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 4, pp. 1235–1243, 2024. doi: 10.57152/malcom.v4i4.1415.

[21] A. F. Salsabilah, A. A. Hanafi, M. S. Nurilhaq, and P. G. D. Wira, "Implementasi algoritma regresi linear untuk memprediksi harga emas," *Jurnal INTRO (Informatika dan Teknik Elektro)*, vol. 3, no. 2, 2024.

[22] R. N. Azizah, U. K. Nisak, and U. Indahyanti, "Analisis jumlah prediksi penyebaran HIV/AIDS di Kabupaten Sidoarjo menggunakan metode multiple linear regression," *Physical Sciences, Life Science and Engineering*, vol. 1, no. 1, pp. 11, 2024. doi: 10.47134/pslse.v1i1.163.

[23] M. Muharrom, "Analisis komparasi algoritma data mining Naïve Bayes, K-Nearest Neighbors dan regresi linier dalam prediksi harga emas," *Bulletin of Information Technology (BIT)*, vol. 4, no. 4, pp. 430–438, 2023. doi: 10.47065/bit.v3i1

ISSN 2085-4552

# Design and Evaluation of an AI-Driven Gamified Intelligent Tutoring System for Fundamental Programming Using the Octalysis Framework

Dzaky Fatur Rahman[1], Fenina Adline Twince Tobing[2], Cian Ramadhona Hassolthine[3]
[1,2] Informatics Department, Universitas Multimedia Nusantara, Tangerang, Indonesia
[3]PJJ Informatika, Universitas Siber Asia, Jakarta, Indonesia
email: dzaky.fatur@student.umn.ac.id [1], fenina.tobing@umn.ac.id[2], cianhassolthine@lecturer.unsia.ac.id[3]

Accepted 04 January 2026
Approved 21 January 2026

***Abstract*— Gamification and Intelligent Tutoring Systems (ITS) are independently proven to enhance student engagement, yet their integration remains fragmented in fundamental programming education. Existing systems often feature either robust gamification with static content or adaptive AI with limited motivational design, leaving a gap in comprehensive frameworks that address both intrinsic motivation and cognitive support simultaneously. This study designs and evaluates "Starcoder," an adaptive learning environment that synthesizes the Octalysis Framework with a Generative AI-driven ITS. The system architecture integrates the Next.js framework with the Gemini AI API to deliver real-time, context-aware feedback and remedial learning paths. The study employs a Research and Development (R&D) methodology, culminating in a quantitative evaluation using the Hedonic-Motivation System Adoption Model (HMSAM). Data from 54 undergraduate respondents were analyzed using descriptive statistics to measure seven core constructs of user experience. The results indicate a positive reception of the integrated approach, with the platform achieving high approval ratings in Perceived Usefulness (86.44%) and Curiosity (85.56%). Comparative analysis reveals a notable increase in Behavioral Intention to Use (+15.56%) relative to the traditional classroom baseline. These findings suggest that coupling narrative-driven gamification with adaptive AI agents effectively fosters student engagement, offering a promising model for next-generation educational technologies.**

***Index Terms*— *Artificial Intelligence (AI); Gamification; Hedonic-Motivation System Adoption Model (HMSAM); Intelligent Tutoring System (ITS); Octalysis Framework.***

## I. INTRODUCTION

Fundamental programming education presents significant pedagogical challenges, particularly in the initial years of informatics study. The curriculum demands a grasp of abstract concepts such as algorithms, logic, and data structures, which are often difficult for novice learners to visualize [1]. Traditional instructional methods, characterized by linear delivery and rigid classroom structures, frequently struggle to foster the intrinsic motivation and engagement necessary for mastering these complex skills [2]. Consequently, students often experience demotivation and a lack of active participation, leading to suboptimal learning outcomes [3].

To address these engagement deficits, gamification has emerged as a powerful pedagogical strategy. Recent meta-analyses indicate that integrating game mechanics into educational settings significantly enhances student motivation and behavioral engagement [4]. Among various gamification models, the Octalysis Framework has proven particularly effective in educational contexts by leveraging eight core drives of human motivation to create immersive learning experiences [5]. However, while gamification improves engagement, it does not inherently address the need for personalized cognitive support. This is where Intelligent Tutoring Systems (ITS) become critical. AI-driven ITS can provide adaptive feedback and tailored learning paths, which have been shown to consistently improve academic achievement by addressing individual learning gaps [6].

Despite the proven efficacy of both gamification and ITS independently, the seamless integration of these two paradigms remains a significant research gap. A systematic review of recent literature suggests that while "personalized gamification" is a crucial technological approach, comprehensive frameworks combining motivational theory (like Octalysis) with adaptive AI systems are scarce [7]. Existing implementations often suffer from a dichotomy: they either feature advanced AI with superficial gamification or robust gamification with static, non-adaptive content. Recent studies confirm that current gamified systems frequently fail to personalize motivational cues dynamically, limiting their long-term effectiveness [8].

This research proposes "Starcoder," a novel learning platform that bridges this gap by integrating

the Octalysis Framework with a generative AI-based ITS. The system was developed using the Next.js framework to ensure high performance and scalability [9], incorporating the Gemini AI API to function as an adaptive mentor named "M.E.C.H.A." The primary objective is to design a cohesive environment where the eight core drives of Octalysis, such as Epic Meaning and Scarcity, are supported by real-time, context-aware AI feedback. The study employs a Research and Development (R&D) methodology, culminating in a comparative evaluation using the Hedonic-Motivation System Adoption Model (HMSAM) with 54 respondents to measure improvements in engagement and perceived usefulness against traditional learning methods.

### *A. Research Gap*

Despite the proven efficacy of both gamification and ITS independently, the seamless integration of these two paradigms remains a significant research gap. A systematic review of recent literature suggests that while "personalized gamification" is a crucial technological approach, comprehensive frameworks combining motivational theory (like Octalysis) with adaptive AI systems are scarce [7]. Existing implementations often suffer from a dichotomy: they either feature advanced AI with superficial gamification (points and badges only) or robust gamification with static, non-adaptive content [19]. Recent studies confirm that current gamified systems frequently fail to personalize motivational cues dynamically, limiting their long-term effectiveness [8]. There is a critical lack of theoretical models and tested implementations that systematically combine the eight core drives of Octalysis with Generative AI-based ITS architectures.

### *B. Research Objectives*

To bridge this gap, this study aims to achieve the following objectives:

1. To design and build "Starcoder," a web-based learning platform that systematically integrates the Octalysis Gamification Framework with an Intelligent Tutoring System architecture.
2. To implement a generative AI agent (M.E.C.H.A.) capable of providing real-time, context-aware feedback and adaptive remedial learning paths.
3. To measure and analyze the difference in user experience—specifically regarding motivation, curiosity, and perceived usefulness—between the proposed gamified platform and traditional classroom learning methods.

### *C. Research Questions*

Based on the identified problems and objectives, this study addresses the following research questions:

1. How can a web-based learning platform be designed to effectively integrate the eight core drives of the Octalysis Framework with an AI-driven Intelligent Tutoring System?
2. To what extent does the proposed gamified system improve user perception, specifically in terms of Behavioral Intention to Use and Curiosity, compared to traditional non-gamified learning methods?

## II. METHOD

This study adopts a Research and Development (R&D) approach utilizing the ADDIE Model (Analysis, Design, Development, Implementation, Evaluation) to ensure a systematic development process. The stages are defined as follows:

1. Analysis: Identifying motivational gaps in current programming pedagogy.
2. Design: Mapping the eight Octalysis core drives to specific system features and architecting the AI-driven ITS logic.
3. Development: Constructing the "Starcoder" platform using Next.js and integrating the Google Gemini API for the "M.E.C.H.A" agent.
4. Implementation: Deploying the system to a pilot cohort of undergraduate students.
5. Evaluation: Assessing user perception and system acceptance using the HMSAM framework.

The platform's frontend is constructed using the Next.js framework, selected for its superior performance in educational web applications through Server-Side Rendering (SSR) and Static Site Generation (SSG). Recent studies indicate that Next.js significantly optimizes load times and SEO, which are critical factors for maintaining learner engagement in digital environments [9], [10]. For the intelligent component, the system integrates the Google Gemini API to power "M.E.C.H.A.," an AI agent capable of natural language processing and code analysis. This integration leverages Generative AI to provide context-aware explanations and adaptive feedback, a method proven to enhance the efficacy of Intelligent Tutoring Systems compared to static rule-based agents [11], [12].

### *A. System Architecture and Technologies*

The operational flow of the "Starcoder" platform is summarized in the mission flowchart shown in Fig. 1. This workflow outlines the process of learning through missions to the adaptive remedial mechanisms triggered by the Intelligent Tutoring System.

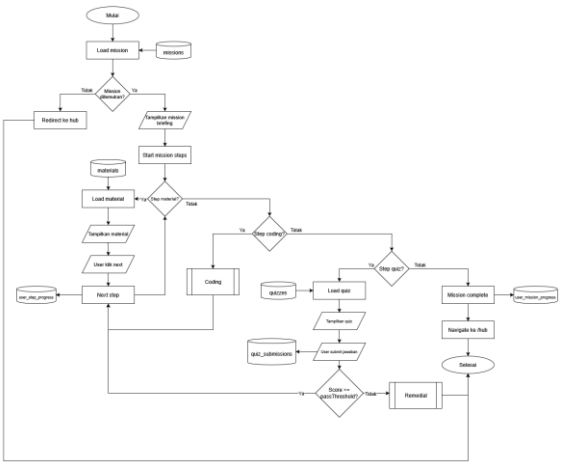


Fig. 1. Mission Flow

The detailed stages in the website are as follows:

1. *User Initialization:* The system begins with user authentication managed by Supabase, ensuring secure session management. Upon registration, a "User Dossier" is automatically generated, initializing gamification metrics such as Experience Points (XP), level, and a daily streak counter.
2. *Gamified Navigation (Nebula Path):* Users access the "Nebula Path," a visual representation of the curriculum where learning modules are depicted as planetary nodes. This interface restricts access to advanced content until prerequisite nodes are successfully completed, ensuring a structured learning trajectory.
3. *Intelligent Instruction:* Upon selecting a mission, the user enters the "Mission Interface." Here, the AI agent "M.E.C.H.A." operates in Tutor Mode, providing contextual explanations of programming concepts via the Gemini API. This interaction model is designed to build trust and facilitate scaffolded learning.
4. *Assessment and Diagnostic:* The learning process culminates in a practical coding challenge. User code is submitted to a secure sandbox environment powered by the Judge0 API, which executes the code against pre-defined test cases to verify functional correctness.
5. *Adaptive Branching:* The system analyzes the assessment results to determine the subsequent learning path:
    a. *Success Scenario:* If the user passes the test cases, the system unlocks the next node, awards XP (Octalysis Core Drive 2), and updates the leaderboard (Core Drive 5).
    b. *Remedial Scenario:* If the user fails, the ITS logic triggers a "System Diagnostic Alert." The system identifies the specific knowledge gap and recommends a "Foundational Mission" (remedial path). This ensures that students master prerequisites before re-attempting the main challenge.
6. *Feedback Loop:* All interactions, including success rates and remedial attempts, are logged in the database to refine future AI recommendations and update the user's "Starchart Log," providing a sense of ownership over their learning progress.

### B. *Octalysis Gamification Implementation*

The gamification strategy for the "Starcoder" platform is grounded in the Octalysis Framework, which categorizes human motivation into eight core drives. Unlike superficial gamification that relies solely on points and badges, Octalysis emphasizes a holistic approach that targets both extrinsic and intrinsic motivation [14].

To ensure sustained engagement, the platform implements specific features corresponding to each of the eight drives. The mapping between the theoretical core drives and their practical implementation in the system is detailed in Table I.

TABLE I. IMPLEMENTATION OF OCTALYSIS CORE DRIVES

| **Octalysis Core Drive** | **Implementation, Risks, and Impact on Evaluation** |
|---|---|
| Epic Meaning & Calling | Implementation: Users act as "Starcoders" saving a digital galaxy.<br>Impact: This narrative wrapper contributed to the 74.52% Focused Immersion (FI) score, helping students enter a flow state by contextualizing abstract code as mission-critical tools. |
| Development & Accomplishment | Implementation: Visual progression bars and immediate feedback loops via the AI agent, rather than just static points.<br>Impact: Supported the 86.44% Perceived Usefulness (PU) score by providing tangible evidence of skill acquisition. |
| Empowerment of Creativity | Implementation: A sandbox coding environment where the AI (M.E.C.H.A.) encourages debugging rather than giving answers.<br>Impact: Directly correlates to the 75.56% Control (CTL) score, as users felt agency over their problem-solving process. |
| Ownership & Possession | Implementation: Unlockable Profile Pictures (PFPs) and a customizable "User Dossier".<br>Impact: Enhances intrinsic value; students maintain the account not just for grades but to preserve their virtual identity. |
| Social Influence & Relatedness | Implementation: "Hall of Legends" leaderboard. |

| | |
|---|---|
| | Risk Analysis: While motivating for top performers, leaderboards can induce performance anxiety or demotivation for lower-ranked students. This risk is mitigated by focusing on "Top 3" highlighting rather than shaming low ranks. |
| Scarcity & Impatience | Implementation: "Daily Challenges" are locked to a 24-hour cycle.<br>Impact: This artificial scarcity contributed to the +15.56% increase in Behavioral Intention to Use (BIU), as students felt a "fear of missing out" (FOMO) on the daily opportunity. |
| Unpredictability & Curiosity | Implementation: Randomized "Gacha" (Lootbox) rewards for quiz completion.<br>Impact: This mechanic was the primary driver for the +12.47% surge in Curiosity (CUR) compared to the baseline, proving that variable rewards are more engaging than fixed rewards. |
| Loss & Avoidance | Implementation: A visual "Streak Counter" that resets upon inactivity.<br>Impact: Acts as a retention hook; while effective for short-term engagement (BIU), reliance on this drive must be balanced to avoid burnout. |

### *C. AI-Driven Intelligent Tutoring System (ITS)*

The platform's cognitive architecture is built upon the classic four-component ITS model—Domain, Student, Pedagogical, and Interface models—enhanced by Generative AI [17]. To ensure pedagogical effectiveness, the system's design is grounded in two core educational theories: Instructional Scaffolding and Mastery Learning.

1. AI Companion Modes: The AI agent, "M.E.C.H.A." (Mission Enhanced Coding Helper & Assistant), does not merely function as a chatbot but operates within the Zone of Proximal Development (ZPD). The agent dynamically adjusts its support level through three context-aware modes to provide appropriate scaffolding:
    a. Tutor Mode (Explicit Instruction): Active during material consumption, this mode lowers Cognitive Load for novice learners by breaking down abstract concepts into digestible analogies and providing direct instruction on syntax and logic.
    b. Co-Pilot Mode (Scaffolding): Active during coding challenges, this mode implements Instructional Scaffolding by offering "faded support." Instead of revealing solutions, the AI provides progressive hints and debugging clues, encouraging students to bridge the gap between their current ability and the target competency independently.
    c. Observer Mode (Formative Assessment): Monitors student inputs during quizzes without interfering, collecting real-time data on misconceptions to update the Student Model.
2. Adaptive Remedial Sequencing (Mastery Learning): The Pedagogical Model implements Bloom's Mastery Learning strategy. The system assumes that students must achieve a high level of proficiency (mastery) in prerequisite concepts before moving to more advanced tasks.
    - Mechanism: When a student fails a core mission (e.g., scoring below the threshold or failing critical test cases), the system triggers a "System Diagnostic Alert."
    - Intervention: Instead of simply allowing a retry, the ITS identifies the specific knowledge gap and activates a "Remedial Training Protocol." This redirects the student to a foundational mission designed to reinforce basic concepts.
    - Outcome: Upon successful completion of the remedial path, the original mission is unlocked. This dynamic adjustment ensures that no student is left behind due to accumulating knowledge gaps.

The flow is as illustrated in Fig. 2.

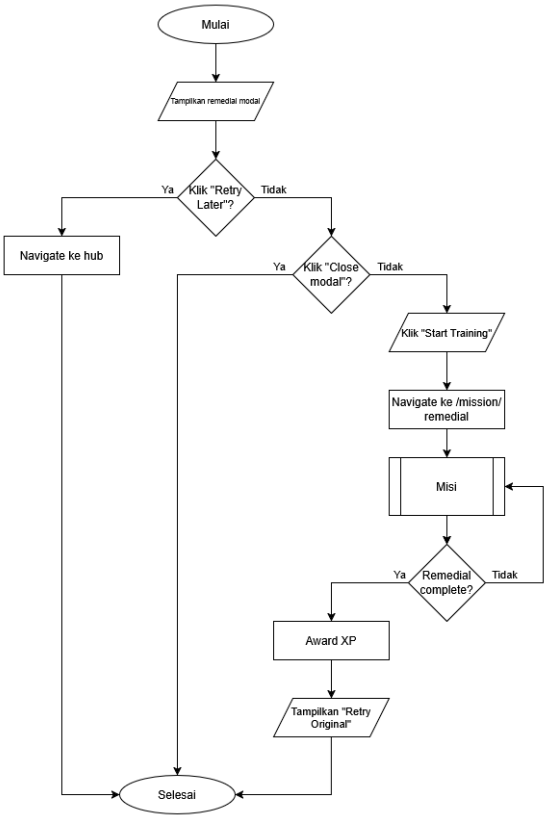



Fig. 2. Remedial Flow

3. Automated Code Assessment via Judge0: To validate practical skills, the system integrates the Judge0 API for secure, real-time code execution. As illustrated in Fig. 3, the live coding workflow follows a strict validation protocol. Upon code submission, the system performs initial validation before sending a batch request to the Judge0 service. The results are compared against pre-defined test cases to provide immediate pass/fail feedback. The live coding flow is as follows:

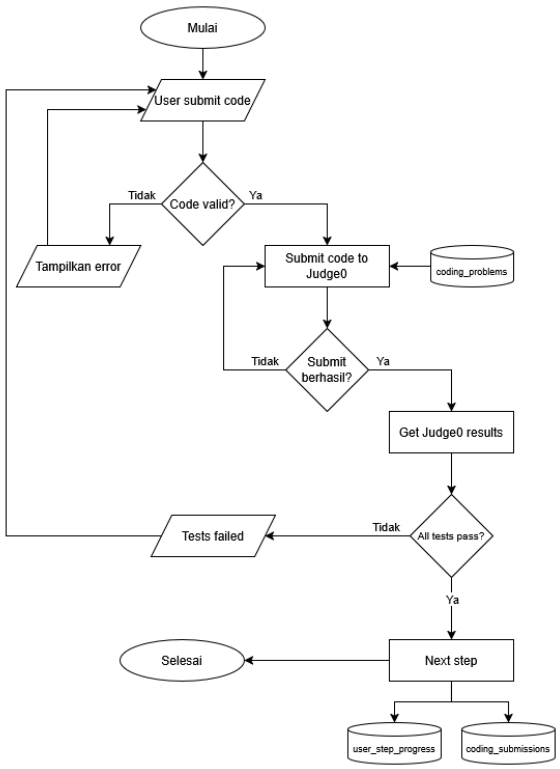



Fig. 3. Live Coding Flow

Upon code submission, the system performs initial validation before sending a batch request to the Judge0 service. The system utilizes a polling mechanism to check submission status asynchronously. Once execution is complete, results are compared against pre-defined test cases . This automated assessment provides immediate pass/fail feedback, which serves as the primary trigger for the system's adaptive logic.

*D. Evaluation Design and Instruments*

To quantitatively assess the platform's impact on student motivation and engagement, this study employs the Hedonic-Motivation System Adoption Model (HMSAM). Unlike traditional Technology Acceptance Models (TAM) that focus primarily on utilitarian aspects, HMSAM is specifically designed to evaluate systems where intrinsic motivation and enjoyment are critical, making it ideal for gamified learning environments [16], [20].

1. Study Design: This research employs a comparative within-subjects design to evaluate the impact of the proposed system on user perception. The same group of respondents (N=54) evaluated two distinct learning conditions to allow for a direct paired comparison:
   - Condition A (Baseline): The traditional non-gamified classroom instruction (standard lectures and IDEs) that the students had previously experienced or were currently undertaking.
   - Condition B (Experimental): The gamified, AI-supported "Starcoder" platform developed in this study.
2. Measurement of Baseline: To establish a valid baseline for comparison, respondents were explicitly asked to rate their engagement levels regarding their *standard classroom experience* using the same HMSAM constructs prior to evaluating the Starcoder platform. This retrospective evaluation method ensures that the comparison of "Intention to Use" and "Curiosity" is relative to their existing educational context, providing a quantifiable gap between traditional methods and the proposed AI-gamified approach.
3. Data Collection: The survey was distributed to undergraduate students who had completed or were currently enrolled in the Fundamental Programming course. The data collection focused on verifying whether the integration of Octalysis drives and AI support resulted in a statistically observable improvement in user perception compared to the established baseline.
4. Measurement Constructs: The evaluation instrument consists of a questionnaire utilizing a 5-point Likert Scale (1 = Strongly Disagree to 5 = Strongly Agree). The instrument measures seven key constructs:
   - Perceived Ease of Use (PEOU): The degree to which using the system is free of effort.
   - Perceived Usefulness (PU): The degree to which the system enhances learning performance.
   - Curiosity (CUR): The extent to which the system evokes user curiosity.
   - Joy (JOY): The perceived enjoyment and fun derived from the system.
   - Control (CTL): The user's sense of agency over the learning interaction.

- Behavioral Intention to Use (BIU): The likelihood of the user continuing to use the platform.
- Focused Immersion (FI): The level of deep engagement and flow experienced.

5. Data Analysis: The study involved respondents from undergraduate programming courses. The collected data was analyzed using a percentage score formula to determine the agreement level for each construct. The percentage score (PS) is calculated as follows:

$$PS = \frac{\sum(R \times F)}{M \times N} \times 100\%$$

Where:

- R = Weight of the choice (1 to 5).
- F = Frequency of the choice.
- M = Maximum score per item (5).
- N = Total number of respondents.

A score above 80% is categorized as "Strongly Agree" (Very Good), while scores between 61-80% are categorized as "Agree" (Good). This quantitative approach allows for a direct statistical comparison between the gamified "Starcoder" platform and the baseline traditional classroom method.

### *E. Data Analysis and Validity*

To ensure the scientific rigor of the findings, the research instrument underwent validity and reliability testing prior to descriptive analysis.

Internal consistency was measured using Cronbach's Alpha (α) for both the Baseline (Classroom) and Experimental (Starcoder) conditions. A threshold of $\alpha \geq 0.70$ was set as the standard for high reliability, while $0.60 \leq \alpha < 0.70$ was considered acceptable for exploratory psychological constructs in early-stage research.

The analysis of the respondent data (N=54) yielded the following reliability statistics for the Starcoder platform:

- High Reliability: The constructs for Perceived Ease of Use (α=0.927), Behavioral Intention to Use (α=0.908), Perceived Usefulness (α=0.870), and Curiosity (α=0.833) demonstrated excellent internal consistency, confirming that the core acceptance metrics are highly reliable.
- Moderate Reliability: The constructs for Joy (α=0.681) and Control (α=0.668) fell within the acceptable range for exploratory affective measures, indicating consistent enough patterns to draw conclusions about user sentiment.
- Variable Reliability: The Focused Immersion (α=0.573) construct showed lower consistency in the experimental condition compared to the baseline (α=0.768). This suggests that the "flow state" experienced by students in the gamified environment was more variable or subjective than in the traditional classroom, likely due to varying levels of familiarity with game mechanics among the respondents.

Following validation, the data was analyzed using descriptive statistics. The percentage scores were calculated to determine the comparative gap between the Baseline and Experimental conditions, as presented in the Results section.

## III. RESULT AND DISCUSSIONS

### *A. Platform User Interface Implementation*

To validate the functional implementation of the design, the key interfaces of the platform were deployed and tested by the respondents. The visual implementation of the gamified elements and AI support is presented below.

To drive exploration and provide a sense of progression, the Nebula Path replaces traditional module lists with an interactive galactic map, allowing students to visualize their learning trajectory as a space exploration mission.


Fig. 4. Nebula Path Navigation Interface

The core learning experience utilizes a clean, mission-focused interface where learning materials are presented as mission briefings to maintain narrative immersion while delivering technical concepts.

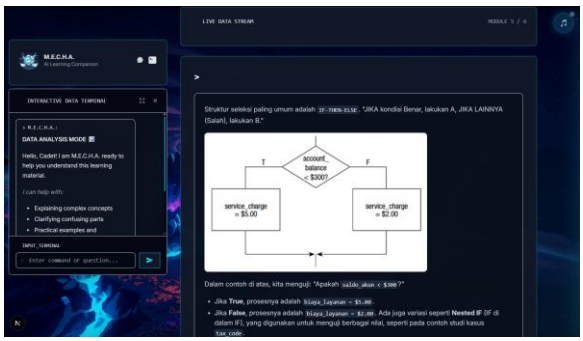



Fig. 5. Mission Learning Material Interface

Interactive assessments are integrated directly into the flow; the Quiz Interface provides immediate

feedback on conceptual understanding, reinforcing knowledge before practical application.

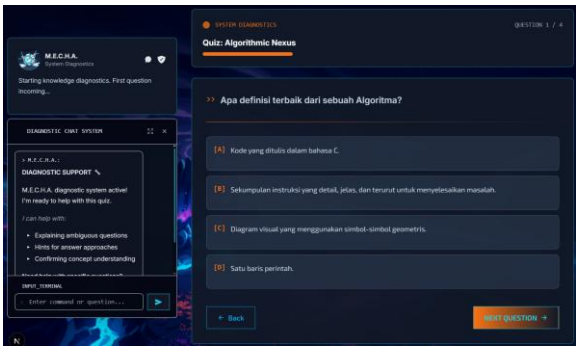



Fig. 6. Interactive Quiz Interface

To validate practical skills, the Live Coding Environment features a split-screen layout with a code editor and real-time execution results, allowing students to write and test code within the browser.

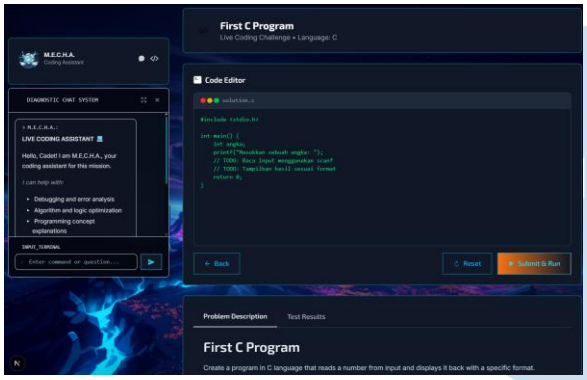



Fig. 7. Live Coding Environment

The AI agent, M.E.C.H.A., is embedded directly within the workspace to offer contextual guidance, answering questions and providing debugging assistance without requiring the user to leave the learning environment.

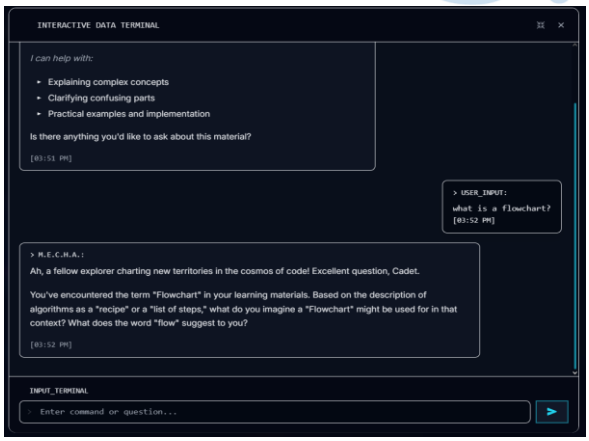



Fig. 8. AI Companion Chat Interface

Finally, when performance thresholds are not met, the system automatically triggers a Remedial Protocol, directing users to foundational exercises to ensure prerequisite knowledge is mastered before proceeding.


Fig. 9. Adaptive Remedial Prompt

## B. *Comparative Analysis of HMSAM Constructs*

The effectiveness of the "Starcoder" platform was evaluated through a quantitative study involving 54 respondents (N=54). The participants were students who had previously taken or were currently enrolled in a Fundamental Programming course. The data was collected using a standard HMSAM questionnaire, ensuring a valid comparison between the traditional classroom experience (Baseline) and the gamified AI-driven platform.

The demographic composition of the respondents provides context for the evaluation results. As illustrated in Fig. 4, the gender distribution was predominantly male (85.2%) compared to female (14.8%), a ratio typical of students. Despite the gender imbalance, the sample size is statistically sufficient to draw initial conclusions regarding system acceptance.

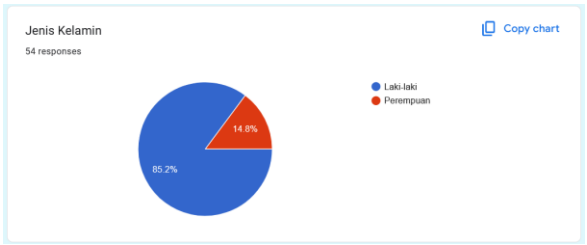



Fig. 10. Gender Distribution of Respondents

In terms of age distribution, shown in Fig. 5, the majority of respondents fell into the 18–20 year age group (61.1%), followed by the 21–23 year group (35.2%). This indicates that the primary evaluators were in the early-to-mid stages of their university education, which aligns perfectly with the target audience for an introductory programming platform. This age group is generally considered "digital natives," possessing a high baseline familiarity with both gaming mechanics and digital learning interfaces.

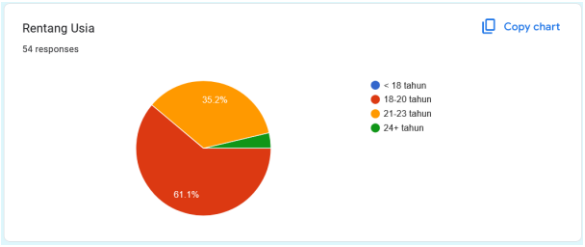



Fig. 11. Age Distribution of Respondents

Furthermore, the study achieved a degree of institutional diversity. While the majority of respondents originated from the host university (87%), there was participation from external institutions (13%), including Binus University and Universitas Terbuka, as shown in Fig. 6.

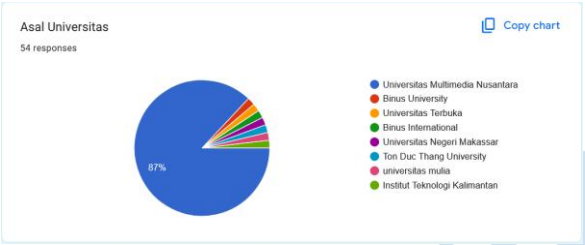



Fig. 12. University Origin of Respondents

The core of the evaluation rests on the comparison between the baseline classroom experience and the "Starcoder" platform across seven HMSAM constructs. The summary of these findings is presented in Table II.

TABLE II. COMPARISON OF EVALUATION SCORES

| HMSAM Constructs | Classroom (Baseline) | Starcoder | Increase | Result Category |
|---|---|---|---|---|
| Behavioral Intention to Use (BIU) | 62.84% | 78.40% | +15.56% | Good (Agree) |
| Curiosity (CUR) | 73.09% | 85.56% | +12.47% | Very Good (Strongly Agree) |
| Perceived Usefulness (PU) | 75.33% | 86.44% | +11.11% | Very Good (Strongly Agree) |
| Control (CTL) | 66.42% | 75.56% | +9.14% | Good (Agree) |
| Perceived Ease of Use (PEOU) | 76.39% | 84.63% | +8.24% | Very Good (Strongly Agree) |
| Focused Immersion (FI) | 67.31% | 74.52% | +7.20% | Good (Agree) |
| Joy (JOY) | 67.47% | 73.58% | +6.11% | Good (Agree) |

The descriptive analysis reveals a consistent positive trend across all metrics, with the "Starcoder" platform receiving higher approval ratings compared to the traditional baseline in every category. The most notable difference was observed in Behavioral Intention to Use (BIU), which showed a difference of 15.56% between the two conditions. In the baseline study, students rated their intention to continue with traditional methods at a moderate 62.84%. However, after interacting with Starcoder, this metric rose to 78.40%. This suggests that the combination of gamification and AI support may help convert a passive learning obligation into an active desire to engage with the material.

The second highest difference was observed in Curiosity (CUR), which was 12.47% higher in the experimental condition, reaching a "Strongly Agree" level of 85.56%. While traditional programming lectures are informative, the data suggests they may struggle to maintain the element of anticipation. By contrast, Starcoder's use of "Black Hat" gamification techniques, specifically Octalysis Core Drive 7 (Unpredictability) via randomized rewards, appears to successfully stimulate student curiosity based on the self-reported responses.

Perceived Usefulness (PU) also saw a substantial increase of 11.11%, achieving the highest overall score of 86.44%. This validates the integration of the AI agent "M.E.C.H.A." unlike static textbooks or pre-recorded videos, the AI provided immediate, context-aware assistance. The ability of the system to diagnose specific errors and offer remedial paths (as detailed in the Methodology) directly contributed to students feeling that the system was practically useful for their learning goals.

The results of this study align with recent literature suggesting that "Personalized Gamification" is superior to generic approaches [17], [20]. The moderate gains in *Control (CTL)* (+9.14%) and *Joy (JOY)* (+6.11%) suggest that the gamified environment provided a stronger sense of agency. The "Nebula Path" allowed students to navigate at their own pace, while the AI agent provided support without taking over the task.

Furthermore, the increase in Focused Immersion (FI) (+7.20%) highlights the potential effectiveness of the narrative wrapper (Epic Meaning). By framing code compilation as a "mission" to repair a digital universe, the platform reportedly induced a deeper state of engagement than standard IDEs. This confirms that wrapping technical tasks in a thematic narrative is a viable strategy for reducing the perceived cognitive load often associated with introductory programming [13], [15].

### C. Limitations

While the results indicate a strong preference for the gamified system, several limitations must be acknowledged to contextualize the findings:

1. Descriptive Nature of Analysis: The comparative analysis relies on descriptive statistics (percentage scores) rather than inferential statistical tests (e.g., t-tests). Consequently, while the differences in means

are distinct, statistical significance ($p<0.05$) was not calculated. The findings should be interpreted as indicative of user preference rather than a confirmed causal effect.

2. Self-Reported Measures: The data is derived entirely from the HMSAM questionnaire, which measures perceived experience. This introduces potential perceptual bias, where a "fun" system is rated highly even if learning gains are not proportional.
3. Lack of Objective Learning Outcomes: This study did not measure cognitive learning gains through pre-test and post-test scores. While Behavioral Intention to Use is a strong predictor of engagement, it does not guarantee improved academic performance (grades). Future research will require longitudinal studies with control groups to validate whether this increased motivation translates into measurable programming proficiency.

## IV. CONCLUSIONS

This study designed and evaluated "Starcoder," an adaptive learning platform integrating the Octalysis Gamification Framework with a Generative AI-driven Intelligent Tutoring System. The development process confirmed that combining modern web technologies (Next.js) with Large Language Model APIs (Gemini) is a technically viable approach for creating responsive, gamified educational environments.

Evaluation results from the pilot cohort (N=54) indicate a positive reception of the system. Descriptive analysis of the HMSAM constructs reveals that the gamified platform received higher approval ratings compared to the traditional classroom baseline, particularly in *Perceived Usefulness* (86.44%) and *Curiosity* (85.56%). The data suggests that the integration of narrative elements and adaptive AI support may effectively foster *Behavioral Intention to Use*, which showed a 15.56% difference relative to the baseline. These findings support the premise that addressing intrinsic motivation through "White Hat" (Meaning, Accomplishment) and "Black Hat" (Scarcity, Curiosity) gamification drives is a promising strategy for enhancing student engagement.

However, these conclusions must be interpreted within the context of the study's limitations. The findings rely on self-reported perception data rather than objective learning outcomes, and the sample size prevents broad generalization. Consequently, future research should prioritize the following areas to validate the pedagogical efficacy of the system:

- Objective Performance Measurement: Future studies must implement a quasi-experimental design with Pre-Test and Post-Test assessments to measure actual cognitive gains and programming proficiency, rather than relying solely on perceived usefulness.
- Longitudinal and Scaled Analysis: To rule out the "novelty effect," longitudinal studies with a larger, multi-institutional sample size are required to determine if the increased motivation translates into sustained long-term engagement and retention.
- Technical Evolution: Beyond validation, technical development should focus on enabling the AI to dynamically adjust gamification parameters (e.g., difficulty curves, reward probabilities) in real-time based on the learner's performance data, moving from a static rule-based system to a fully adaptive motivational engine.

## REFERENCES


[1] L. Christopher and A. Waworuntu, "Java programming language learning application based on octalysis gamification framework," IJNMT (International Journal of New Media Technology), vol. 8, no. 1, pp. 65-69, 2021.

[2] E. Ratinho and C. Martins, "The role of gamified learning strategies in student's motivation in high school and higher education: A systematic review," Heliyon, vol. 9, 2023.

[3] N. A. Pratama, A. C. Sari, and P. K. N. Bara, "The effect of online learning on student motivation and achievement during the pandemic," Edumaniora: Jurnal Pendidikan dan Humaniora, vol. 3, no. 2, pp. 33-38, 2024.

[4] T. Karakose, R. Yirci, S. Papadakis, T. T. Ozdemir, M. Demirkol, and H. Polat, "Examining the effectiveness of gamification as a tool promoting teaching and learning in educational settings: a meta-analysis," Frontiers in Psychology, vol. 14, p. 1253549, 2023.

[5] E. Rivera and C. Garden, "Gamification for student engagement: a framework," Journal of Further and Higher Education, vol. 45, pp. 999–1012, 2021.

[6] A. Létourneau, M. Dupuis, and S. Tremblay, "A systematic review of ai-driven intelligent tutoring systems: Effects on k-12 learning and retention," International Journal of Artificial Intelligence in Education, vol. 35, no. 2, pp. 215-238, 2025.

[7] S. D. Ristianto, A. Putri, and Y. Rosmansyah, "Personalized gamification a technological approach for student education: A systematic literature review," IEEE Access, vol. 13, pp. 55 712-55 726, 2025.

[8] F. Naseer, M. Khan, A. Addas, Q. Awais, and N. Ayub, "Game mechanics and artificial intelligence personalization: A framework for adaptive learning systems," *Education Sciences*, vol. 15, p. 301, 2025.

[9] R. Pati, N. Sharma, and J. Lee, "Building scalable intelligent tutoring systems with next.js and ai microservices," Journal of Web Engineering, vol. 19, no. 2, pp. 123-145, 2025.

[10] Aghaei and S. Kumar, "Performance and seo benefits of next.js in educational web applications," in *Proceedings of the International Conference on Educational Technology*, 2024, pp. 78-89.

[11] H. Almetnawy, A. Orabi, A. Alneyadi, T. Ahmed, and A. Lakas, "An adaptive intelligent tutoring system powered by generative ai," 2025, pp. 1-10.

[12] M. Mukherjee, J. Le, and Y. Chow, "Generative ai-enhanced intelligent tutoring system for graduate cybersecurity programs," *Future Internet*, vol. 17, no. 4, p. 154, 2025.

[13] G. Papagni, J. de Pagter, S. Zafari, M. Filzmoser, and S. T. Koeszegi, "Artificial agents' explainability to support trust: considerations on timing and context," *AI & SOCIETY*, vol. 38, no. 2, pp. 947–960, 2022.

[14] Octalysis Group, "Framework," Sep 2023. [Online]. Available: https://octalysisgroup.com/framework/

[15] F. Marisa, S. S. S. Ahmad, Z. I. M. Yusoh, A. L. Maukar, R. D. Marcus, and A. A. Widodo, "Evaluation of student core drives on e-learning during the covid-19 with octalysis gamification framework," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 11, pp. 104-116, 2020.

[16] Suwarno, D. Deli, and M. Christian, "Exploring the impact of octalysis gamification in japanese m-learning using the technology acceptance model," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, vol. 9, no. 1, pp. 77-88, 2024.

[17] P. Sharma and M. Harkishan, "Designing an intelligent tutoring system for computer programing in the pacific," *Education and Information Technologies*, vol. 27, pp. 6197 – 6209, 2022.

[18] T. Shah, "Ai-powered personalised learning plans for intelligent tutoring systems," *International Journal for Research in Applied Science and Engineering Technology*, vol. 12, no. 5, pp. 5450-5462, 2024.

[19] F. Keshtkar, N. Rastogi, S. Chalarca, and S. Ahmad Chan Bukhari, "Ai tutor: Student's perceptions and expectations of ai-driven tutoring systems: A survey-based investigation," in The International FLAIRS Conference Proceedings, vol. 37, 2024.

[20] R. Pati, N. Sharma, and J. Lee, "Building scalable intelligent tutoring systems with next.js and ai microservices," Journal of Web Engineering, vol. 19, no. 2, pp. 123-145, 2025

ISSN 2085-4552

# Integration of Internet of Things Technology in Digital-Based Residential Security Application

Rajib Ghaniy[1], Binanda Wicaksana[2], Fahmi Arnes[3], Laras Melati[4], Helena Septiana[5]
[1,2,3,4,5]Faculty of Informatics and Computer Science, Universitas Binaniaga Indonesia, Bogor, Indonesia
[1]rajib@unbin.ac.id, [2]binanda@unbin.ac.id, [3]fahmi@unbin.ac.id,
[4]larasmlt05@gmail.com, [5]helenaseptiana13@gmail.com

Accepted 04 January 2026
Approved 22 January 2026

***Abstract*— This study evaluates the usability of an IoT-based residential security application designed to improve guest registration and access control in housing environments. The system integrates multiple user roles, including administrators, residents, guests, and security officers, and utilizes QR Code verification to streamline entry procedures. Usability testing was conducted using the System Usability Scale (SUS) with 30 respondents. The results show an average SUS score of 74.83, indicating that the application falls within the "Good" usability category. Most users reported that the interface is intuitive, the functions are well integrated, and system navigation is easy to understand. Although minor improvements are still required—such as notification speed and icon clarity—the system is considered acceptable for public use. These findings demonstrate that IoT integration can enhance residential security operations while maintaining positive user experience.**

***Index Terms*— Internet of Things; residential security; QR Code verification; System Usability Scale; usability evaluation; user experience.**

## I. INTRODUCTION

According to Law of the Republic of Indonesia Number 1 of 2011, housing is defined as a group of houses equipped with various environmental infrastructure and facilities [1]. One of the essential environmental facilities in a housing area is a security system. The conventional security system commonly found in housing complexes involves placing guard posts along with security personnel to ensure that access in and out of the area remains controlled. Various standard operating procedures are implemented to maintain security within a housing area, where each housing complex may have different procedures, but all share the same goal—ensuring the safety of the residential environment [2].

According to data from the National Criminal Information Center (PUSIKNAS), as of July 2024, theft remains the most frequently occurring type of crime [3]. Criminal acts are not only influenced by the personal factors of the perpetrators but are also strongly affected by environmental conditions [4]. Additionally, the high crime rate indicates that accessibility and the design of residential areas play a major role in determining the level of vulnerability to crime. Housing located near main roads and lacking adequate security systems is more likely to become a target for criminals. Therefore, crime prevention efforts must combine physical environmental planning (architecture and spatial design) with strengthened social control within the community [5].

The development of digital technology based on the Internet of Things (IoT) is increasingly widespread across various fields, including residential security systems. IoT enables physical devices such as cameras, motion sensors, and smart locks to connect to the internet and be controlled through mobile applications. This innovation allows real-time monitoring of home conditions, thereby enhancing residents' sense of security [6].

Although offering significant benefits, the implementation of IoT-based security applications is not without challenges. Some users still doubt the usefulness of new technologies compared to conventional security systems, particularly regarding their effectiveness in preventing criminal acts [7]. Moreover, issues of privacy and data security have also become concerns, as the potential for information leakage can reduce user trust [8].

In the context of Human–Computer Interaction (HCI), user acceptance of an application is influenced by its ease of use and perceived usefulness. Applications considered difficult to use may reduce user interest and positive attitudes, even when the features offered are beneficial [9]. Therefore, user experience becomes one of the key factors determining the success of implementing IoT-based digital security systems.

In previous studies [6], the application of IoT has been utilized in residential security systems. However, those studies did not implement usability measurement

based on the System Usability Scale (SUS) in evaluating the application of IoT. This gap serves as the basis for identifying the novelty of this study, as it introduces a usability evaluation using the SUS framework, which has not been addressed in prior research.

Based on the above explanation, this study formulates the research problems to be addressed: How high is the usability value of IoT-based residential security systems? And how can the usability of IoT-based residential security systems be measured?

## II. THEORY

### A. *Internet of Things*

The Internet of Things is a concept in which physical devices such as sensors, cameras, actuators, and various other electronic devices are able to connect and interact with each other through the internet [10].

In general, the architecture of an IoT ecosystem follows a hierarchical pattern consisting of three main layers:

1. Edge, This layer contains connected devices and sensors that directly interact to process data in real time and make quick decisions.
2. Fog, This layer serves as an intermediary between the edge and the cloud layers by managing and optimizing data traffic, as well as performing a significant portion of distributed data processing.
3. Cloud, This layer stores data and applies advanced data analytics to generate deeper insights.

Through this architecture, data from various devices and sensors is collected, processed, and integrated to create useful solutions, ensuring that IoT-based services can be effectively utilized by users [11].

Sensor technology in IoT serves as the fundamental foundation for IoT development. With sensors, devices are able to detect and measure their surroundings. The key stages in IoT data and control flow include data collection, data processing, and the resulting processed output, which ultimately impacts the real world.

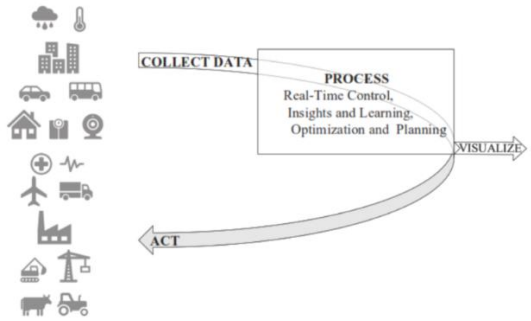



Fig 1. Functional flow in IoT system [11]

### B. *Qr Code*

A QR Code is an abbreviation for *Quick Response Code* [12]. A QR Code is a type of two-dimensional (2D) barcode, which distinguishes it from traditional one-dimensional barcodes: instead of straight lines, a QR Code consists of a square pattern filled with numerous small black-and-white modules. It was first created in 1994 by Denso Wave, previously part of the Toyota Group, to assist in tracking components in manufacturing [13]. Since then, the QR Code has become an international standard—widely adopted for various modern applications [14].

QR Codes work by storing data through encoding information into a pattern of black-and-white modules (squares) arranged in a grid/matrix. When scanned with a smartphone camera or QR scanner, the device interprets the pattern into digital information (such as URLs, text, contacts, etc.) [15]. The three large squares located at three corners of the QR Code are position markers (also called *finder patterns*). These markers help the device determine the code's orientation so it can be read from various angles [16].

Fig 2. Qr Code

QR Codes can store various types of data: numeric, alphanumeric, binary/byte data, and even more complex characters such as Kanji [17]. QR Codes also incorporate an error correction mechanism, allowing them to remain readable even if part of the code is scratched, damaged, or obstructed—making them resistant to physical wear or imperfect printing [18].

### C. *System Usability Scale*

The System Usability Scale (SUS) is a standardized instrument widely used to measure the usability level of a system, application, or device, and it has remained the most popular method over the past five years due to its simplicity, reliability, and validity [19]. Numerous studies have shown that SUS provides consistent evaluative performance across various contexts, including digital interfaces [20], interactive environments [21], and cross-device evaluations involving IoT technologies [22].

SUS consists of ten items rated on a 1–5 Likert scale, and the score is calculated using the standard formula: (1) for odd-numbered items: contribution score = response value − 1; (2) for even-numbered items: contribution score = 5 − response value. All contribution scores are then summed and multiplied by

2.5 to produce a final value ranging from 0 to 100. The interpretation of SUS scores has been refined through contemporary approaches, including grading scales (A–F), adjective ratings, and acceptability ranges [19].

In general, a score of ≥ 80.3 is categorized as Excellent (Grade A) and indicates a very high level of usability; scores between 68–80.3 fall into the Good (Grade B) category and are generally considered to reflect good usability; scores within 65–68 are classified as OK / Marginal High (Grade C); scores from 50–65 are categorized as Marginal / Poor (Grade D); while scores < 50 fall into the Not Acceptable (Grade F) category. These interpretations allow researchers to determine whether a system meets usability standards and to compare a system's performance with industry and research benchmarks.

Thus, SUS remains a relevant and empirically robust tool for evaluating usability in IoT systems as well as other digital technologies [23].

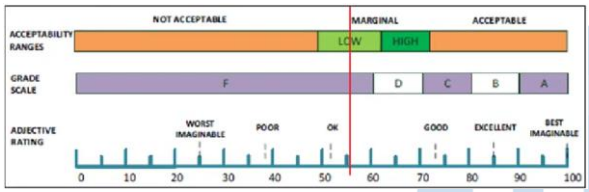



Fig 3. Acceptance rate SUS [19]

## III. METHOD

### A. Research Design

Studies using the System Usability Scale (SUS) generally employ a descriptive quantitative approach, as SUS produces numerical scores that objectively represent the usability level of a system [24]. This research aims to evaluate the usability level of a technology-based system/product, where data are collected once after users have interacted with the system.

### B. Population and Sample

The population in this study consists of prospective users of the IoT-based residential security application in urban housing environments. The sampling technique used slovin with the following criteria:

- Owning an Android or iOS smartphone.
- Having used or had prior experience with technologies.
- Residing in a housing/residential area.

The sample size was determined using Slovin's formula with a margin of error of 10%. Based on a population of 43 respondents, a total of 30 samples were obtained and considered sufficient to represent the population for user evaluation purposes.

$$n = \frac{N}{1 + Ne^2}$$

$$n = \frac{43}{1 + 43(0{,}1)^2}$$

$$n = 30$$

The total sample consists of 30 respondents drawn from residents across five housing complexes located in Bogor City.

### C. Research Instrument

The research instrument used is the SUS questionnaire, which consists of 10 statements with responses measured using a 1–5 Likert scale, where:

1 = Strongly Disagree

2 = Disagree

3 = Neutral

4 = Agree

5 = Strongly Agree

The statements are arranged alternately as positive and negative items to maintain validity. The SUS items are as follows:

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.
4. I think I would need the support of a technical person to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very cumbersome to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system.

It should be noted that the odd-numbered items (1, 3, 5, 7, 9) are positive statements, while the even-numbered items (2, 4, 6, 8, 10) are negative statements.

SUS Scoring Procedure:

- Positive items = (Response value − 1)
- Negative items = (5 − Response value)
- Total SUS Score = (Sum of contributions) × 2.5

The final SUS score ranges from 0 to 100, with the following interpretation:

- > 80.3 = Excellent
- 68 – 80.3 = Good / Acceptable
- 50 – 68 = Marginal
- < 50 = Not Acceptable

## IV. Result and Discussion

### A. Result

#### 1) System Analysis

Before the system was developed and the IoT components were designed, a system analysis was carried out. The researcher analyzed what kind of system was needed and how the system would be used. The results of the system analysis included the system scenarios and a use case diagram.

Table 1. Actor and Role in System

| Actor Name | Role in the System |
|---|---|
| Super Admin | An actor responsible for validating RT Head accounts. |
| RT Head | An actor who functions as the system administrator. |
| Security Officer | An actor responsible for validating guest presence/arrival. |
| Resident | An actor who serves as the primary business actor, registering guests who plan to visit. |
| Guest | An actor who receives the QrCode provided by the resident. |

**System Scenario**

The system is used by actors, and each actor has their own role within the system.

**Super Admin**

The Super Admin logs into the system using registered credentials. The primary responsibility of the Super Admin is to register RT Head accounts by entering required information such as National ID Number, full name, phone number, home address, term of office, and the official appointment letter issued by the authorized authority. Upon completing system operations, the Super Admin may log out.

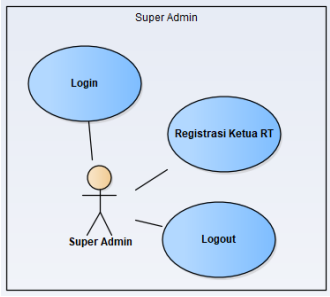



Fig 4. Usecase Diagram Super Admin

**RT Head**

The RT Head accesses the system using a registered username and password. The RT Head is responsible for managing resident data, including the Head of Family's National ID Number, name, home address, phone number, number of occupants, list of occupants, and resident status (local resident or newcomer).

The system notifies the RT Head when residents submit information regarding planned guest visits or their temporary absence from home. The RT Head validates guest visit requests, particularly when the guest intends to stay for more than 24 hours. Additionally, the RT Head can view the list of residents, security officers, and guest visit records within a specified period. The RT Head may log out after completing required tasks.

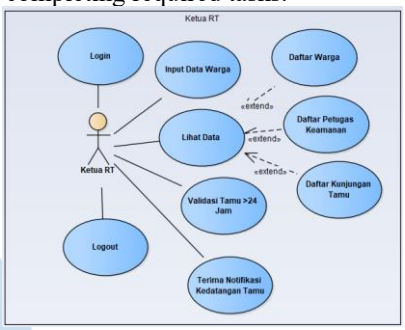



Fig 5. Usecase Diagram RT Head

**Security Officer**

The Security Officer logs into the system using registered credentials. The officer performs verification by scanning the QrCode presented by guests who have been registered by residents. The system also provides notifications of planned guest visits immediately after residents submit the data. The Security Officer is able to view the list of planned guest visits for the current day. The officer may log out after completing system-related duties.

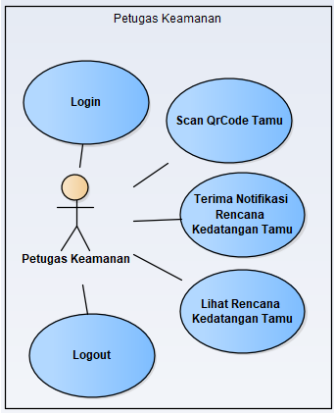



Fig 6. Usecase Diagram Security Officer

**Resident**

Residents access the system using registered accounts. They can submit planned guest visit information, including guest name, the resident being visited, number of guests, type of vehicle,

and duration of the visit. The system automatically generates a QrCode based on the submitted data, which residents can share with guests through messaging applications such as WhatsApp. The QrCode is temporary and remains valid according to the visit duration.

Residents may also provide information regarding travel plans or indicate unavailability for receiving guests. This allows security personnel to inform unexpected visitors about the resident's absence or unavailability. Residents can view records of guest visits within a given time range and may log out after completing their actions.

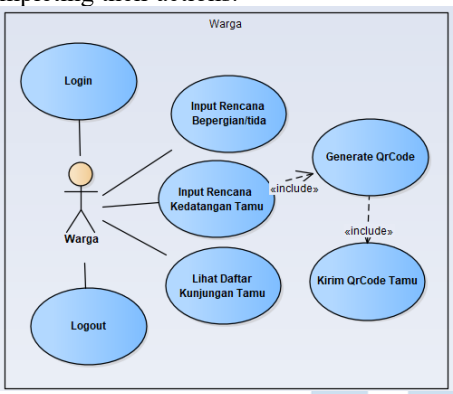



Fig 7. Usecase Diagram Resident

**Guest**

Guests receive a QrCode from the resident who has registered a planned visit. Upon arrival, guests present the QrCode to the Security Officer for verification.

*2) System Design*


Fig 8. Guest Registration Form

This form is used by residents to register guests who plan to visit their homes. Residents input detailed information based on the required fields.

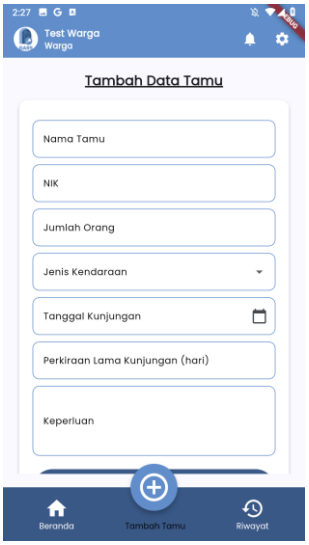



Fig 9. Guest Data Form Fields

This form is filled out by residents to register the guests who will be visiting.

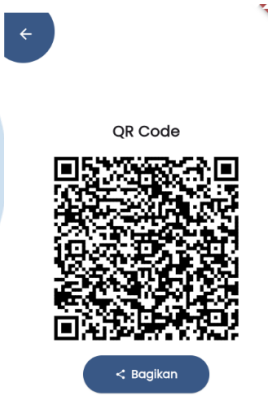



Fig 10. QR Code Form

This is the QR code generated from the guest data entered. Residents will share this QR code with the guests who plan to visit.

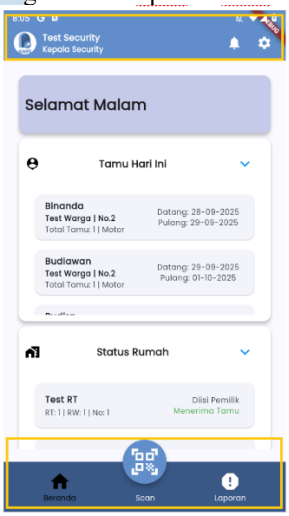



Fig 11. Security Officer Form

This form contains the main feature: QR code scanning. Security officers will scan the QR code shown by the guest.

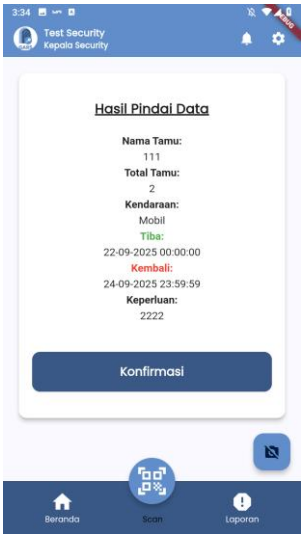



Fig 12. QR Code Scan Result Form

After the guest's QR code is scanned by the security officer, the detailed contents of the QR code will appear, including the name, date and time of visit, and the vehicle used.

*3) IoT Architect*

Fig 13. ESP32 Dev Kit

**ESP32 Dev Kit** is a development board designed to simplify the use of the ESP32 microcontroller (System on a Chip/SoC) developed by Espressif Systems. This board typically includes all the essential components (such as a voltage regulator, USB-to-UART chip, and pin headers) so you can directly program and test the ESP32 chip without needing to build a complex supporting circuit.

The **ESP32 chip** itself is the core of this kit and is known for its main features:

- **Integrated Wi-Fi and Bluetooth:** This is the key feature that makes it a top choice for IoT projects. ESP32 supports Wi-Fi (802.11 b/g/n) and Bluetooth (Classic and Low Energy/BLE).
- **Powerful Processor:** Most variants have a Tensilica Xtensa Dual-Core 32-bit LX6 processor with speeds up to 240 MHz, offering better performance compared to its predecessor, the ESP8266.
- **Low Power Consumption:** Designed for low-power applications with various sleep modes to conserve battery life.

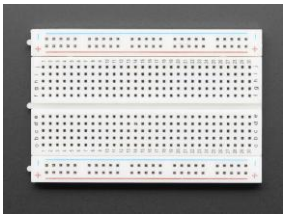


Fig 14. Breadboard

**Breadboard** is a board used for assembling and testing electronic circuits temporarily without requiring any soldering.

The name "breadboard" originates from early inventors who used wooden bread-cutting boards to attach nails and connect electronic components. Modern breadboards are made of plastic and contain small holes with internal conductive metal clips or rails that serve as connectors.

Its main functions are:

- **Rapid Prototyping:** Allows users to quickly assemble, test, modify, and disassemble electronic circuits.
- **Non-Destructive:** Components can be inserted and removed repeatedly without damage.
- **Learning Tool:** Ideal for beginners and educational purposes because it helps users understand circuit paths, voltage, and component connections.


Fig 15. Ultrasonic Sensor HC-SR04

**HC-SR04** is a widely used, low-cost ultrasonic distance sensor. It works based on sonar principles by using ultrasonic waves (sound waves above 20 kHz) to determine the distance of an object. It is commonly used in various microcontroller projects (such as Arduino or ESP32).

The working principle of the HC-SR04 is based on measuring the travel time of sound waves:

- **Trigger:** The microcontroller sends a HIGH pulse for at least 10 microseconds to the sensor's Trig pin.
- **Wave Transmission:** After receiving the trigger signal, the module emits eight ultrasonic pulses at 40 kHz through the transmitter.
- **Echo Reflection:** The ultrasonic waves travel through the air at the speed of sound (343 m/s or 0.0343 cm/microsecond) and reflect off any solid object in front of it.

- **Reception:** The reflected wave (echo) is received by the sensor's receiver.
- **Time Measurement:** The Echo pin stays HIGH from the moment the wave is sent until the reflected wave is received. The microcontroller measures the duration of this HIGH pulse, which represents the total travel time.

Fig 16. Direct Current Motor

**DC Motor** (Direct Current Motor) is an electric machine that converts direct current (DC) electrical energy into mechanical energy in the form of rotational motion.

DC motors are widely used in many applications such as toys, household appliances (blenders, car wipers), robotics systems, and industrial machines that require speed and torque control.

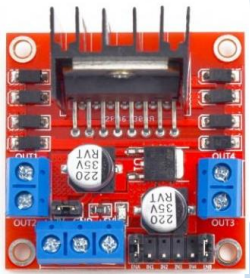


Fig 17. L298N Motor Driver

**L298N Motor Driver** is an electronic module based on the L298N Dual H-Bridge Integrated Circuit (IC). Its main functions include:

- **Controlling Motor Direction:** Converts low-voltage logic signals (e.g., from Arduino) into sufficiently large power signals to reverse motor voltage polarity, enabling forward or reverse rotation.
- **Controlling Motor Speed:** Allows DC motor speed control using PWM (Pulse Width Modulation).
- **Current Amplification:** Microcontrollers can only output very small current (around 20 mA). The L298N acts as a bridge, enabling motors requiring higher current (up to 2A per channel) to be powered by an external supply without damaging the microcontroller.

Fig 18. Limit Switch

**Limit Switch** is a type of mechanical sensor used to detect the presence, position, or movement boundary of an object or machine.

Simply put, a limit switch is an electrical switch activated (opened or closed) by physical contact when a moving object reaches its predetermined limit.

Working Principle:

- **Actuator Contact:** When a moving object contacts the actuator (lever, roller, or push button) of the limit switch.
- **Contact Change:** The actuator pushes the internal microswitch mechanism.
- **Electrical Signal:** This changes the electrical contact state:
  - ✓ **NO (Normally Open):** Becomes closed, allowing current to flow.
  - ✓ **NC (Normally Closed):** Becomes open, cutting off the current.

Fig 19. Jumper Wires

**Jumper Wires** are short conductive cables used to connect two points in an electronic circuit. They allow connections without soldering, making them ideal for assembly, testing, and prototyping.

Fig. 20 QR Code Scanner

**QR Code Scanner** is a hardware or software tool designed to read and decode information stored in a QR Code (Quick Response Code).

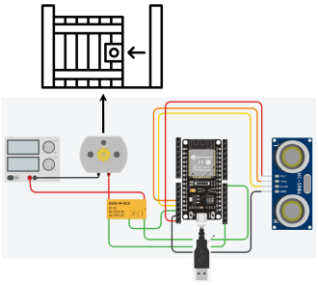


Fig 21. IoT Prototype Architect

Figure 21 explains that the IoT system operates alongside the existing system that has been developed. The system is initiated by residents, who register planned guest visits to their homes. This data is then

generated in the form of a QR code. The IoT system becomes active when the QR code is scanned by a QR code reader. The sensor system, consisting of an ESP32 microcontroller and an L298N motor driver, detects the QR code and activates the gate drive motor (DC motor) to open the gate. When the gate opens, its movement stops upon reaching the limit switch. Another sensor, namely the ultrasonic sensor HC-SR04, ensures that the gate will automatically close after an object identified as a vehicle passes through the sensor, which then triggers the gate to close again.

### *B. Discussion*

The following is the recap of the SUS questionnaire that was distributed to 30 respondents. The respondents were given questions based on the standard SUS question set.

Table 2. Recap of the SUS questionnaire

| | q1 | q2 | q3 | q4 | q5 | q6 | q7 | q8 | q9 | q1 |
|---|---|---|---|---|---|---|---|---|---|---|
| r1 | 4 | 4 | 3 | 3 | 4 | 3 | 4 | 4 | 3 | 4 |
| r2 | 4 | 5 | 3 | 4 | 3 | 5 | 3 | 5 | 3 | 4 |
| r3 | 3 | 5 | 3 | 5 | 5 | 4 | 3 | 3 | 5 | 4 |
| r4 | 5 | 4 | 4 | 4 | 4 | 5 | 4 | 3 | 5 | 5 |
| r5 | 5 | 4 | 3 | 4 | 4 | 3 | 5 | 4 | 4 | 5 |
| … | … | … | … | … | … | … | … | … | … | … | … |
| r28 | 3 | 4 | 4 | 4 | 5 | 3 | 4 | 3 | 5 | 4 |
| r29 | 3 | 3 | 5 | 5 | 3 | 5 | 4 | 4 | 3 | 3 |
| r30 | 4 | 5 | 5 | 4 | 4 | 4 | 5 | 3 | 3 | 5 |

*1) Usability Measurement Results Using the System Usability Scale (SUS)*

Usability testing on the IoT-based residential security system application was conducted using the System Usability Scale (SUS) involving 30 respondents. The SUS instrument consists of 10 statements rated on a 1–5 Likert scale. The score calculation follows the standard procedure: odd-numbered statements are calculated as score – 1, while even-numbered statements are calculated as 5 – score, and the total result is then multiplied by 2.5 to produce a final score ranging from 0–100. Based on the calculations from 30 respondents, individual SUS scores ranged from 65 to 92.5. The average SUS score reached 74.83, with a median of 75, a minimum score of 65, a maximum score of 92.5, and a standard deviation of 8.12. With an average score of 74.83, the application falls into the "Good" category according to SUS interpretation standards.

This value indicates that, in general, the application meets user expectations in terms of ease of use, clarity of functions, and interaction efficiency. An average score above the threshold of 68 signifies that the application has an acceptable level of usability and is suitable for use by the general public.

*2) Distribution and User Perception Patterns*

The score distribution shows that most respondents gave ratings in the range of 70–85, indicating a consistent perception that the application is easy to operate and supports users' needs in monitoring home security. Only a few respondents provided scores close to the lower threshold (65–68), although these values still fall within the acceptable category.

The maximum score of 92.5 indicates that there is a group of users who find the application highly intuitive, with clear navigation flow and fast system responses. This may suggest that the application design is already highly optimized for certain user profiles.

*3) SUS Category Interpretation*

The maximum score of 92.5 indicates that there is a group of users who find the application highly intuitive, with clear navigation flow and fast system responses. This may suggest that the application design is already highly optimized for certain user profiles.

- 68–80.3 = Good / Acceptable
- 80.3 = Excellent

Thus, the application is categorized as Good, meaning that it:

- Can be used easily by the general public without requiring special training.
- Provides interactions that are easy to understand, resulting in a relatively low cognitive load for users.
- Demonstrates good user satisfaction, particularly in navigation, interface clarity, and the effectiveness of the application's core functions.

The "Good" category also implies that the application is sufficiently competitive compared to similar applications in terms of usability and only requires minor improvements to reach the "Excellent" level.

*4) Qualitative Analysis Based on Rating Patterns*

Although the overall score falls within the "Good" category, respondent rating patterns reveal several important points:

- Ease of use during first-time interaction is rated highly by most users, indicating that the application interface is quite intuitive.
- IoT function integration performs well, with few users experiencing confusion during the device pairing or monitoring process.

- Visual consistency also received positive responses, suggesting that the interface design is comfortable and facilitates navigation.
- Some respondents indicated the need for minor improvements in notification speed and the clarity of certain icons, although these issues did not significantly affect the overall score.

## V. CONCLUSION

The SUS score in the Good category implies that the application has successfully achieved an acceptable level of usability and can serve as a solid foundation for further development. However, to reach the Excellent category, several areas can be improved, including:

1. Optimizing response time in updating IoT sensor status.
2. Refining icon designs to make them more understandable for general users.
3. Simplifying several menus that are considered less frequently used.

Improvements in these aspects have the potential to increase the SUS score in the future and enhance the overall quality of the user experience.

This study demonstrates that the implementation of IoT-based security systems can be further expanded into a more comprehensive security framework. Beyond access control, the proposed system has the potential to be developed into an integrated residential security system by incorporating additional features, such as a panic button function that can be activated by residents and directly connected to security personnel. Furthermore, future development may include integration with existing infrastructure, such as residential CCTV systems, to enhance real-time monitoring, incident logging, and overall situational awareness.

## REFERENCES


[1] Pemerintah Republik Indonesia, *Undang-undang Republik Indonesia Nomor 1 Tahun 2011 Tentang Perumahan dan Kawasan Permukiman* . Indonesia: https://bphn.go.id/data/documents/11uu001.pdf, 2011.

[2] R. Prabowo and M. R. Lubis, "Tinjauan Hukum Pidana terhadap Kebijakan Keamanan di Perumahan Berdasarkan Peraturan Perumahan," *JBH: Jurnal Begawan Hukum*, vol. 3, no. 2, pp. 1–10, Sep. 2025, doi: https://doi.org/10.62951/jbh.v3i2.133.

[3] A. D. Ayuningtyas, "Pencurian Jadi Kejahatan Paling Masif di Indonesia," *GoodStats*, Jakarta, 2024.

[4] K. Rinaldi, D. Prayoga, and H. Mianita, "Enviromental Criminology : Penerapan Defensible Space Sebagai Alternatif Pencegahan Kejahatan," *Jurnal Hukum Pidana dan Kriminologi*, vol. 3, no. 1, pp. 14–29, Apr. 2022, doi: 10.51370/jhpk.v3i1.66.

[5] Z. Hidayati and M. Noviana, "KORELASI AKSES PERUMAHAN DAN KRIMINALITAS DI PERUMAHAN KOTA SAMARINDA," *Jurnal Kreatif : Desain Produk Industri dan Arsitektur*, vol. 1, no. 1, p. 10, Oct. 2020, doi: 10.46964/jkdpia.v1i1.113.

[6] G. Vardakis, G. Hatzivasilis, E. Koutsaki, and N. Papadakis, "Review of Smart-Home Security Using the Internet of Things," *Electronics (Basel)*, vol. 13, no. 16, p. 3343, Aug. 2024, doi: 10.3390/electronics13163343.

[7] L. Y. Rock, F. P. Tajudeen, and Y. W. Chung, "Usage and impact of the internet-of-things-based smart home technology: a quality-of-life perspective," *Univers Access Inf Soc*, vol. 23, no. 1, pp. 345–364, 2024, doi: 10.1007/s10209-022-00937-0.

[8] J. M. Haney, S. M. Furman, and Y. Acar, "Research report : User Perceptions of Smart Home Privacy and Security," Nov. 2020. doi: 10.6028/NIST.IR.8330.

[9] C. Zhou, Y. Qian, and J. Kaner, "A study on smart home use intention of elderly consumers based on technology acceptance models," *PLoS One*, vol. 19, no. 3, p. e0300574, Mar. 2024, doi: 10.1371/journal.pone.0300574.

[10] Erwin *et al.*, *Pengantar & Penerapan Internet of Things Konsep dasar & Penerapan IoT di berbagai Sektor* . Jambi: PT. Sonpedia Publishing Indonesia, 2023.

[11] L. P. I. Kharisma *et al.*, *Internet of Things Pengenalan dan Penerapan Teknologi Internet of Things*. Jambi: PT. Sonpedia Publishing Indonesia, 2024.

[12] A. Francois, A. Danilova, D. Jagna, D. Caraig, and V. Petrov, "How Do QR Codes Work?," Techslang . Accessed: Nov. 30, 2025. [Online]. Available: https://www.techslang.com/how-do-qr-codes-work/

[13] FORTINET, "What Is A QR Code?," FORTINET. Accessed: Nov. 30, 2025. [Online]. Available: https://www.fortinet.com/resources/cyberglossary/what-is-a-qr-code?

[14] The QR STOCK Editorial Team, "What is a QR Code (2D Code)? A Complete Guide from its Official Name to Creation Methods," QR STOCK. Accessed: Nov. 30, 2025. [Online]. Available: https://qr-stock.com/articles/knowledge/about-qrcode?

[15] ITU, "What is a Quick Response Code (QR Code)?," ITU Online. Accessed: Nov. 30, 2025. [Online]. Available: https://www.ituonline.com/tech-definitions/what-is-a-quick-response-code-qr-code/

[16] Malwarebytes Team, "QR Codes: How they work and how to stay safe," Malwarebytes . Accessed: Nov. 30, 2025. [Online]. Available: https://www.ituonline.com/tech-definitions/what-is-a-quick-response-code-qr-code/

[17] S. Roychoudhury, "What is a QR Code, how does it work, and are they safe?," The QR Code Generator. Accessed: Nov. 30, 2025. [Online]. Available: https://www.the-qrcode-generator.com/whats-a-qr-code?

[18] C. Bustos, "How Do QR Codes Work—A Simple Guide," QR.io. Accessed: Nov. 30, 2025. [Online]. Available: https://qr.io/blog/how-do-qr-codes-work/

[19] A. Bangor, P. Kortum, and J. Miller, "Determining what individual SUS scores mean: adding an adjective rating scale," *J. Usability Studies*, vol. 4, no. 3, pp. 114–123, May 2009.

[20] Ahire, Ahmad, Bhatt, and Patel, "Evaluating the usability of emerging digital interfaces using the System Usability Scale (SUS)," *Int J Hum Comput Stud*, 2021.

[21] Rempel and Moffatt, "Applying the System Usability Scale to playful and interactive systems: A domain validity analysis," *ACM Transactions on Computer–Human Interaction (TOCHI)*, vol. 29, no. 4, pp. 1–27, 2022.

[22] Yoon and Ji, "Psychometric evaluation of the System Usability Scale across device types," *Int J Hum Comput Interact*, vol. 40, no. 2, pp. 150–165, 2024.

[23] Singh, Kumar, and Verma, "Usability assessment of mobile game applications using the System Usability Scale," *Mobile Information Systems*, 2024.

[24] J. Hair, G. T. M. Hult, C. Ringle, and M. Sarstedt, *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)*. 2022

ISSN 2085-4552

# Customer Service Chat Application Design Raden Inten II Airport, Lampung

M. Alif Ridho Setiawan[1], Indra Gunawan[2], Mezan el-Khaeri Kesuma[3], Fiqih Satria[4]
[1234]Prodi Sistem Informasi, Universitas Negeri Islam Raden Intan Lampung, Bandar Lampung, Indonesia
mezan@radenintan.ac.id

Accepted 04 January 2026
Approved 22 January 2026

***Abstract*— This study aims to design a web-based customer service application with live chat and chatbot features that implement Retrieval-Augmented Generation (RAG) technology at Raden Inten II Airport, Lampung. The main problems encountered are the slow response of conventional services and limited access to information by users located far from the airport. The system development uses the Waterfall model which includes requirements analysis, system design, implementation, testing, and maintenance. The main features in the application include user authentication, live chat, RAG chatbot, and data management through the admin dashboard. System testing was conducted using the black-box method, indicating that all features function according to specifications. The results of the study indicate that this system can significantly improve the efficiency and quality of customer service. For further development, it is recommended to use WebSocket to improve real-time communication performance.**

***Index Terms*— Customer Service; Live Chat; Chatbot; RAG; Web-based Application.**

## I. INTRODUCTION

The development of artificial intelligence (AI) technology has had a significant impact on increasing the efficiency of customer service, including in the air transportation sector.[1] Artificial Intelligence (AI) is a computer or machine technology that possesses human-like intelligence. One form of AI implementation in industry is the use of chatbots as a customer relations tool.[2] One rapidly developing technology is chatbots, AI-based software capable of interacting with users in real-time.[3] One modern approach to chatbot development is Retrieval-Augmented Generation (RAG)[4], the chatbot in this study is built using a RAG-based model.[5] Retrieval-Augmented Generation technology is expected to be integrated into chatbots,[6] which combines relevant data retrieval and generative capabilities to generate contextual and accurate responses.[7]

Live chat has become an increasingly popular way to provide real-time service in today's customer service environment.[8] Live chat enhances communication between customers and agents. Live chat is crucial for creating a positive customer experience. Good communication can enable information exchange, problem-solving, and relationship building.[9]

Public service organizations are experiencing an increase in digital requests from citizens (e.g., web pages, social media, or service apps), which has increased during times of social distancing. To effectively address this surge, we need to combine existing resources with new ways that go beyond existing service models. Chatbots are an example of an Artificial Intelligence application that has been widely used to address the surge in service demand, performing communication tasks previously performed by humans.

As one of the main transportation hubs in Lampung Province, Raden Inten II Airport in Lampung faces challenges in providing responsive and informative customer service. Conventional systems that rely on direct interaction often result in delays in providing information and services. Web-based applications offer numerous benefits and advantages for supporting business and service operations. They provide cross-platform accessibility, allowing users to access services anytime and anywhere via browsers without the need for installation. They are also easier to update and maintain, ensure consistent performance across devices, and enable seamless integration with various digital services such as databases, payment systems, and customer support tools[10], [11].

To address these challenges, developing customer service chat applications is a potential solution. Many companies are also starting to provide chatbot features to automate communications with people who use computers as a tool to interact with customers.[12] RAG technology enables chatbots to provide more accurate and relevant responses by leveraging existing databases and dynamically generating answers[13], [14], [15]. With this technology, customers can obtain information and assistance quickly and conveniently, accessible wherever they are. Good service with real-time feedback can improve customer satisfaction. The novelty of this study lies in three main aspects: (1) integration of a web-based live chat system with an AI chatbot into a single seamless platform; (2) the use of

the Gemini + RAG model to develop a domain-specific chatbot tailored to the operational and informational context of an airport; and (3) the implementation of a knowledge base system directly managed by administrators via a web interface, enabling real-time updates of information and responses.

The technical integration of RAG in this study is designed not merely for computational accuracy, but to address the operational challenges of customer service. Through text normalization and semantic search (FAISS/ChromaDB), the system is capable of comprehending informal customer complaints and mapping technical issues to the appropriate SOP solutions. Customer data security is ensured through anonymization mechanisms and local LLM execution (Ollama), while fallback and feedback loop features ensure that any complaints unresolved by the bot are escalated to human agents, thereby maintaining high customer satisfaction standards.

## II. METHOD

This research uses the Waterfall Model approach, namely a structured and sequential software development method. The Waterfall Model is the oldest and the most wellknown SDLC model. This model is widely used in government projects and in many major companies.[16] This model ensures that design errors can be detected before product development begins. It is well-suited for projects where quality control is a primary concern, as it emphasizes intensive documentation and planning. The stages include:

This study adopts the Waterfall model rather than iterative frameworks like Agile or Scrum. This choice is justified by the strict sequential dependency inherent in developing Retrieval-Augmented Generation (RAG) pipelines. The output quality of the Generator module (LLM) is strictly dependent on the performance of the Retriever module, which in turn relies on the integrity of the Data Preprocessing and Embedding phases.

Unlike Agile, which is optimized for volatile requirements, this project operates under fixed, critical constraints regarding data privacy and local execution. Consequently, the intensive System and Software Design phase provided by the Waterfall model is imperative to finalize the security architecture and vectorization schema before implementation begins, thereby minimizing the risk of costly architectural revisions later in the development cycle.

Although Waterfall is linear, we incorporated a feedback mechanism during the 'Operation and Maintenance' phase to allow for minor refinements based on user escalation data, providing a balance between structural rigidity and operational adaptability[17], [18], [19].

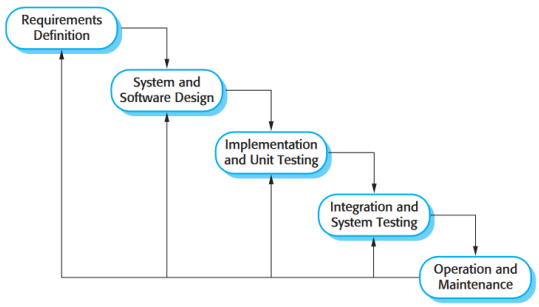



**Figure 1:** Waterfall Method

Figure 1 explain that:

- Requirements Definition: This stage encompasses the identification of customer complaint challenges (such as slang and typos) and data privacy requirements, which served as the foundation for selecting Local LLM technology.
- System and Software Design: This phase involves designing the technical RAG architecture, including the selection of PDF document chunking schemes and the vector database design (FAISS/ChromaDB).
- Implementation and Unit Testing: The coding stage (using FastAPI/Next.js backend) and isolated unit testing (e.g., verifying the correct functionality of text normalization).
- Integration and System Testing: A critical phase where the retriever module is integrated with the generator (LLM) and evaluated for accuracy (utilizing metrics such as BLEU/ROUGE)
- Operation and Maintenance: The deployment phase where the system actively serves customers. The feedback arrow indicates that data from real-world interactions (e.g., escalation tickets to the Admin) is utilized to update and refine the system in the initial phases.

## III. RESULT AND DISCUSSION

### *A. Requirement Definition*

Data was collected through observations and interviews with customer service personnel at Raden Inten II Airport, Lampung. It was found that users needed the following services:

1. User authentication
2. Live chat with admins
3. RAG-based chatbot
4. Admin dashboard for data management.

The purpose of this identification process was to ensure that the designed system truly met the needs of

end users, both customers and customer service officers.

*B. System and Software Design*

System design is done through use case diagrams, class diagrams, system architecture and system usage flow.

1. Use Case Diagram

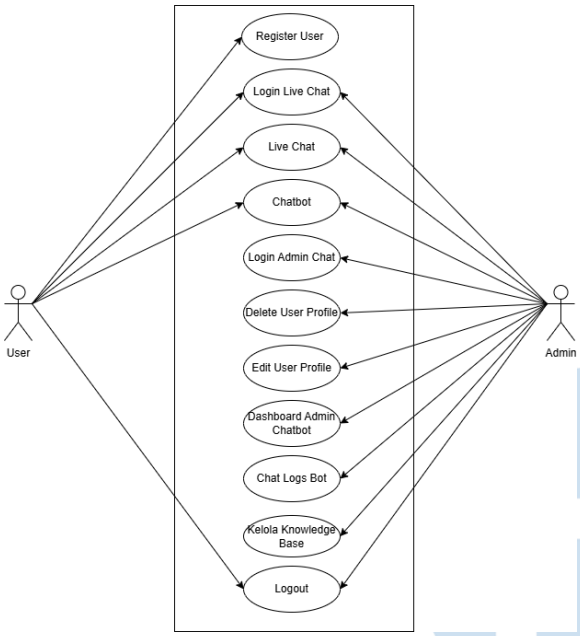



**Figure 2:** Usecase Diagram

Based on the use case diagram in the figure 2 above, the design of this customer service web application system involves two main actors: the User and the Admin. From the User perspective, the system provides essential functionality, including User Registration to create a new account and Live Chat Login to access the system. Once logged in, Users can choose to interact directly with staff via Live Chat or receive a quick, automated response from a Chatbot. The User session ends with a Logout function.

Admins, on the other hand, have more comprehensive access rights to manage the entire system. In addition to interacting with Users via Live Chat, Admins have a dedicated panel (Login Admin Chat) that grants them the authority to perform user management functions, such as Editing User Profiles and Deleting User Profiles. The Admin's core functionality is chatbot management, which includes access to the Chatbot Admin Dashboard to monitor activity, view conversation history (Chat Logs Bot), and most importantly, manage the knowledge base (Manage Knowledge Base), which serves as a source of information and answers for the chatbot. Like Users, Admins also have a Logout function to exit the system. Overall, this workflow is designed to create efficient and interactive customer service.

2. Class Diagram

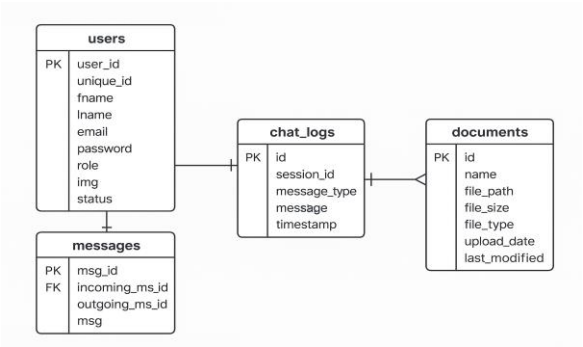



**Figure 3:** Class Diagram

The database structure used in designing this system consists of four main entities: users, messages, chat_logs, and documents. These four entities represent two main types of interactions in the system: live chat communication between users and admins, and Retrieval-Augmented Generation (RAG)-based chatbot interactions.[20], [21], [22], [23] The following explains the relationships between these entities:

a. Relationship between users and messages

The users table stores both user and system admin data. This table has a one-to-many relationship with the messages table, where a single user can send and receive multiple messages. This is represented by two foreign key attributes in messages: incoming_msg_id and outgoing_msg_id, each of which refers to the unique_id attribute in the users table. This relationship is necessary to support the two-way live chat feature between users and admins.

b. Relationship between users and chat_logs

The chat_logs table stores the history of interactions between users and the chatbot. Although there is no explicit foreign key to the users table, logically, each session_id in chat_logs is generated based on the identity of the currently active user. Therefore, this relationship is implicitly one-to-many, where a single user can have multiple chat sessions recorded in the system log.

c. Relationship between chat_logs and documents

In the RAG approach, chatbot responses are generated not only from the generative model but also from relevant documents retrieved from the documents table. Therefore, there is a many-to-one relationship from chat_logs to documents, reflecting that a single chatbot log entry can reference a single source document. Advanced implementations can extend this relationship to many-to-many, if a single chatbot response references more than one document.

d. The documents table as a knowledge base

The documents table stores the entire knowledge base that can be used by the chatbot. This entity does not have direct inbound or outbound foreign keys, but it is a crucial component in the RAG system's retrieval pipeline. Document files are stored along with metadata such as file name, size, type, and upload and update times.

With this ERD structure, the system is able to clearly separate user interactions (live chat) from interactions with the chatbot (chat_logs), and supports factual information retrieval through integration with the knowledge base.

3. System Architecture

The system architecture describes how the system is built from a technical and infrastructure perspective. This application is a web application consisting of several main components:

a. Client Side: The user and admin browsers serve as the main interface.
b. Application Server: The backend web server that processes requests from users/admins.
c. Live Chat Engine: Uses XMLHttpRequest for its live chat feature.
d. Chatbot Engine (LLM/RAG): The chatbot engine that generates automated responses using the integration of LLM models such as Gemini and the Retrieval-Augmented Generation approach.
e. Database: Stores user data, conversation history, and the chatbot's knowledge base.

4. System Usage Flow

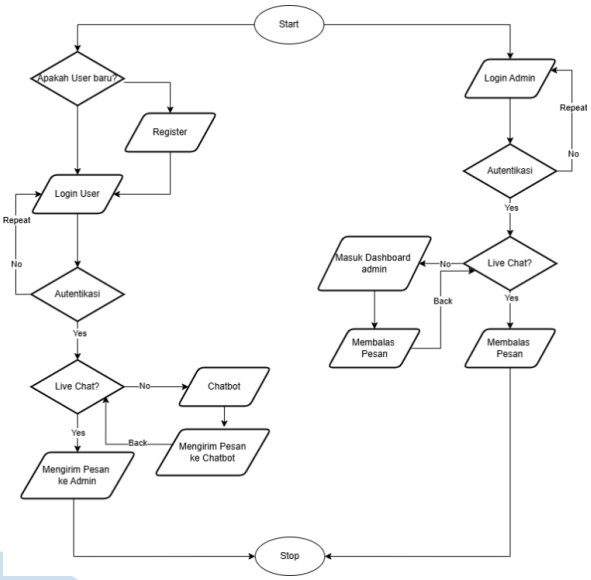



**Figure 4:** System Usage Flow

On figure 4 shows that System flowchart demonstrating the hybrid interaction flow with fallback mechanism to human agents.

*C. Implementation and Unit Testing*

After the system analysis and design are complete, the next stage is the implementation and testing of the customer service chat application. Implementation includes the design and coding phase, aligning with the business flow and integrating it with the database. The goal of this phase is to ensure that the customer service application functions properly according to specifications in a real-world environment. The implementation and testing of the customer service chat application are explained in the following page:

1. Register User Page

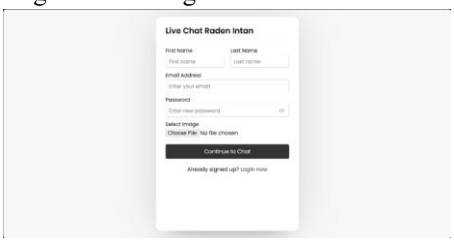



**Figure 5:** Register User page

The Figure 5 shows the registration page is used by new users to register before accessing the live chat or chatbot features. On this page, users are asked to enter their full name, email address, and

password, which will be used to log in to the system.

Once all the data has been entered, the user can click the Register button to save the data to the system. Validation is performed to ensure the email address is not duplicated and the password meets security standards.

2. Login User and Admin Page

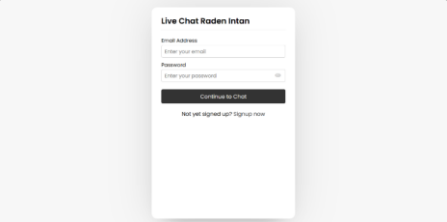



**Figure 6:** Login User and Admin

The figure shows After successfully registering, the user will be directed to the login page. This page is used for authentication by entering the email and password.

If the login data is valid, then the user will enter the main dashboard. If it is not valid, the system will display an error message.

3. Dashboard Chat User Page

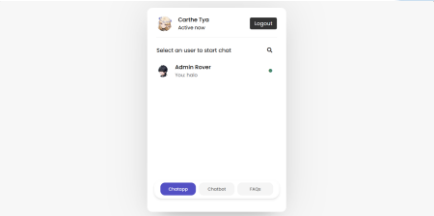



**Figure 7:** Dashboard Chat User Page

The figure shows The dashboard is the first page that appears after successfully logging in. It features a menu with options for live chat or a chatbot.

Users can choose the communication features they need. The interface is simple and intuitive for easy access.

4. Dashboard Chat Admin Page

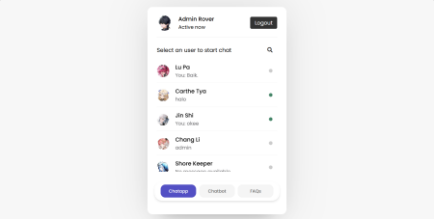



**Figure 8:** Dashboard Chat Admin Page

The figure shows The live chat admin dashboard page is the main interface for admins responsible for handling user messages in real-time.

Admins can view a list of recent conversations and access the live chat menu to respond to user messages directly.

5. Chat Interface Page

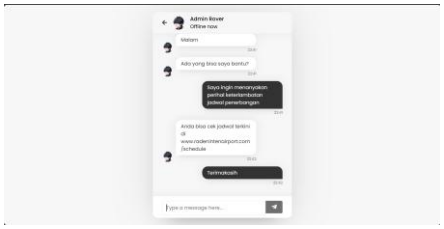


**Figure 9:** Chat Interface Page

The figure shows the page displays a conversation interface between the admin and the user. The admin can reply to user messages directly through the input box.

Messages are displayed chronologically and arranged like a chat window, making it easy for admins to monitor conversations in real-time.

6. Chatbot Page

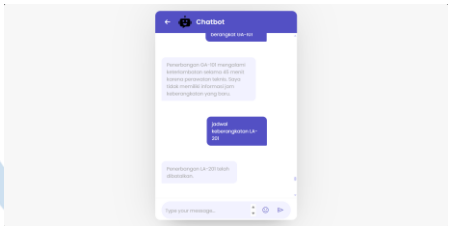



**Figure 10:** Chatbot Page

The figure shows The chatbot page allows users to interact with the system automatically. The chatbot works using a Retrieval-Augmented Generation (RAG) approach that can retrieve relevant documents and generate LLM-based answers.

Users simply type a question, and the system will respond quickly with accurate and contextual information.

7. Login Admin Chatbot Page

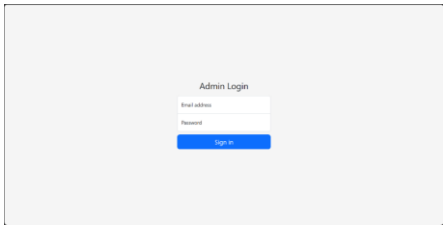



**Figure 11:** Login Admin Chatbot Page

The figure shows The login page is used by admins who manage the chatbot system. The login process requires valid credentials to access the chatbot dashboard.

8. Dashboard Admin Bot Page

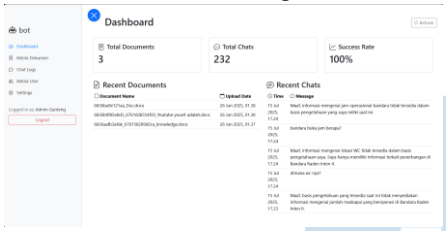



**Figure 12:** Dashboard Admin Bot Page

The figure shows The dashboard presents information such as API status, number of user interactions with the chatbot, and performance logs.

Admins can monitor the efficiency of chatbot responses and perform troubleshooting if necessary.

9. Management Knowledge Base Bot Page

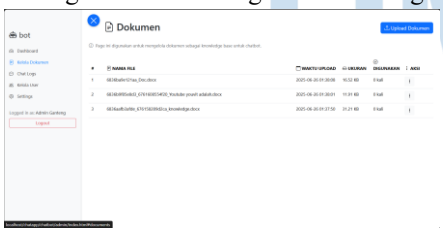



**Figure 13:** Management Knowledge Base Bot Page

The figure shows page allows you to upload, view, delete, or update documents that serve as sources of information for the chatbot during the retrieval process.

The managed documents will be used by the RAG system to factually answer user questions.

10. Management Chat Logs Bot Page

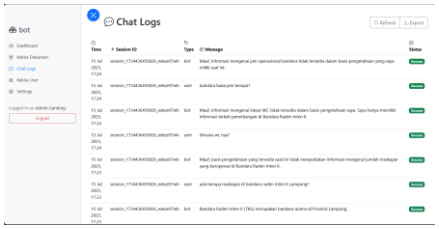



**Figure 14:** Management Chat Logs Bot Page

The figure shows Admins can monitor the entire history of user interactions with the chatbot through this page. Log data includes user input, chatbot output, interaction time, and success status.

This feature helps admins evaluate chatbot performance and identify responses that need to be improved or followed up.

11. Management User Page

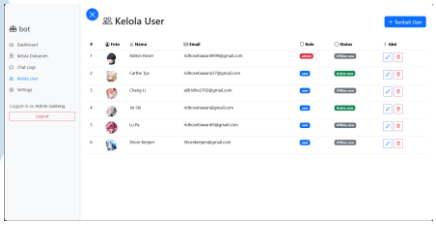



**Figure 15:** Management User Page

The figure shows Admin can manage available users.

12. API Management

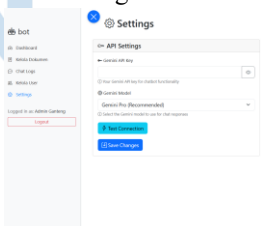



**Figure 16:** API Management

The figure shows Admin can manage the Gemini API if an error occurs.

### *D. Integration and System Testing*

Testing is done using the black-box method. System testing is crucial because everyone makes mistakes when creating software. The errors in each piece of software will be different.[24] Therefore, we use blackbox testing to verify the functionality of the:

1. Table 1: Testing of Login Page

| No | Skenario Test | Input | Test Step | Expected Output | Test Result | Status |
|---|---|---|---|---|---|---|
| 1 | Login with valid data | Valid Username & Password | Enter input and click login | User is redirected to the chat dashboard page | Output is correct | Successful |
| 2 | Login does not match valid data | Valid Username/Password | Enter input and click login | Login failed error alert appears | Output is correct | Successful |
| 3 | Leave login input blank | Empty | Click login without entering input | Please fill in all fields | Output is correct | Successful |

2. Table 2: Testing of Live Chat

| No | Skenario Test | Input | Test Step | Expected Output | Test Result | Status |
|---|---|---|---|---|---|---|
| 1 | User sends message to admin | Message text | Type and send | Incoming message to admin | Output is correct | Successful |
| 2 | Admin replies to message | Message text | Type and send | Incoming message to user | Output is correct | Successful |
| 3 | Send empty message | Empty | Click send | Shown please fill in all fields | Output is correct | Successful |

3. Table 3: Testing of Chatbot (RAG)

| No | Skenario Test | Input | Test Step | Expected Output | Test Result | Status |
|---|---|---|---|---|---|---|
| 1 | User asks for flight schedule information | Jadwal hari ini? | Type and send | Chatbot responds with the schedule information provided in the link. | Output is correct | Successful |
| 2 | User asks about airports | Dimanakah bandara terletak? | Type and send | Chatbot responds with a relevant answer | Output is correct | Successful |
| 3 | Random question other than airports | Siapa presiden Indonesia? | Type and send | “Maaf, saya adalah chatbot layanan pelanggan yang menyediakan informasi penerbangan di Bandara Raden Inten II (TKG). Saya tidak memiliki informasi mengenai presiden Indonesia.” | Output is correct | Successful |

*E. Operation and Maintenance*

The system is monitored and evaluated regularly. User feedback serves as the basis for continuous improvement and development.

*F. System Features*

1. User Authentication: Users can register and log in to access the system.
2. Live Chat: Direct communication with a customer service representative for complex questions.
3. RAG Chatbot: An automated chatbot system that responds to questions based on relevant documents using a retrieval and generation approach.
4. Admin Dashboard: Provides user management, document management, chatbot settings, and conversation log monitoring.

*G. System Testing*

Table 4: The result of features tested

| No | Features tested | Result |
|---|---|---|
| **1** | Login | Valid |
| **2** | Registration | Valid |
| **3** | Live Chat | Valid |
| **4** | Chatbot RAG | Valid |
| **5** | Dashboard Admin | Valid |

All features functioned as intended. The chatbot successfully responded to questions with a high level of relevance, while live chat provided personalized assistance options when needed.

To complement functional (black-box) testing, this study implements system performance evaluation and human assessment to ensure the chatbot's operational readiness. Latency measurement is adopted from Albert & Voutama [13] to guarantee real-time responsiveness, which is crucial for customer satisfaction, while interface quality is evaluated using Lighthouse, as in the study by Pratama & Sisephaputra [14]. Furthermore, to mitigate the risk of fatal information hallucination in public services, human evaluation is conducted involving expert staff using Accuracy, Completeness, and Clarity parameters, referring to the methodology of Putro et al[15].

While initial unit testing confirmed the functional validity of all system modules (success/fail), this study further employed quantitative metrics to rigorously assess system performance. Response latency was measured to evaluate efficiency[13], ROUGE-L scores were calculated to quantify answer relevance[15], and Lighthouse metrics were utilized to assess interface quality[14], thereby providing a comprehensive, data-driven evaluation beyond simple binary validation.

*H. Finding*

The system architecture is engineered with a primary emphasis on operational resilience and contextual accuracy. The core mechanism initiates when user input undergoes processing via a normalization module to mitigate linguistic noise. Subsequently, the system activates a Dual-Path Routing mechanism (as illustrated in the Flowchart): the primary path leverages Retrieval-Augmented Generation to extract technical solutions from the vector knowledge base, while the secondary path functions as a safety net, automatically escalating tickets to the Admin interface should the AI resolution prove inadequate. This entire process executes within a local infrastructure to guarantee data sovereignty.

The integration of RAG into a chatbot system has proven effective in providing contextual information. Compared to conventional chatbot methods, RAG can reduce generic responses and provide more specific answers. However, the system lacks multilingual support and doesn't utilize real-time protocols like WebSocket, which could potentially be added in future development.

RAG is not merely a technological enhancement, but a strategic solution for operational challenges and data security. The transition to RAG provides three strategic leaps: First, From keyword matching to semantic understanding. Second, From generic responses to grounded, internal fact-based responses. Third, From cloud dependency to local infrastructure sovereignty (privacy-preserving)."

## IV. CONCLUSION

This research successfully designed and implemented a customer service web application at Raden Inten II Airport, Lampung, integrating live chat and the RAG chatbot. This system provides easy access to information, fast responses, and increased operational efficiency for customer service. Testing results showed that all features performed according to specifications.

Recommendations for further development could consider the proposed improvements for the system include using WebSocket to increase the speed and responsiveness of the live chat feature, adding multilingual capabilities to accommodate users from diverse linguistic backgrounds, and integrating the platform with mobile applications to achieve a broader reach and enhance accessibility.

## ACKNOWLEDGMENT

The author would like to express his gratitude to Allah SWT for His blessings so that this scientific article can be completed well. Thanks are also extended to Customer Service of Raden Inten II Airport Lampung who has provided support and data for research purposes, as well as to Mr. Indra Gunawan and Mr. Mezan el-Khaeri Kesuma as supervisors who

have guided him patiently. Thanks are also extended to his beloved parents and all parties who have helped directly or indirectly in the process of compiling this scientific article.

.

## REFERENCES


[1] N. Ali dan M. Alfayez, "The impact of E-CRM on customer loyalty in the airline industry: the mediating role of customer experience," *Cogent Business & Management*, vol. 11, no. 1, hlm. 2364838, Des 2024, doi: 10.1080/23311975.2024.2364838.

[2] T. Destiany dan Y. Iskandar, "ANALISIS CUSTOMER SERVICE MENGGUNAKAN CHATBOT BERBASIS ARTIFICIAL INTELLIGENCE (Suatu Studi pada Daya Motor Honda Ciamis)," vol. 4, 2022.

[3] C. A. Oktavia, "Implementasi Chatbot Menggunakan Dialogflow dan Messenger Untuk Layanan Customer Service Pada E-Commerce," *JIMP*, vol. 4, no. 3, Jan 2020, doi: 10.37438/jimp.v4i3.230.

[4] Y. Tribber dan M. Asfi, "Implementasi Retrieval Augmented Generation untuk Layanan Informasi Kampus dengan Chatbot Berbasis AI," 2024.

[5] Muhammad Irfan Syah, Nazruddin Safaat Harahap, Novriyanto, dan Suwanto Sanjaya, "PENERAPAN RETRIEVAL AUGEMENTED GENERATION MENGGUNAKAN LANGCHAIN DALAM PENGEMBANGAN SISTEM TANYA JAWAB HADIS BERBASIS WEB," *zn*, vol. 6, no. 2, hlm. 370–379, Mei 2024, doi: 10.31849/zn.v6i2.19940.

[6] Universitas Internasional Batam, Y. Christian, M. Siahaan, dan H. Hansvirgo, "Designing a Web-Based Light Novel Application with an LLM-Powered Chatbot Recommendation System Using Scrum Methodology," *jmika*, vol. 8, no. 2, hlm. 174–186, Okt 2024, doi: 10.46880/jmika.Vol8No2.pp174-186.

[7] J. Miao, C. Thongprayoon, S. Suppadungsuk, O. A. Garcia Valencia, dan W. Cheungpasitporn, "Integrating Retrieval-Augmented Generation with Large Language Models in Nephrology: Advancing Practical Applications," *Medicina*, vol. 60, no. 3, hlm. 445, Mar 2024, doi: 10.3390/medicina60030445.

[8] M. Adam, M. Wessel, dan A. Benlian, "AI-based chatbots in customer service and their effects on user compliance," *Electron Markets*, vol. 31, no. 2, hlm. 427–445, Jun 2021, doi: 10.1007/s12525-020-00414-7.

[9] J. Heinonen dan E. Sthapit, "Service Agent Driven Co-Created Caring in Chat-Based Customer Service Encounters," *Services Marketing Quarterly*, vol. 45, no. 1, hlm. 1–24, Jan 2024, doi: 10.1080/15332969.2023.2288733.

[10] A. Waworuntu, "Rancang Bangun Aplikasi e-Commerce Dropship Berbasis Web," *Ultimatics*, vol. 12, no. 2, hlm. 118–124, Des 2020, doi: 10.31937/ti.v12i2.1823.

[11] H. Wijaya, D. Supriyanti, dan A. Saefullah, "Penggunaan Teknologi Web 2.0 dan Dampak Perubahannya pada Aplikasi Website berbasis Rich Internet Application (RIA)," *Ultimatics*, vol. 9, no. 2, hlm. 72–81, Sep 2017, doi: 10.31937/ti.v9i2.621.

[12] F. E. Office, "AI-based Chatbot Service for Financial Industry," vol. 54, no. 2, 2018.

[13] G. D. Albert dan A. Voutama, "PENGEMBANGAN CHATBOT BERBASIS PDF MENGGUNAKAN LOCAL RETRIEVAL-AUGMENTED GENERATION (RAG) DAN OLLAMA," *JITET*, vol. 13, no. 2, Apr 2025, doi: 10.23960/jitet.v13i2.6361.

[14] I. I. R. Pratama dan B. Sisephaputra, "Pengembangan Sistem Helpdesk Menggunakan Chatbot Dengan Metode Retrieval-Augmented Generation (RAG)," *JINACS: Journal of Informatics and Computer Science*, vol. 06, no. 3, 2024.

[15] I. P. H. Putro, J. Antoni, M. K. Adhitya, N. A. Herawati, A. Purwarianti, dan N. P. Utama, "Retrieval-Augmented Generation (RAG) Chatbot for Handling Customer Complaints in the Energy Sector," *Jurnal Infomedia : Teknik Informatika, Multimedia, dan Jaringan*, vol. 10, no. 2, hlm. 105–111, 2025.

[16] "A_Comparison_Between_Three_SDLC_Models_W."

[17] P. G. S. C. Nugraha, "Rancang Bangun Sistem Informasi Software Point of Sale (Pos) Dengan Metode Waterfall Berbasis Web," *JST (Jurnal Sains dan Teknologi)*, vol. 10, no. 1, hlm. 92–103, 2021, doi: 10.23887/jstundiksha.v10i1.29748.

[18] F. Nurdiansyah, E. Daniati, dan A. Ristyawan, "Pengembangan Sistem Informasi Kasir Apotek Dengan Metode Waterfall," *EDUSAINTEK*, vol. 9, no. 3, hlm. 752–773, Agu 2022, doi: 10.47668/edusaintek.v9i3.550.

[19] R. Susanto dan A. D. Andriana, "Perbandingan model waterfall dan prototyping untuk pengembangan sistem informasi," *Majalah Ilmiah UNIKOM*, vol. 14, no. 1.

[20] "Methods Of Implementing Retrieval-Augmented Generation In Combination With Modern Large Language Models," *SNSUT*, vol. 7, no. 12, 2025, doi: 10.31673/2786-8362.2025.012843.

[21] G. D. Marcantonio, "Intelligenza artificiale, Large Language Models (LLMs) e Retrieval-Augmented Generation (RAG).1 Nuovi strumenti per l'accesso alle risorse archivistiche e bibliografiche," *Intelligenza artificiale*.

[22] X. Li *dkk.*, "Building A Coding Assistant via the Retrieval-Augmented Language Model," 2 November 2024, *arXiv*: arXiv:2410.16229. doi: 10.48550/arXiv.2410.16229.

[23] J. Chen, H. Lin, X. Han, dan L. Sun, "Benchmarking Large Language Models in Retrieval-Augmented Generation," *AAAI*, vol. 38, no. 16, hlm. 17754–17762, Mar 2024, doi: 10.1609/aaai.v38i16.29728.

[24] F. C. Ningrum, D. Suherman, S. Aryanti, H. A. Prasetya, dan A. Saifudin, "Pengujian Black Box pada Aplikasi Sistem Seleksi Sales Terbaik Menggunakan Teknik Equivalence Partitions," *JIUP*, vol. 4, no. 4, hlm. 125, Des 2019, doi: 10.32493/informatika.v4i4.3782.

ISSN 2085-4552

# Sales Prediction at PT.World Infinite Network: A Comparative Study Of Naïve Bayes and Adaptive Neuro-Fuzzy Inference System

Jenie Sundari[1], Aden Irman[2]
[1] Universitas Bina Sarana Informatika: Faculty of Technology and Informatics, Jakarta, Indonesia
[2] Universitas Nusa Mandiri Faculty of Technology and Informatics, Jakarta, Indonesia
[1] Jenie.jni@bsi.ac.id, [2]12190102@nusamandiri.ac.id

Accepted 08 January 2026
Approved 23 January 2026

***Abstract*— This study analyzes historical sales transaction data from PT. World Infinite Network to support sales prediction and pattern discovery in the IT product domain. The dataset consists of structured transaction records containing product categories, sales volume, and time-based attributes, which are processed using data mining techniques. Two predictive methods—Naïve Bayes and Adaptive Neuro-Fuzzy Inference System (ANFIS)—are applied to model sales trends and classify purchasing behavior. Naïve Bayes is utilized for probabilistic classification due to its computational efficiency, while ANFIS is employed to capture nonlinear relationships through fuzzy inference and neural learning. Model performance is evaluated using classification accuracy. Experimental results show that the Naïve Bayes algorithm achieves an accuracy of 19.05%, indicating limited predictive capability on the given dataset and suggesting that sales patterns exhibit high variability and weak conditional independence among features. Despite the low accuracy, the findings highlight the comparative behavior of probabilistic and neuro-fuzzy approaches and provide insights into the suitability of different data mining methods for sales prediction in small or noisy transactional datasets. This study contributes empirical evidence for method selection in sales analytics and offers a baseline for improving predictive performance through feature engineering and data enrichment.**

***Index Terms*— sales, adaptive, neuro, data mining.**

## I. INTRODUCTION

The development of digital technology has transformed communication patterns through the emergence of new media that accelerate information dissemination. This shift promotes a digital-oriented lifestyle and marks the rise of the information society, supported by Information and Communication Technology (ICT) infrastructure. Broader access to information enhances public participation through more open two-way communication and feedback mechanisms. [1].

One of the techniques for data processing is **data mining**, which aims to discover connections between data that are not yet known to users and to make them easily accessible. Based on these informational relationships, decisions can be derived. Description, estimation, prediction, classification, clustering, and association are among the main categories that constitute data mining. In short, classification refers to assigning an individual or entity into a specific category, generally referred to as a "class." According to previous studies, the **Naïve Bayes** method possesses several advantages: it performs rapid calculations, uses a simple algorithm, and yields highly accurate results. These features make it one of the most widely used methods in classification. To estimate the parameters required for classification, the Naïve Bayes method needs only a small amount of training data [2].

The **Apriori algorithm** is used to calculate association rules between objects, which indicate how two or more items interact with one another. In other words, association rules describe how two or more products are related—for example, evaluating whether customers who buy product A are also likely to buy product B. This algorithm, based on association rules, was developed by R. Agrawal and R. Srikant in 1994, specifically for analyzing shopping cart data and identifying products that are frequently purchased together. In the healthcare field, Apriori can also be applied to determine patient responses to medications.

Data mining using the Apriori algorithm is designed to find data that most frequently appear in a database. Event information within the database is recorded as items [3].

**PT. World Infinite Network** is a system integrator company that provides hardware, software, and maintenance services for information technology (IT) devices. Since its establishment in 2011, PT. World Infinite Network has collaborated with various

governmental and non-governmental institutions. Some of its government partners include the Department of Communication, Informatics, and Statistics (DISKOMINFOTIK) of the DKI Jakarta Provincial Government, the Human Resources Division of the Indonesian National Police (SSDM Polri), the Central Statistics Agency (BPS), Bank Negara Indonesia (BNI), Angkasa Pura, POS Indonesia, and the National Traffic Management Center (NTMC) of the Indonesian National Police. The company's product and service offerings include servers, personal computers (PCs), notebooks, laptops, storage devices, IT peripherals, and maintenance services [4]

## II. THEORY

**Prediction** is a systematic method used to estimate future events based on past and present information [5]. Prediction is employed to obtain insights about future changes that may affect policy implementation and its consequences. For example, governments may use prediction to estimate the impact of economic policies on economic growth, inflation, and unemployment rates [6].

**Sales** refer to an integrated effort in designing strategies aimed at fulfilling the needs and desires of buyers, with the goal of achieving profitable transactions. Sales serve as the primary source of revenue for a company, as profits are generated from these activities [7].

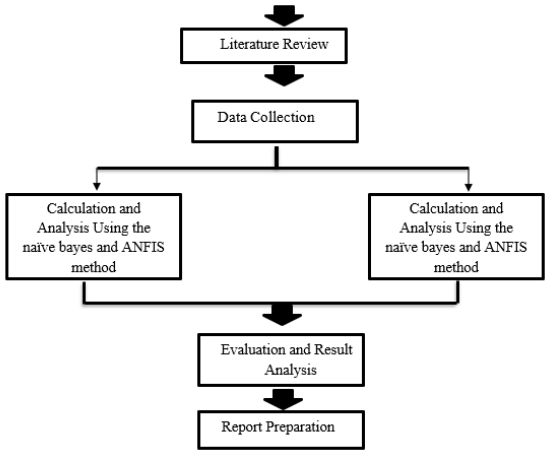
**Data mining** is an instrument that enables users to quickly access large volumes of data by extracting information from massive datasets through statistical, mathematical, and artificial intelligence approaches [8]. It is a method in the field of computer science used to discover knowledge from data, transforming it into useful and meaningful information.

In pattern recognition, the Bayesian Decision Theorem is an important statistical technique. Based on a simplifying assumption, the **Naïve Bayes** method assumes that if the output value is given to the attributes, these attributes are conditionally independent of one another. For categorical data features, Naïve Bayes is easy to apply—for example, when calculating the "gender" feature with values such as "male" and "female." However, additional methods are required before applying Naïve Bayes to handle numerical features. Consequently, the calculation of probability (likelihood) values for numerical (continuous) features differs from that for categorical (discrete) features[9].

The **Adaptive Neuro-Fuzzy Inference System (ANFIS)** is a hybrid approach that combines a fuzzy inference system with the adaptive learning capabilities of an artificial neural network. Its purpose is to create a model capable of capturing the relationship between the input and output of a complex system, even when such relationships are difficult to describe mathematically[10]

## III. METHOD

The research process serves as an essential stage to facilitate the study, as it follows a structured, standardized, sequential, systematic, and logical framework. A systematic research procedure helps researchers conduct their studies effectively and in a well-directed manner. The following are the methods employed by the authors.




## IV. RESULT AND DISCUSSION

In conducting this research, the sample data in the **Sales Data Table of PT. World Infinite Network** consisted of 22 product items, along with sales data covering a two-year period (2021–2022). The use of the **ANFIS** aimed to fulfill all the rules defined within the algorithm by meeting the minimum **support** and combining each item within the database, as well as the minimum **confidence** requirement, which represents the strength of the relationship between association rules and items.

Data processing was also carried out using the **Naïve Bayes Algorithm** in accordance with the standard implementation procedures of the algorithm. This approach enables the company to determine appropriate strategies based on the calculation results derived from the database that has been implemented using both the ANFIS and Naïve Bayes algorithms.

The following is a list of product items from the sales data at **PT. World Infinite Network:**

Table 1. List of Product Sales Items

| No | Nama Item | Kode |
|---|---|---|
| 1 | Dell Latitude 3420 | DL3 |
| 2 | Dell Latitude 5420 | DL5 |

| | | |
|---|---|---|
| 3 | Dell Latitude 3590 | DL35 |
| 4 | Acer Travelmate P214 | ATP214 |
| 5 | HP Envy X360 | HPEX3 |
| 6 | HP Probook 430 | HPP43 |
| 7 | HP Pav Plus 14 | HPPP14 |
| 8 | Lenovo V14 | LV14 |
| 9 | Lenovo Thinkpad X1 Carbon | LTPX1C |
| 10 | Lenovo Ideapad S530 | LIDS530 |
| 11 | Acer VM4960 | AVM4960 |
| 12 | Acer Veriton | AV |
| 13 | Dell Optiplex 3000 SFF | DO3000 |
| 14 | Dell Optiplex 5260 | DO5260 |
| 15 | HP Pro A G3 | HPPAG3 |
| 16 | Lenovo Thinksystem ST250 | LTST250 |
| 17 | Dell PowerEdge R740 | DPR740 |
| 18 | Dell PowerEdge R540 | DPER540 |
| 19 | Microsoft Office Home And Business 2019 | MOHB2019 |
| 20 | Microsoft Office 365 | MO365 |
| 21 | Backpack | BPCK |

Source: Research Results, 2023

Table 2. data training

| Item | Brand | Type | Sales |
|---|---|---|---|
| **DL3** | Dell | Notebook | 38 |
| **DL5** | Dell | Notebook | 22 |
| **DL35** | Dell | Notebook | 16 |
| **ATP214** | Acer | Notebook | 9 |
| **HPEX3** | HP | Notebook | 13 |
| **HPP43** | HP | Notebook | 18 |
| **HPPP14** | HP | Notebook | 14 |
| **LV14** | Lenovo | Notebook | 21 |
| **LTPX1C** | Lenovo | Notebook | 6 |
| **LIDS530** | Lenovo | Notebook | 30 |
| **AVM4960** | Acer | PC | 112 |
| **AV** | Acer | PC | 1 |
| **DO3000** | Dell | PC | 6 |
| **DO5260** | Dell | PC | 2 |
| **HPPAG3** | HP | Server | 2 |
| **LTST250** | Lenovo | Server | 6 |
| **DPR740** | Dell | Server | 0 |
| **DPER540** | Dell | Server | 13 |

Table 3. data testing

| Item | Brand | Type | Sales |
|---|---|---|---|
| **DL3** | Dell | Notebook | 8 |
| **DL5** | Dell | Notebook | 4 |
| **DL35** | Dell | Notebook | 5 |
| **ATP214** | Acer | Notebook | 9 |
| **HPEX3** | HP | Notebook | 2 |
| **HPP43** | HP | Notebook | 1 |
| **HPPP14** | HP | Notebook | 1 |
| **LV14** | Lenovo | Notebook | 0 |
| **LTPX1C** | Lenovo | Notebook | 0 |
| **LIDS530** | Lenovo | Notebook | 90 |
| **AVM4960** | Acer | PC | 10 |
| **AV** | Acer | PC | 4 |
| **DO3000** | Dell | PC | 3 |
| **DO5260** | Dell | PC | 0 |
| **HPPAG3** | HP | Server | 2 |
| **LTST250** | Lenovo | Server | 4 |
| **DPR740** | Dell | Server | 0 |
| **DPER540** | Dell | Server | 1 |
| **MOHB2019** | Microsoft | Software | 4 |
| **MO365** | Microsoft | Software | 1 |
| **BPCK** | Unbrand | Accessories | 120 |

In the data grouping process, the dataset was divided into a ratio of **80% for training data** and **20% for testing data**.

Table 4 naïve bayes calculation

| | true_DL3 | true_DL5 | true_ATP | true_HPP | true_HPE | true_HPPP | true_LV14 | true_LTP | true_LIDS530 |
|---|---|---|---|---|---|---|---|---|---|
| pred_DL3 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| pred_DL5 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| pred_DL35 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| pred_ATP | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| pred_HPEX3 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| pred_HPP43 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| pred_HPPP14 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| pred_LV14 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| pred_LTPX1C | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| pred_LIDS530 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Based on the image above, the results of the Naïve Bayes Algorithm calculation using RapidMiner Studio show an outcome of **19.05%**.

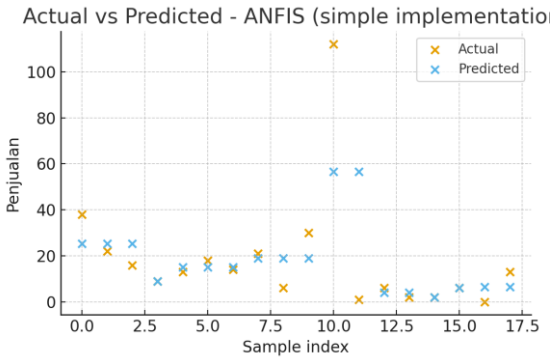



figure 1. ANFIS prediction

Model: Simple ANFIS (2 inputs: Brand & Item Type; 2 Gaussian MFs per input → 4 rules; first-order consequent).

Training method: Alternating Least Squares for consequent parameters combined with numerical gradient descent for premise parameters (mean & sigma).

Final RMSE (Root Mean Square Error): ≈ **19.4646** (sales units).

## V. CONCLUSION

Based on the experimental results, the Naïve Bayes algorithm achieved an accuracy of **19.05%**, indicating limited predictive capability for this dataset. In contrast, the **Simple ANFIS model** with two input variables (Brand and Item Type), four fuzzy rules, and first-order consequents produced a **Final RMSE of approximately 19.4646 sales units**. This suggests that the ANFIS model was able to represent the relationship between inputs and sales output with a moderate level of error. Overall, the ANFIS approach provides a more flexible and adaptive modeling framework compared to the Naïve Bayes classifier for this type of sales prediction problem.

## REFERENCES

[1] I. Astrid Faidlatul Habibah, "The Information Society Era as an Impact of New Media," *J. Teknol. dan Inf. Bisnis*, vol. 3, no. 2, pp. 350–363, 2021, doi: https://doi.org/10.47233/jteksis.v3i2.255.

[2] S. J. P.Bhargavi, "Applying Naive Bayes Data Mining Technique for Classification of Agricultural Land Soils," *Int. J. Comput. Sci. Netw. Secur.*, vol. 9, no. 8, pp. 117–122, 2009.

[3] R. A. Mahessya and S. Indrawati, "Implementasi Metode Anfis Data Mining Dalam Menyeleksi Beasiswa Di Smpn 7 Sorolangun," *J. Process.*, vol. 12, no. 1, pp. 904–915, 2017, [Online]. Available: http://ejournal.stikom-db.ac.id/index.php/processor/article/view/379

[4] L. Genetic and A. Viewer, "Artificial life Genetic Algorithm Viewer 1 . 0," pp. 11–15.

[5] S. Adiguno, Y. Syahra, and M. Yetri, "Implementasi Data Mining C.45 LINEAR REGRESI DAN KMEANS DENGAN MENGGUNAKAN FRAMEWORK DJANGO PYTHON," *J. Sist. Inf. TGD*, vol. 1, no. 4, pp. 275–281, 2022.

[6] G. Galih, "Data Mining di Bidang Pendidikan untuk Analisa Prediksi Kinerja Mahasiswa dengan Komparasi 2 Model Klasifikasi pada STMIK Jabar," *J. Teknol. Sist. Inf. dan Apl.*, vol. 2, no. 1, p. 23, 2019, doi: 10.32493/jtsi.v2i1.2643.

[7] A. Safitri Sembiring, N. Laila, and A. Wahyuni Lubis, "Analisis Harga Pokok Penjualan dan Laba Kontribusi terhadap Volume Penjualan pada Perum Bulog Divre Sumut," *ILTIZAM J. Shariah Econ. Res.*, vol. 7, no. 1, pp. 109–123, 2023, doi: 10.30631/iltizam.v7i1.1841.

[8] M. S. Mochammad Faid, Ahmad Supri, "Implementasi Data Mining C.45 LINEAR REGRESI DAN KMEANS DENGAN MENGGUNAKAN FRAMEWORK DJANGO PYTHON," pp. 2–3, 2023.

[9] A. Virrayyani and S. Sutikno, "Prediksi Penjualan Barang Menggunakan Metode Adaptive Neuro-Fuzzy Inference System (ANFIS)," *Khazanah Inform. J. Ilmu Komput. dan Inform.*, vol. 2, no. 2, pp. 57–63, 2016, doi: 10.23917/khif.v2i2.2554.

[10] Fahrillah Fahrillah and Zaehol Fatah, "Pengelompokan Data Nilai Siswa Madrasah Ta'Hiliyah Menggunakan Metode K-Means Clustering," *J. Ris. Sist. Inf.*, vol. 2, no. 1, pp. 53–59, 2025, doi: 10.69714/0v1pkz05.

ISSN 2085-4552

# Hybrid V-Net and Swin Transformer–Based Deep Learning Model for Brain Tumor Segmentation in Low-Quality MRI

Fajar Astuti Hermawati[1], Andre Pramudya[2]
[1,2] Departement of Informatics, Universitas 17 Agustus 1945 Surabaya, Surabaya, Indonesia
[1] fajarastuti@untag-sby.ac.id, [2]andrepramudya3@gmail.com

Accepted 04 January 2026
Approved 23 January 2026

***Abstract*— Brain tumor segmentation from low-quality magnetic resonance imaging (MRI) remains a challenging task due to noise, resolution variation, and low contrast between tumor and healthy tissues. Improving segmentation accuracy is essential to support more precise diagnosis and treatment planning. This study proposes a hybrid deep learning model that integrates V-Net and Swin Transformer–based architecture (Swin UNETR) for automatic brain tumor segmentation in multimodal MRI images. The MICCAI BraTS 2020 dataset was used, consisting of T1, T1c, T2, and FLAIR sequences with corresponding segmentation labels. The preprocessing pipeline includes resampling, skull stripping, intensity normalization, and data augmentation. V-Net is employed to extract local spatial features from 3D volumetric data, while the Swin UNETR captures global spatial relationships through a self-attention mechanism. Postprocessing procedures such as thresholding, morphological refinement, and false-positive removal are applied to enhance segmentation quality. The proposed hybrid model achieves Dice scores of 0.8635 for Whole Tumor (WT), 0.7179 for Tumor Core (TC), and 0.8073 for Enhancing Tumor (ET), with additional evaluation using precision, recall, and IoU further confirming its effectiveness. These results highlight the model's potential to improve automated brain tumor segmentation in low-quality MRI images and support its applicability as an efficient AI-assisted diagnostic tool in clinical practice.**

***Index Terms*— Brain Neoplasms; MRI; Deep Learning; Segmentation; V-Net, Transformer.**

## I. INTRODUCTION

Magnetic Resonance Imaging (MRI) is a non-invasive medical imaging technology that is crucial for detecting and diagnosing various diseases, particularly brain tumors. MRI's advantage lies in its ability to produce high-resolution images with good contrast against soft brain tissue. Imaging modalities such as T1-weighted, T2-weighted, and FLAIR can provide comprehensive information about brain structure [1]. The multimodal MRI approach has proven effective in improving diagnostic accuracy because each modality provides distinct information about the structure and morphology of brain tissue [2].

However, segmenting brain tumors from MRI images is a significant challenge. This is due to the complexity of tumor shape and size, irregular boundaries, differences in intensity between tissues, and the presence of noise and imaging artifacts [3]. Therefore, automated methods based on artificial intelligence, particularly deep learning, are needed to improve segmentation efficiency and accuracy.

Recent advancements in brain tumor segmentation and classification from MRI scans highlight the shift toward sophisticated deep learning and hybrid models. Early methods, like the one proposed by [4], utilized a classical approach combining a Modified Region Growing (MRG) algorithm for segmentation with Adaptive Support Vector Machine (ASVM) and Grasshopper Optimization Algorithm (GOA) feature selection to manage computational complexity. However, the field has rapidly moved toward Convolutional Neural Networks (CNNs) and Transformers. Key developments include 3D U-Net models for accurate volumetric segmentation [5], and advanced U-Net variants like the Trans U-Net [6] and UNETR [2], which leverage the Transformer's self-attention mechanism to capture long-range spatial dependencies. Further innovation includes hybrid approaches such as the 3D U-Net with Contextual Transformer and Double Attention [1], multi-pathway 3D FCNs for multimodal data fusion [7], and the introduction of computational efficiency techniques like QuantSR [8] for high-resolution medical imaging. These studies collectively demonstrate a trend of integrating advanced architectures and multi-modal data processing to achieve superior segmentation and classification accuracy for clinical application.

One deep learning architecture that has proven effective is the U-Net, which uses a symmetric encoder-decoder approach with skip connections. The U-Net performs well in medical image segmentation, but is less than optimal when handling images with high noise [9]. On the other hand, Swin UNETR is capable of

capturing global and local relationships in medical images, but requires significant computational resources [10]. Other approaches such as 3D U-Net and Modified Region Growing (MRG) have also been explored. 3D U-Net can process volumetric images, but still faces challenges when intensity is non-uniform [11]. V-Net, which uses a 3D convolutional neural network, is specifically designed for volumetric data such as MRI. V-Net is effective in understanding spatial context between layers, but has limitations in comprehensively capturing global features [12].

Combining V-Net with Swin UNETR in a hybrid approach is expected to overcome the weaknesses of each method. This combination allows for the integration of the local strengths of V-Net and the global strengths of Swin UNETR, thereby improving the accuracy of brain tumor segmentation in low-quality MRI images. This system uses multimodal image input (T1, T1c, T2, FLAIR) from the BraTS dataset that has undergone preprocessing stages, including noise removal, intensity normalization, and data augmentation. The system outputs a label map (mask) that clearly shows the brain tumor area. The segmentation results were then compared with ground truth to evaluate performance. With this approach, the research is expected to significantly contribute to the development of more accurate and efficient automated segmentation tools, accelerate medical diagnosis, and enrich the academic literature in the field of deep learning-based medical image segmentation.

## II. METHOD

### A. Data

The dataset used in this study is the MICCAI Brain Tumor Segmentation Challenge (BraTS) 2020 dataset, which provides multimodal MRI scans and expert-annotated ground truth labels for brain tumor segmentation [13]. The data consists of four main imaging modalities: T1, T1 with contrast (T1c), T2, and FLAIR, as shown in Fig.1. Segmentation labels are provided in three categories: Whole Tumor (WT), Tumor Core (TC), and Enhancing Tumor (ET).

The dataset described in Table I consists of two main parts: Training Data, used to train the segmentation model, and Validation Data, used to evaluate the model's performance. The dataset can be used for the development of deep learning-based segmentation methods because it provides tumor segmentation labels that include Whole Tumor (WT), Tumor Core (TC), and Enhancing Tumor (ET).

TABLE I. DATASET SPECIFICATIONS

| Specifications | Description |
|---|---|
| Amount of MRI Images | 369 total (295 for training and 74 for validation) |
| Amount of Image slices | 2,349 total images (1,847 for training and 502 for validation) |
| Resolution | 240×240×155 voxel. |
| Modalities | T1, T1c, T2, FLAIR. |
| Color Depth | 16-bit per channel. |
| Format | NIfTI (.nii). |

Computations were performed using a laptop with the following specifications: an Intel Core i5-13000 processor, an NVIDIA RTX4050 GPU, 24 GB of RAM, and Windows OS. Programming was performed using Python via the Google Colab and Jupyter Notebook platforms, with libraries such as PyTorch, MONAI, and Scikit-image for medical image processing and deep learning model implementation.

### B. The Proposed Methods

This research was conducted through several main stages systematically arranged to ensure optimal segmentation results, as shown in Fig. 2. These stages include data preprocessing, segmentation using a hybrid model, post-processing to refine the results, and testing scenarios to evaluate model performance. Each stage is interconnected and designed to address common issues encountered in segmenting low-quality MRI images, such as noise, intensity variations, and low contrast between tissues.

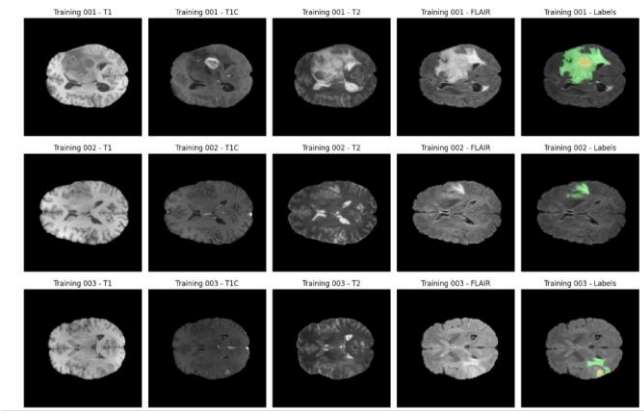



Fig. 1. Some Examples of Images from the BraTS 2020 MRI Dataset

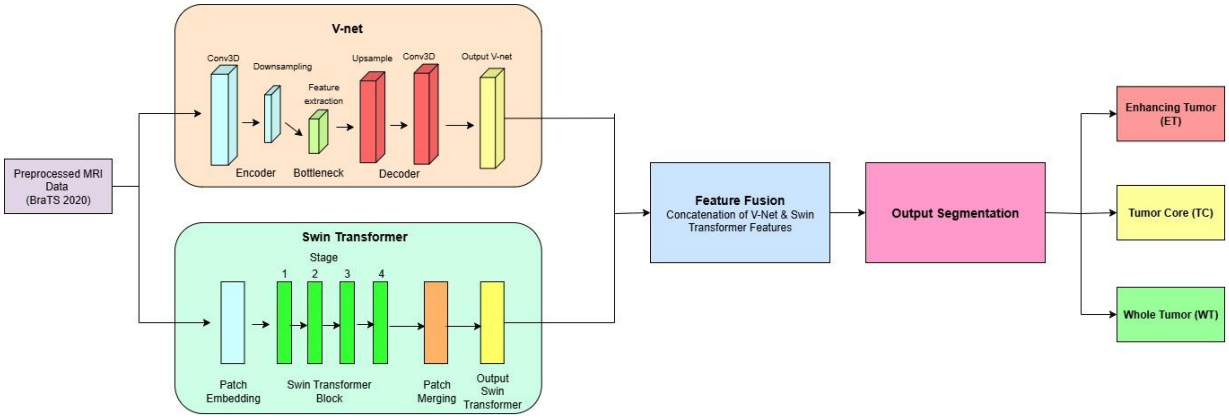



Fig. 2. Architecture Diagram of Hybrid Model of V-Net and Swin UNETR for Brain Tumor Segmentation

1). *Preprocessing*

Preprocessing is key to producing consistent and optimal input images. First, resampling is performed to standardize voxel resolution, given that the MRI data originate from different institutions. Next, skull stripping is performed using threshold-based and connected components algorithms to remove non-brain tissue [14]. The third step is intensity normalization, applying *z*-score normalization to each modality. Normalized intensity value, $I_{norm}$, is calculated based on (1) as follow:

$$I_{norm} = \frac{I-\mu}{\sigma} \quad (1)$$

Where *I* is an original voxel intensity value in the MRI image, $\mu$ is a mean intensity value within an MRI volume, $\sigma$ is a standard deviation of intensity within an MRI volume.

The modalities are then combined into a 3D tensor consisting of four channels. The dataset is then converted to tensor format for compatibility with deep learning architectures [15]. The final step is converting the image into a 3D tensor format ready for processing by the model. Once these steps are complete, the data is ready to be fed into the Hybrid V-Net and Swin UNETR, where V-Net handles local spatial features, while Swin UNETR focuses on broader spatial relationships. With proper preprocessing, the model is expected to perform more accurately in brain tumor segmentation.

2). *Segmentation Approach*

The hybrid V-Net and Swin UNETR model was designed using a dual-path approach. V-Net, as a 3D convolutional network, focuses on local spatial features through an encoder-decoder with skip connections [12]. Swin UNETR with a hierarchical architecture based on local self-attention, is used to extract global spatial context [10]. After feature extraction, feature fusion is performed through concatenation and convolution layers to combine the representations from both models. To combine the feature outputs from V-Net (local) and Swin UNETR (global), a concatenation operation is used followed by a 3D convolution to reduce the dimensionality and fuse the features, as shown in (2).

$$F_{fusion} = \text{Conv3D}([F_v||F_t] \quad (2)$$

$F_v$ is features extracted from the V-Net pipeline, which captures local spatial information from volumetric MRI (e.g., shape, texture around the tumor). $F_t$ is the extracted features from the Swin UNETR pipeline, which brings global context through a self-attention mechanism (long-range relations).

The training process uses *n* epochs, with a loss function based on a combination of Dice Loss and Cross Entropy Loss [16]. Equation (3) is formula of the Dice Loss as:

$$\mathcal{L}_{Dice} = 1 - Dice \quad (3)$$

where Dice is as presented in follow equation:

$$Dice = \frac{\sum_i p_i g_i + \epsilon}{\sum_i p_i + \sum_i g_i + \epsilon} \quad (4)$$

with $p_i$ = model prediction at the *i*-th voxel (0 or 1, or probabilistic 0–1), $g_i$ = ground truth at the *i*-th voxel (0 or 1), $\sum_i p_i g_i$= number of correctly detected voxels (intersection) and $\epsilon$ = small value to prevent division by zero.

Equation (5) is formula of Cross Entropy Loss:

$$\mathcal{L}_{CE} = -\sum_i \sum_{c=1}^{C} g_{i,c} \log(p_{i,c}) \quad (5)$$

with *C* = number of classes, $g_{i,c}$= 1 if the *i*-th voxel belongs to class *c* and 0 otherwise, $p_{i,c}$= predicted probability that the *i*-th voxel belongs to class *c*.

The final segmentation results are grouped into three sections, namely: Enhancing Tumor (ET) that is the active tissue after contrast, Tumor Core (TC) that is the

interior of the tumor without edema and Whole Tumor (WT) that is the entire tumor mass.

3). *Postprocessing*

The postprocessing stage begins with thresholding and probability masking, where a threshold value is set (usually between 0.3 and 0.7) to filter predictions based on the probabilities generated by a model, such as the Swin UNETR. Only voxels with probabilities above the threshold are considered valid, thus reducing noise and preventing minor misclassifications at the tumor edge [17]. Next, morphological refinement is performed using closing and dilation techniques to address contour roughness, fill small holes, and strengthen segmentation boundaries to align with the original anatomical structure [9]. This refinement is crucial because initial segmentation results are often discrete and not perfectly connected.

The third step is the removal of false positives, which are areas of the image incorrectly identified as tumors. This process uses Connected Component Analysis (CCA) to eliminate small predicted regions that are not spatially related to the main tumor structure, thereby increasing the model's specificity [13]. To refine the final results, smoothing using Gaussian or median filtering is applied, which is useful for smoothing segmentation edges and reducing unnatural intensity variations due to noise or unstable predictions [6]. This stage also improves the accuracy of volume measurements and facilitates 3D visualization.

As a final step, the segmentation results are converted into standard medical formats, namely NIfTI (.nii) and DICOM (.dcm). The NIfTI format is very commonly used in neuroimaging research because it is compatible with software such as FSL and SPM, while DICOM is a universal format in clinical medical practice and supports integration with hospital PACS systems [18]. This conversion makes the segmentation results ready for further analysis and clinical applications, bridging research findings with real-world applications.

4). *Testing Scenario*

The test scenario in this study was designed to evaluate the performance and reliability of a hybrid V-Net and Swin UNETR model in brain tumor segmentation based on the MICCAI BraTS 2020 dataset. The dataset includes various MRI imaging modalities such as T1, T1c, T2, and FLAIR, equipped with ground truth labels, allowing for objective evaluation of prediction accuracy. Initial testing was conducted by applying the trained model to validation data to measure the model's ability to identify and separate tumor structures from healthy brain tissue. Next, model performance was analyzed using evaluation metrics such as the Dice Score, Jaccard Index, sensitivity, and specificity. The Dice Score measures the similarity between the predicted segmentation and the reference label, while the Jaccard Index measures the degree of overlap between the two. Sensitivity assesses the model's ability to correctly detect tumors, while specificity assesses its accuracy in avoiding misclassification of healthy tissue. In addition to the Dice Score, other commonly used segmentation metrics, including Intersection over Union (IoU), Precision, Recall (Sensitivity), and Specificity, were employed to ensure broader comparability with existing brain tumor segmentation studies.

To assess the model's robustness to variations in image quality, testing was conducted on noisy or low-resolution MRI data. This testing is crucial for assessing the model's resilience under less-than-ideal imaging conditions. Furthermore, the resulting segmentations were also analyzed post-processing, using techniques such as morphological refinement and removal of false positives to ensure the final results were more accurate and freer from false predictions.

## III. Result and Discussions

### A. *Training Result*

The training process was carried out using a stepwise approach. Initially, the model was trained for 5 epochs to test the stability of the architecture and data pipeline. The results of this initial testing showed that the loss value was still relatively high and the segmentation performance was not optimal. The segmentation produced at this stage appeared coarse, with a very low Dice Score (WT = 0.159, TC = 0.0, ET = 0.0), and was not able to differentiate well between Whole Tumor (WT), Enhancing Tumor (ET), and Tumor Core (TC). After increasing the number of epochs from 5 to 40, there was an improvement in both the Loss and Dice score for predicting brain tumors, as seen in Fig. 3 and 4.

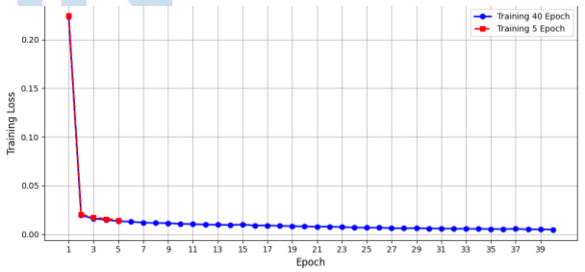



Fig. 3. Loss Comparison Graph between 5 epochs and 40 epochs.

Fig. 3 shows a comparison of the loss graphs between the model training for 5 epochs and 40 epochs. It can be seen that in the initial training with 5 epochs, the loss value decreased quite drastically but stopped before reaching stable convergence. Conversely, in the training for 40 epochs, the loss decrease was more consistent and sustained, reaching a value approaching 0.0048 at the end of the training. This graph shows that increasing the number of epochs provides a longer learning period for the model, allowing it to better

adjust its weights and resulting in more accurate and stable segmentation performance.

### *B. Segmentation Result*

Segmentation visualization for a representative MRI slice is shown in Fig. 4. For this particular sample, the model achieved Dice Scores of 0.8932 for Whole Tumor (WT), 0.9327 for Tumor Core (TC), and 0.8304 for Enhancing Tumor (ET). These values represent the segmentation quality on a single example image and are intended to illustrate the model's behavior visually. Table II summarizes the quantitative performance of the proposed model across multiple evaluation metrics, including Dice Similarity, Precision, and Recall, aggregated over the entire validation set.

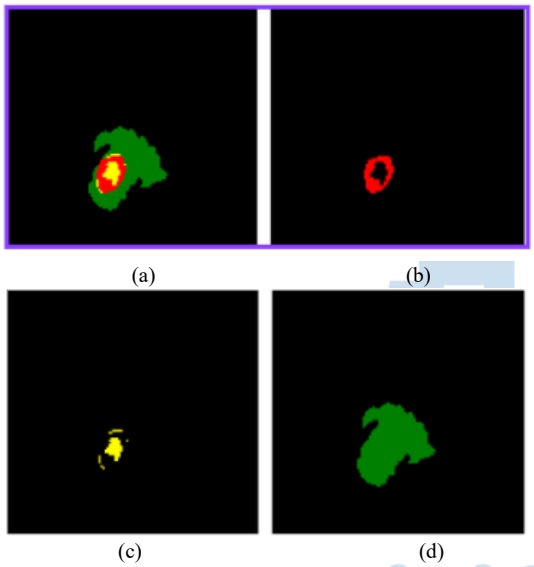



Fig. 4. Prediction results example for 40 epochs: (a) Prediction of the entire tumor, (b) Enhancing tumor, (c) Tumor core, (d) Whole tumor.

TABLE II. SEGMENTATION PERFORMANCE EVALUATION RESULTS FOR EACH BRAIN TUMOR SUBREGION MAP

| Brain Tumor Subregions | Dice Similarity | Precision | Recall |
|---|---|---|---|
| Tumor Core (TC) | 0.7179 | 0.8346 | 0.6675 |
| Whole Tumor (WT) | 0.8635 | 0.8622 | 0.8677 |
| Enhancing Tumor (ET) | 0.8073 | 0.7904 | 0.8392 |

Based on the test results, Whole Tumor (WT) achieved the highest scores across almost all evaluation metrics, with a Dice Score of 0.8635, Precision 0.8622, and Recall 0.8677. This indicates that the model is capable of identifying the entire tumor area with good accuracy and sensitivity.

Meanwhile, Enhancing Tumor (ET) also demonstrated quite solid performance with a Dice Score of 0.8073, indicating the model's ability to detect active tumor regions or those experiencing contrast enhancement following contrast agent administration in MRI. However, the relatively small variation in shape and size of ET compared to WT makes it more difficult to fully segment.

For the Tumor Core (TC), the Dice Score of 0.7179 indicates that the model still faces challenges in precisely detecting the tumor core. This may be due to the similarity in intensity between the TC and the surrounding tissue, as well as the more limited distribution of TC data compared to WT.

Overall, this evaluation results indicate that the hybrid V-Net and Swin UNETR approach is capable of providing competitive segmentation performance on low-quality MRI images. However, accuracy improvements, particularly for the TC segment, can still be achieved through strategies such as adding various data augmentations, adjusting the loss function (e.g., a combination of Dice Loss and Focal Loss), and implementing more adaptive post-processing techniques to reduce segmentation errors in small areas.

### *C. Qualitative Evaluation Results*

Visual evaluation was performed by displaying axial MRI image slices along with predicted segmentation results and ground truth labels. This visualization demonstrates that the model is able to map tumor areas with relatively accurate shapes, although there are minor inaccuracies at the edges of small tumors.

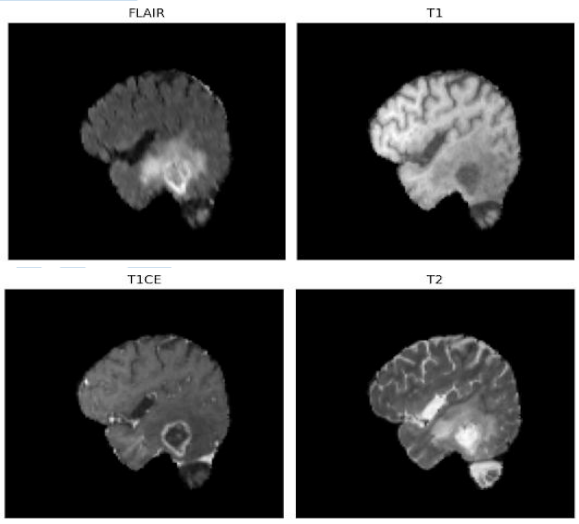



Fig. 5. MRI Modality Output: FLAIR, T1, T1CE, T2

Fig. 5 shows the four main MRI modalities that have undergone preprocessing and postprocessing, used as input for the segmentation process: FLAIR, T1, T1CE, and T2. Each modality provides different information about brain tissue structures, such as edema, active tumor contrast, and anatomical brain boundaries. The combination of these four modalities is crucial in providing a complete representation of various types of brain tumor tissue.

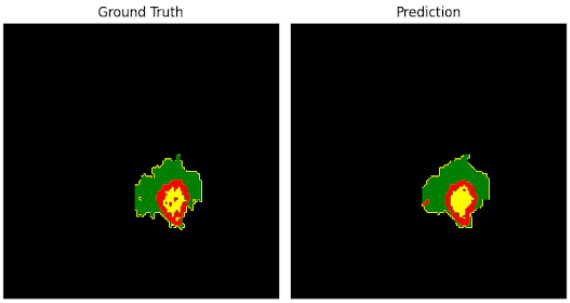



Fig. 6. Ground Truth and Segmentation

Fig. 6 displays a comparison between the segmentation results generated by the model and the ground truth labels. It can be seen that the model's predictions successfully follow the tumor shape and area quite accurately. Despite slight differences in tumor edges, the model was generally able to identify relevant tumor locations and shapes, including their internal structures, such as Tumor Core (TC) and Whole Tumor (WT).

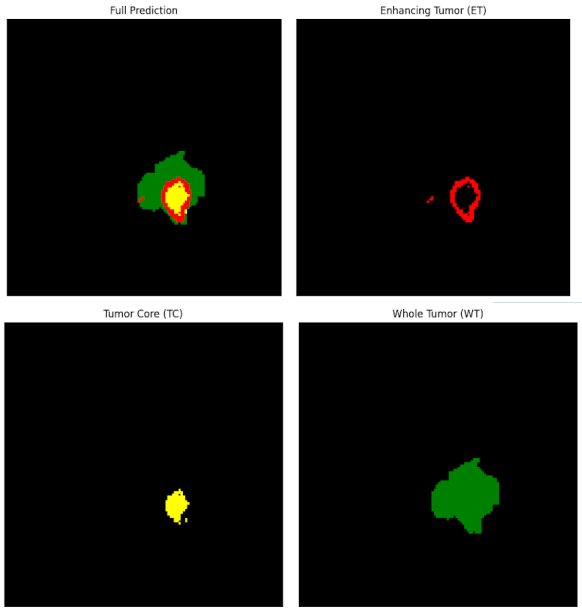



Fig. 7. Tumor Class Mask: ET (red), TC (yellow), WT (green)

Fig. 7 clarifies the classification of tumor classes predicted by the model. Red indicates the Enhancing Tumor (ET) region, yellow represents the Tumor Core (TC), and green indicates the Whole Tumor (WT). This mask helps assess how well the model can spatially distinguish and characterize each tumor subregion and highlights the model's ability to detect complex tumor structures with precise segmentation.

### *D. 3D Visualization*

As part of the qualitative evaluation, a three-dimensional visualization of the brain tumor segmentation results was performed using the *Plotly* library. The purpose of this visualization was to provide a comprehensive understanding of the spatial structure of the tumor predicted by the *FusionModel* model, while also more intuitively evaluating the accuracy, integrity, and distribution of each tumor component. The visualization was performed by mapping each voxel classified as tumor into 3D space based on the ($x$, $y$, $z$) coordinates of the *pred_mask*, which is the final segmentation prediction result. Each voxel is displayed as a point scatter in 3D space and assigned a different color to distinguish the tumor components: yellow for Tumor Core, green for Whole Tumor, and red for Enhancing Tumor.

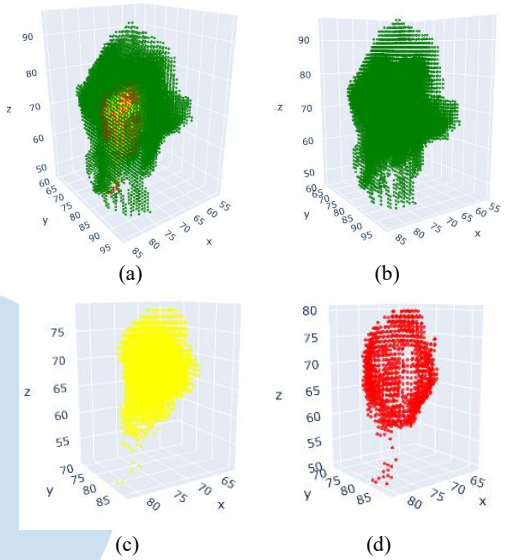



Fig. 8. 3D visualization of (a) the entire tumor (b) the Whole Tumor (c) the Tumor Core and (d) Enhancing Tumor (ET)

This visualization consists of four main sections. First, the full tumor prediction visualization displays all voxels classified as part of the tumor, colored based on their respective labels. This visualization illustrates the overall structure and distribution of the tumor, including irregular borders, asymmetric distribution, and areas of high density that may indicate tumor dominance. The clarity of the color labels allows identification of spatial relationships between tumor components and helps assess whether the model's predictions logically follow the biological pattern of the brain tumor.

The Whole Tumor (WT) visualization focuses on the entire tumor mass regardless of label type. In this stage, all voxels with a label greater than zero are displayed in green, reflecting the total size and shape of the tumor. This visualization is useful for evaluating whether the model covers the entire tumor volume as intended, or whether it is missing important areas (under-segmentation) or over-segmenting.

The third visualization displays only the Tumor Core (TC), the deepest area of the tumor, typically composed of dense tissue and crucial for diagnosis. The TC is displayed in yellow to highlight whether the model accurately and consistently identifies the tumor core, consistent with the general pattern of tumor

growth. The distinctive shape and location of the TC can help assess the potential for compression of vital brain structures.

The Enhancing Tumor (ET) was visualized, which is the area of the tumor that shows increased signal on T1-weighted contrast (T1c) imaging. This area is often associated with high levels of biological activity and increased vascularization, making it important for diagnosis and therapy planning. Voxels in the ET are visualized in red, which helps observe the distribution and potential aggressiveness of the tumor in 3D space.

Overall, this 3D visualization not only strengthens the quantitative evaluation results but also provides a more realistic spatial representation of model predictions, making it very useful for medical practitioners in understanding and analyzing brain tumor development more comprehensively.

## Discussion

### *A. Findings*

This study demonstrates that combining the Swin UNETR and V-Net architectures into a single hybrid model (Fusion model) can improve the accuracy of brain tumor segmentation in volumetric MRI images. Quantitative evaluation results demonstrate that the average Dice Score for three tumor types, Whole Tumor (WT), Tumor Core (TC), and Enhancing Tumor (ET), reaches 0.8635, 0.7179, and 0.8073, respectively. These values are considered high, indicating that the model successfully recognizes and maps tumor areas accurately, even in low-resolution MRI images.

Although several recent studies report higher Dice scores, particularly for Whole Tumor segmentation, these methods often rely on extensive architecture tuning, large-scale computational resources, or ideal imaging conditions. In contrast, the proposed hybrid V-Net and Swin UNETR model demonstrates balanced performance across Dice, Precision, and Recall metrics, especially under low-quality MRI conditions. This indicates that the proposed approach prioritizes robustness and generalizability rather than solely optimizing a single metric.

When compared conceptually to non-hybrid baselines, a standalone V-Net effectively captures local volumetric features but lacks global contextual awareness, often leading to fragmented boundaries. Conversely, Swin UNETR models emphasize global spatial relationships but may miss fine-grained local details critical for small tumor regions. The proposed hybrid architecture integrates both strengths, resulting in improved segmentation consistency across WT, TC, and ET regions.

This achievement aligns with a previous study by [2] which demonstrated that the use of a transformer-based architecture like Swin UNETR is able to capture global spatial context better than conventional CNN models, especially for 3D segmentation tasks. Furthermore, the V-Net-based encoder-decoder approach proved effective in extracting local spatial features from medical volumes, as also demonstrated by [12] in their original study on V-Net for internal organ segmentation.

The success of the fusion model in this study also demonstrates that an architectural ensemble approach can mitigate the weaknesses of each model when used alone. This is reinforced by findings [16] in nnU-Net, which suggest that appropriate architecture and pipeline adaptation, including fusion strategies and post-processing, significantly impact segmentation quality. Furthermore, qualitative evaluation through 3D visualization demonstrated that the model's predictions were not only numerically accurate but also morphologically and spatially consistent. The Tumor Core (TC) and Enhancing Tumor (ET) areas were successfully mapped with shapes and distributions consistent with the general biological structure of brain tumors.

However, several challenges and potential improvements remain. One is the reliance on training data, due to the lack of publicly available labels in the official BraTS validation set, evaluation was conducted using an internal validation split. This opens up the possibility of evaluation bias. Furthermore, some samples exhibited lower Dice scores in the Enhancing Tumor (ET) class, indicating that the model still has limitations in capturing small, mixed, low-contrast tumor areas. A similar finding was also reported by [19], who emphasized that ET segmentation is a major challenge because its intensity contrast often overlaps with normal tissue.

The implications of these findings are important in the context of developing AI-based clinical decision support systems. Fusion models such as those proposed in this study have the potential to be used as non-invasive and efficient early diagnostic tools, particularly for brain tumor screening in 3D images. However, further validation against external datasets and integration with feedback from healthcare professionals are needed for the system to be adopted clinically.

### *B. Limitation*

Although the results are promising, this study still has several limitations that should be considered for future development. While multiple evaluation metrics, including Dice, Precision, and Recall, were employed, the inclusion of additional distance-based metrics such as the Hausdorff Distance could provide deeper insight into boundary accuracy. Furthermore, clinical evaluation involving radiologists or specialist physicians would offer more comprehensive validation of the segmentation results.

A second limitation lies in the lack of segmentation labels in the validation data, so the evaluation was conducted only on the training data. This results in a lack of objective testing of the model's performance on data the model has never encountered before. Therefore, future research is recommended to perform manual splitting or add an external validation dataset for more comprehensive and accurate model evaluation.

Furthermore, the current loss formulation can be further improved by incorporating advanced loss combinations, such as Dice Loss with Focal Loss, to enhance sensitivity for small tumor regions. Combining *DiceLoss* with *CrossEntropyLoss* is recommended to improve segmentation performance, especially for minority classes. Finally, future research is expected to explore hyperparameter optimization in more depth, such as variations in learning rate and batch size, as well as the application of spatial data augmentation such as rotation, flipping, and elastic deformation. These steps have the potential to improve the model's generalization and robustness to variations in tumor shape and size in MRI images.

By considering these various suggestions and improvements, it is hoped that future research will produce a more reliable, accurate, and applicable brain tumor segmentation system in real-world clinical settings.

## IV. CONCLUSIONS

The hybrid V-Net-Swin UNETR model successfully improves brain tumor segmentation performance on the BraTS 2020 MRI dataset. By combining comprehensive preprocessing, a dual-path feature extraction strategy, and adaptive postprocessing, the model achieves Dice scores of 0.8635 for Whole Tumor (WT), 0.8073 for Enhancing Tumor (ET), and 0.7179 for Tumor Core (TC), demonstrating its ability to integrate local volumetric features with global contextual information effectively. These results highlight the model's potential as a reliable AI-based diagnostic support tool in clinical workflows. For future development, further validation on external datasets is needed to assess generalization across imaging protocols, along with enhancements in detecting small tumor regions through improved loss functions and augmentation strategies. Incorporating uncertainty estimation, developing lightweight versions for real-time or resource-limited settings, and enabling interactive or semi-supervised segmentation could also enhance clinical usability. Additionally, integrating imaging data with clinical or molecular information offers opportunities for more comprehensive tumor characterization.

## ACKNOWLEDGMENT

The authors would like to thank all parties who supported this research, particularly the Lembaga Penelitian dan Pengabdian Kepada Masyarakat (LPPM) of Universitas 17 Agustus 1945 Surabaya

## REFERENCES

[1] T. B. Nguyen-Tat, N. H. Nghia, and V. M. Ngo, "Enhancing Brain Tumor Segmentation in MRI Images: A Hybrid Approach Using UNet, Attention Mechanisms, and Transformers," Egyptian Informatics Journal, vol. 7, 2024, doi: 10.13140/RG.2.2.18164.36485.

[2] A. H. Nvidia et al., "UNETR: Transformers for 3D Medical Image Segmentation," arXiv:2003.10504v3, Oct. 2021, [Online]. Available: https://monai.io/research/unetr

[3] E. Sami, H. Ebied, S. Amin, and M. Hassaan, "Brain Tumor Segmentation Using Modified U-Net," Res Sq, no. preprint, May 2022, doi: 10.21203/rs.3.rs-1653006/v2.

[4] A. Srinivasa Reddy and P. Chenna Reddy, "MRI brain tumor segmentation and prediction using modified region growing and adaptive SVM," Soft comput, vol. 25, no. 5, pp. 4135–4148, Mar. 2021, doi: 10.1007/s00500-020-05493-4.

[5] P. Agrawal, N. Katal, and N. Hooda, "Segmentation and classification of brain tumor using 3D-UNet deep neural networks," International Journal of Cognitive Computing in Engineering, vol. 3, pp. 199–210, Jun. 2022, doi: 10.1016/j.ijcce.2022.11.001.

[6] J. Chen et al., "3D TransUNet: Advancing Medical Image Segmentation through Vision Transformers," arXiv:2310.07781v1, Oct. 2023, [Online]. doi: 10.48550/arXiv.2310.07781

[7] Y. Ding, L. Gong, M. Zhang, C. Li, and Z. Qin, "A multi-path adaptive fusion network for multi-modal brain tumor segmentation," Neurocomputing, vol. 412, pp. 19–30, Oct. 2020, doi: 10.1016/j.neucom.2020.06.078.

[8] H. Qin et al., "QuantSR: Accurate Low-bit Quantization for Efficient Image Super-Resolution," in 37th Conference on Neural Information Processing Systems (NeurIPS 2023), 2023. [Online]. Available: https://github.com/htqin/QuantSR.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in Navab N., Hornegger J., Wells W., Frangi A. (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015., Springer, Cham, 2015, ch. Lecture No, pp. 1–8.

[10] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021. doi: https://doi.org/10.1109/ICCV48922.2021.00986.

[11] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation," in International Conference on Medical Image Computing and Computer-Assisted Intervention, Athens, Greece, Jun. 2016. [Online]. Available: http://arxiv.org/abs/1606.06650

[12] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," arXiv:1606.04797v1, Jun. 2016, [Online]. Available: http://arxiv.org/abs/1606.04797

[13] B. H. Menze et al., "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," IEEE Trans Med Imaging, vol. 34, no. 10, pp. 1993–2024, Oct. 2015, doi: 10.1109/TMI.2014.2377694.

[14] A. Hoopes, J. S. Mora, A. V. Dalca, B. Fischl, and M. Hoffmann, "SynthStrip: skull-stripping for any brain image," Neuroimage, vol. 260, Oct. 2022, doi: 10.1016/j.neuroimage.2022.119474.

[15] M. Havaei et al., “Brain tumor segmentation with Deep Neural Networks,” Med Image Anal, vol. 35, pp. 18–31, Jan. 2017, doi: 10.1016/j.media.2016.05.004.

[16] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, “Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioin-formatics), Springer Verlag, 2017, pp. 240–248. doi: 10.1007/978-3-319-67558-9_28.

[17] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation,” Nat Methods, vol. 18, no. 2, pp. 203–211, Feb. 2021, doi: 10.1038/s41592-020-01008-z.

[18] A. A. Taha and A. Hanbury, “Metrics for evaluating 3D medical image segmentation: Analysis, se-lection, and tool,” BMC Med Imaging, vol. 15, no. 1, Aug. 2015, doi: 10.1186/s12880-015-0068-x.

[19] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, “UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation,” IEEE Trans Med Imaging, vol. 39, no. 6, pp. 1856–1867, Jun. 2020, doi: 10.1109/TMI.2019.2959609

ISSN 2085-4552

# Multiclass Emotion Detection on YouTube Comments Using IndoBERT

## A Web-Based Incremental Learning System with Multiple Data Split Evaluation

Naufal Syarifuddin[1], Nurirwan Saputra[2]
[1,2] Prodi Informatika, Fakultas Sains dan Teknologi, Universitas PGRI Yogyakarta
[1] naufalsrfdn@gmail.com, [2] nurirwan@upy.ac.id

Accepted 05 January 2026
Approved 23 January 2026

***Abstract*— YouTube comment sections provide rich textual data that reflect users' emotional responses to various social, political, and entertainment-related issues. However, the large volume of user-generated comments makes manual emotion analysis inefficient and impractical. This study proposes a multiclass emotion classification approach for Indonesian YouTube comments using the IndoBERT model integrated with a database-driven incremental learning system. Comment data were collected through the YouTube Data API and manually labeled into six emotion categories: anger, sadness, happiness, fear, surprise, and neutral. Text preprocessing included lowercasing, text cleaning, and normalization of informal Indonesian words. The IndoBERT model was fine-tuned using three training–testing split scenarios (60:40, 70:30, and 80:20) to evaluate model performance and stability. Experimental results indicate that the 80:20 split achieved the best performance with an accuracy of 68%. This performance is influenced by a highly imbalanced class distribution, where minority emotion classes such as fear (2%) and surprise (3%) are significantly underrepresented compared to dominant classes. In addition to classification performance, the proposed system incorporates a database to continuously store newly collected and validated data and supports incremental retraining, enabling the model to learn from new data without discarding previously acquired knowledge. This adaptive learning mechanism allows the system to improve over time and represents a key advantage over conventional static emotion classification approaches.**

***Index Terms*— *emotion classification; IndoBERT; incremental learning; social media analysis; YouTube comments.***

## I. INTRODUCTION

YouTube is one of the largest video-sharing platforms and has evolved into a major space for public interaction through its comment sections. Users actively express opinions, attitudes, and emotional reactions toward various social, political, and entertainment-related issues in textual form[1]. These emotional expressions reflect public perception and collective responses, making YouTube comments an important data source for large-scale emotion analysis and social media research [2], [3].

Despite their analytical potential, the massive and continuously growing volume of YouTube comments makes manual emotion analysis inefficient, time-consuming, and difficult to maintain consistently[4]. In addition, the language used in YouTube comments is highly informal, frequently containing slang, abbreviations, spelling variations, and rapidly evolving expressions[5]. These characteristics increase the complexity of emotion classification and pose significant challenges for automated text analysis systems [6].

Recent advances in Natural Language Processing (NLP) have enabled the automation of emotion analysis through machine learning techniques. Early studies commonly employed traditional classifiers such as Naïve Bayes and Support Vector Machine [7]. Although these approaches achieved moderate performance, they rely heavily on handcrafted features and have limited ability to capture contextual and semantic relationships within text, particularly in informal social media environments [8].

Transformer-based models have emerged as a more effective solution by leveraging self-attention mechanisms to capture long-range contextual dependencies. Bidirectional Encoder Representations from Transformers (BERT) has demonstrated strong performance across various text classification tasks, including sentiment and emotion analysis [9]. For the Indonesian language, IndoBERT was developed as a pre-trained transformer model using large-scale Indonesian corpora and has shown superior results compared to conventional machine learning approaches in multiple NLP tasks [10].

Although IndoBERT and other transformer-based models have achieved promising results, most existing studies on emotion classification still adopt a static learning paradigm. In this setting, models are trained once using a fixed dataset and deployed without further updates. Such an approach is less suitable for social media platforms like YouTube, where language usage,

slang expressions, and emotional patterns continuously evolve over time[11], [12]. Models that are not updated with new data risk becoming less relevant and experiencing performance degradation as data characteristics change [13].

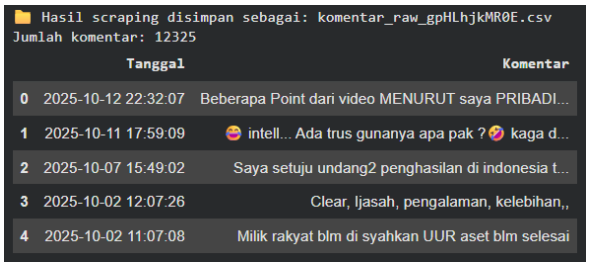
Based on the above discussion, several key problems can be identified. First, the continuously increasing volume of YouTube comments makes manual emotion analysis impractical and inconsistent. Second, the informal and dynamic nature of YouTube comment language poses significant challenges for conventional and static machine learning models. Third, most existing emotion classification studies do not address the need for adaptive learning mechanisms that allow models to evolve as new data become available. These problems indicate the necessity of an emotion classification system that is not only accurate but also adaptive to the dynamic nature of social media data.

To address the identified problems, this study aims to achieve the following objectives. First, to evaluate the performance of the IndoBERT model in multiclass emotion classification of Indonesian YouTube comments using different training–testing data split scenarios. Second, to analyze the impact of class imbalance on classification performance, particularly for minority emotion categories. Third, to develop and implement a web-based emotion classification system that integrates a database-driven incremental learning mechanism, enabling the model to continuously learn from newly validated data without discarding previously acquired knowledge.

## II. METHOD

### A. Data Collection

This study utilized public comments collected from a YouTube video on the Rakyat Bersuara channel. The data were obtained using the YouTube Data API v3, which provides structured access to publicly available comment data on YouTube[14]. Only comments written in Indonesian were retained, while spam content, hyperlinks, and emoji-only comments were removed to ensure data relevance and quality. This filtering process resulted in a clean dataset suitable for emotion classification tasks.


Hasil scraping disimpan sebagai: komentar_raw_gpHLhjkMR0E.csv
Jumlah komentar: 12325

| | Tanggal | Komentar |
|---|---|---|
| 0 | 2025-10-12 22:32:07 | Beberapa Point dari video MENURUT saya PRIBADI... |
| 1 | 2025-10-11 17:59:09 | intell... Ada trus gunanya apa pak ? kaga d... |
| 2 | 2025-10-07 15:49:02 | Saya setuju undang2 penghasilan di indonesia t... |
| 3 | 2025-10-02 12:07:26 | Clear, ljasah, pengalaman, kelebihan,, |
| 4 | 2025-10-02 11:07:08 | Milik rakyat blm di syahkan UUR aset blm selesai |

Fig 1. YouTube Comment Collection Process

### B. Data Labeling

The collected comments were manually annotated using a supervised labeling approach. Each comment was assigned to one of six predefined emotion categories, namely anger, sadness, happiness, fear, surprise, and neutral. The labeling process was conducted by considering the dominant emotional expression conveyed in each comment, following emotion classification schemes commonly adopted in previous studies[15]. The labeled dataset served as ground truth data for model training and evaluation.

TABLE I. EMOTION CATEGORIES

| Text | Emotion |
|---|---|
| Terus gimana? Coba lu jelasin, jgn nyocot doang | Anger |
| Ferry di pojokin terus. Bahkan host juga pojokin ferry. kasihan. | Sadness |
| Diskusi mantulll | Happiness |
| Indonesia hancur karna di adudomba | Fear |
| Daginggg semuaa gelooo | Surprise |
| Banyakk yang dipotong videonyaa | Neutral |

### C. Text Preprocessing

Text Text preprocessing was applied to reduce noise and standardize raw comment data prior to model training. The preprocessing steps included text lowercasing, removal of punctuation, numbers, URLs, and emojis, as well as normalization of informal Indonesian words using a slang dictionary[16], [17]. These steps are necessary due to the informal and unstructured nature of social media text.

Tokenization was performed using the WordPiece tokenizer provided by IndoBERT. This tokenizer represents words at the subword level and effectively handles out-of-vocabulary terms that frequently appear in user-generated content[9].

TABLE II. EXAMPLES OF TEXT PREPROCESSING RESULTS

| Text | Cleaned Text |
|---|---|
| Terus gimana? Coba lu jelasin, jgn nyocot doang | terus gimana coba kamu jelasin jangan bicara doang |
| Ferry di pojokin terus. Bahkan host juga pojokin ferry. kasihan. | ferry di pojokin terus bahkan host juga pojokin ferry kasihan |
| Diskusi mantulll | diskusi mantul |
| Indonesia hancur karna di adudomba | indonesia hancur karena di adu domba |
| Daginggg semuaa gelooo | daging semua gila |
| Banyakk yang dipotong videonyaa | banyak yang dipotong videonya |

### D. IndoBERT Fine-Tuning

The IndoBERT-base model was employed as the core architecture for multiclass emotion classification. IndoBERT is a pre-trained transformer model specifically designed for the Indonesian language and trained on large-scale Indonesian corpora, enabling it to

capture linguistic characteristics unique to Indonesian text [10]. Fine-tuning was performed by adding a fully connected classification layer on top of the [CLS] token representation, followed by a Softmax activation function to predict six emotion classes.

To evaluate model robustness and generalization capability, three training–testing data split scenarios were applied, namely 60:40, 70:30, and 80:20. These split configurations were selected to examine the impact of training data proportion on classification performance while maintaining a consistent evaluation framework. All experiments were conducted using identical preprocessing steps and hyperparameter settings to ensure fair comparison across scenarios.

Model optimization was carried out using the AdamW optimizer with categorical cross-entropy as the loss function, which is widely adopted for multiclass text classification tasks involving transformer-based models [18], [19].The learning rate was set to $2\times10^{-5}$, as this value is commonly recommended for fine-tuning BERT-based architectures to allow gradual parameter updates without disrupting the pre-trained representations. A batch size of 16 was selected to balance gradient stability and computational efficiency, particularly under hardware memory constraints typical in transformer fine-tuning.

The model was trained for three epochs, which was considered sufficient to adapt the pre-trained IndoBERT model to the emotion classification task while minimizing the risk of overfitting. Training for a larger number of epochs may lead to performance degradation on unseen data, especially in social media text classification tasks where class imbalance is present [20]. Overall, this hyperparameter configuration was chosen to provide stable convergence, efficient training, and reliable performance comparison across different data split scenarios.

TABLE III. FINE-TUNING HYPERPARAMETERS

| Hyperparameter | Value |
|---|---|
| Epoch | 3 |
| Batch Size | 16 |
| Learning Rate | 2e-5 |
| Optimizer | AdamW |
| Max Sequence Length | 256 |
| Loss Function | Categorical Cross-Entropy |

### *E. Evaluation Metrics*

Model performance was evaluated using a confusion matrix and standard classification metrics, including accuracy, precision, recall, and F1-score. Since the classification task involved multiple emotion categories with imbalanced class distribution, macro-averaged metrics were used to provide a balanced evaluation across all classes[21]. The best-performing model was selected based on overall accuracy and macro F1-score[22].

### *F. Database-Driven Incremental Retraining Strategy*

To accommodate the dynamic characteristics of social media data, this study implements a database-driven incremental retraining strategy as a key system contribution. A relational database (MySQL) is employed to store YouTube comments together with their predicted emotion labels, confidence scores, validation status, and timestamps. This structured data management enables the continuous accumulation of newly collected and human-validated samples, ensuring that training data can be systematically reused over time [23].

The use of a relational database is motivated by the structured nature of the stored data and the need for data consistency in the learning process. Unlike document-oriented NoSQL databases such as MongoDB, which prioritize schema flexibility, MySQL provides predefined schemas and relational constraints that ensure data integrity and traceability. These characteristics are particularly important in incremental learning scenarios, where training data must be accurately filtered, validated, and retrieved based on specific conditions.

Incremental retraining is performed by reusing the previously fine-tuned IndoBERT model as the initial model state and further fine-tuning it using newly validated data retrieved from the database. This approach allows the model to adapt to evolving linguistic patterns and emerging emotional expressions without discarding previously acquired knowledge. Compared to training from scratch, incremental retraining reduces computational cost and mitigates the risk of catastrophic forgetting, making the proposed system more suitable for long-term deployment in real-world emotion analysis applications [24], [25].

## III. RESULT AND DISCUSSIONS

### *A. Dataset Distribution*

The final dataset consisted of 5,500 Indonesian YouTube comments that were manually labeled into six emotion categories: anger, sadness, happiness, fear, surprise, and neutral. The distribution of emotion classes was imbalanced, with anger and happiness dominating the dataset, while fear and surprise appeared less frequently. Such imbalance is a common characteristic of user-generated social media data and may influence classification performance, particularly for minority emotion classes.

TABLE IV. EMOTION DISTRIBUTION

| Emotion | Count | Percentage |
|---|---|---|
| Anger | 2091 | 38% |
| Happiness | 1927 | 35% |
| Fear | 112 | 2% |
| Surprise | 164 | 3% |
| Sadness | 216 | 4% |
| Neutral | 990 | 18% |

| Total | 5.500 | 100% |
|---|---|---|

*B. Performance Comparison Across Data Split Scenarios*

To evaluate model robustness, the IndoBERT model was trained and tested using three different data split scenarios: 60:40, 70:30, and 80:20. All experiments were conducted using identical preprocessing steps and hyperparameter settings to ensure a fair comparison.

The experimental results indicate a consistent improvement in model performance as the proportion of training data increased. The 80:20 split achieved the highest accuracy, followed by the 70:30 and 60:40 splits. This finding suggests that IndoBERT benefits from larger labeled datasets, enabling the model to better capture contextual information and emotional patterns within YouTube comments.

TABLE V. PERFORMANCE COMPARISON UNDER DIFFERENT DATA SPLIT SCENARIOS

| Data Split | Accuracy |
|---|---|
| 60 : 40 | 63.0% |
| 70 : 30 | 65.7% |
| 80 : 20 | 68.0% |

*C. Classification Report Analysis*

A detailed classification report was generated to analyze model performance across individual emotion categories. The results show that the anger and happiness classes consistently achieved higher precision, recall, and F1-score values compared to other classes. This performance can be attributed to the larger number of training samples and more distinctive linguistic patterns associated with these emotions.

In contrast, the fear and surprise classes exhibited lower F1-scores due to their limited sample size and semantic overlap with other emotion categories. Similar observations have been reported in previous emotion classification studies on Indonesian social media text. Among the evaluated scenarios, the 80:20 split produced the most balanced performance, as reflected by the highest macro-averaged F1-score.

TABLE VI. CLASSIFICATION REPORT FOR THE 80:20 DATA SPLIT

| Emotion | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Anger | 0.72 | 0.76 | 0.74 | 418 |
| Sadness | 0.52 | 0.35 | 0.42 | 43 |
| Happiness | 0.75 | 0.74 | 0.74 | 385 |
| Fear | 0.69 | 0.48 | 0.56 | 23 |
| Surprise | 0.58 | 0.21 | 0.31 | 33 |
| Neutral | 0.51 | 0.58 | 0.54 | 198 |
| ***Accuracy*** | — | — | **0.68** | **1100** |
| *Macro Average* | 0.63 | 0.52 | 0.55 | 1100 |
| *Weighted Average* | 0.68 | 0.68 | 0.67 | 1100 |

*D. Confusion Matrix Analysis*

The confusion matrix for the best-performing 80:20 data split reveals that the IndoBERT model performs well primarily on two dominant emotion classes, namely anger and happiness. Most predictions for these classes are correctly located along the diagonal, indicating that the model is able to capture their distinctive linguistic patterns. In contrast, minority emotion classes such as fear, surprise, and sadness exhibit noticeably higher misclassification rates, suggesting limited predictive capability for these categories.

This imbalance in performance is mainly attributed to the highly skewed distribution of emotion classes in the dataset. While anger and happiness account for a large proportion of the training data, fear and surprise represent only a small fraction of the samples. As a result, the model is exposed to insufficient examples of minority emotions during training, which restricts its ability to learn robust and discriminative representations for these classes. When encountering ambiguous expressions, the model tends to favor dominant classes that it has learned more confidently.

Additionally, misclassifications frequently occur between semantically related emotion pairs. For instance, comments expressing fear are often confused with sadness, while surprise is sometimes misclassified as neutral. This phenomenon reflects the linguistic characteristics of informal YouTube comments, where emotional cues are often subtle, context-dependent, and expressed through shared vocabulary or short phrases. Even transformer-based models such as IndoBERT may struggle to distinguish fine-grained emotional differences under such conditions, particularly when training data are limited.

Although the overall accuracy of 68% indicates reasonable performance, the confusion matrix highlights the limitations of relying solely on aggregate metrics in the presence of class imbalance. The strong performance on dominant classes masks weaker recognition of minority emotions. These findings suggest that future improvements may be achieved through data balancing strategies or by leveraging the proposed database-driven incremental learning mechanism to gradually enrich underrepresented emotion classes over time, thereby improving classification robustness and fairness.

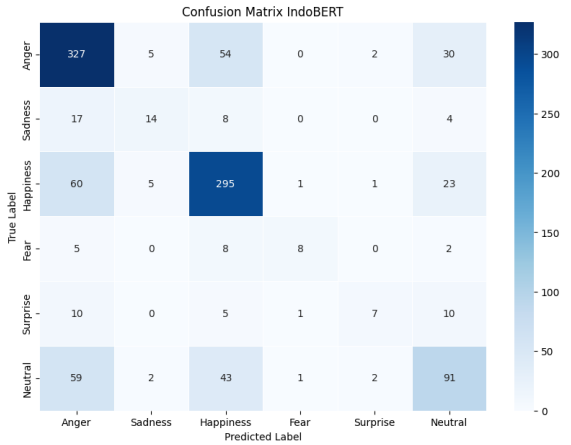



Fig. 2. Confusion Matrix of IndoBERT for the 80:20 Split

### E. *Discussion of Model Performance*

Overall, the experimental results demonstrate that IndoBERT is effective for multiclass emotion classification on Indonesian YouTube comments. The observed performance improvement as the proportion of training data increases, particularly under the 80:20 split scenario, is consistent with previous studies on transformer-based models, which rely on sufficient contextual examples to learn robust text representations.

Nevertheless, the evaluation results also reveal a noticeable performance gap between majority and minority emotion classes. As shown in Table IV, emotion categories such as fear (2%) and surprise (3%) are significantly underrepresented compared to dominant classes such as anger (38%). This extreme class imbalance limits the model's ability to learn discriminative patterns for minority classes, leading to lower recall and F1-score values despite strong overall accuracy.

Although this study does not apply data balancing techniques, the findings indicate that future performance improvements may be achieved by incorporating data augmentation or class rebalancing strategies. Techniques such as Random Oversampling or Synthetic Minority Over-sampling Technique (SMOTE) could be explored to increase the representation of minority emotion classes. In addition, the continuous data accumulation enabled by the proposed database-driven incremental learning system provides a practical pathway to gradually reduce class imbalance as more labeled data become available over time.

### F. *Database-Driven Incremental Retraining Analysis*

Beyond conventional model evaluation, this study introduces a database-driven system architecture that enables continuous data management and incremental learning for emotion classification. The trained IndoBERT model is deployed within a web-based environment that serves as an interface for data collection, human validation, and analysis. A relational database is employed to store YouTube comments along with their predicted emotion labels, confidence scores, and validation status. This structured storage mechanism allows newly collected comments to be accumulated systematically and prepared for subsequent learning processes.

Table structure | Relation view

| # | Name | Type | Collation | Attributes | Null | Default | Comments | Extra |
|---|---|---|---|---|---|---|---|---|
| 1 | id | int | | | No | None | | AUTO_INCREMENT |
| 2 | komentar | text | utf8mb4_0900_ai_ci | | Yes | NULL | | |
| 3 | emosi | varchar(50) | utf8mb4_0900_ai_ci | | Yes | NULL | | |
| 4 | confidence | float | | | Yes | NULL | | |
| 5 | created_at | timestamp | | | Yes | CURRENT_TIMESTAMP | | DEFAULT_GENERATED |
| 6 | label_benar | varchar(50) | utf8mb4_0900_ai_ci | | Yes | NULL | | |

Fig. 3. Database Structure

The database plays a central role in supporting system adaptability and long-term learning. By maintaining both historical data and newly validated samples, the system ensures data consistency, traceability, and controlled data growth throughout the learning lifecycle. Unlike static emotion classification approaches that rely on a fixed dataset, the proposed system is designed to evolve over time, reflecting changes in language usage, emerging slang, and shifting emotional expressions commonly observed in YouTube comments.

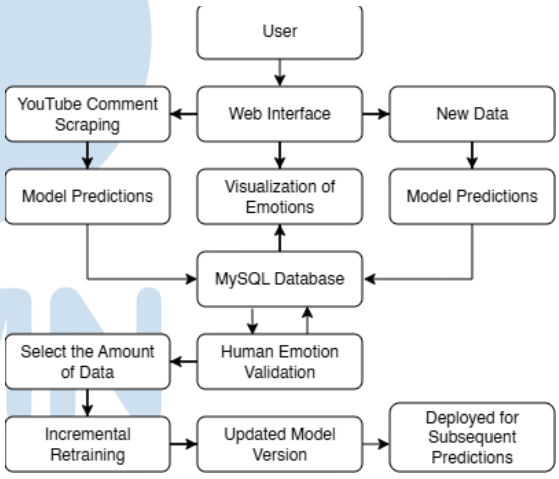



Fig. 4. Integrated System Architecture and Database-Driven Incremental Retraining Workflow

Incremental retraining is performed by reusing the previously fine-tuned IndoBERT model as the initial model state. Instead of reinitializing the model from scratch, the retraining process refines existing model parameters using newly validated data retrieved from the database. This approach enables the model to adapt to new emotional expressions while preserving previously acquired contextual knowledge.

A critical challenge in incremental learning is catastrophic forgetting, where a model loses previously learned knowledge when trained on new data. In the proposed system, this issue is mitigated by retaining the previously learned model parameters and performing controlled fine-tuning using incremental data batches. Rather than replacing existing knowledge, the retraining process refines and extends the learned

representations. As a result, the incremental retraining mechanism functions as a knowledge accumulation process rather than a knowledge replacement process, enabling stable long-term learning.

In the proposed system, the retraining process is triggered after a predefined number of newly validated comments have been accumulated in the database. Specifically, retraining is performed when at least 100 new validated samples are available, ensuring that sufficient new information is provided to meaningfully update the model. The retraining procedure can be executed manually by an administrator to maintain data quality control. This design choice balances learning efficiency and computational cost, preventing excessive retraining while enabling periodic model updates as new data are collected.

Furthermore, the incremental retraining mechanism offers significant computational advantages compared to full retraining. As the volume of stored data increases—particularly for underrepresented emotion classes—the model can progressively improve its classification performance. This adaptive learning capability distinguishes the proposed system from prior emotion classification studies that employ static models and demonstrates its suitability for scalable, real-world emotion analysis on dynamic social media platforms such as YouTube.

## IV. CONCLUSIONS

This study presents an adaptive multiclass emotion classification system for Indonesian YouTube comments using the IndoBERT model integrated with a database-driven incremental learning framework. Experimental results demonstrate that the best performance is achieved under the 80:20 training–testing split, with an overall accuracy of 68%, confirming that a larger proportion of labeled training data improves contextual understanding in informal social media text.

Beyond classification accuracy, the proposed system introduces a human-in-the-loop validation mechanism that selectively targets only incorrect or uncertain model predictions. Rather than validating all predictions, human annotators correct misclassified emotion labels, and these validated samples are stored back into the database. This selective validation strategy ensures that subsequent learning focuses on informative errors, improving data quality and training efficiency.

Incremental retraining is then performed by reusing the previously fine-tuned IndoBERT model as the initial state and updating it with the validated samples. This targeted retraining process allows the model to refine its decision boundaries and progressively align its predictions with human judgment, particularly for ambiguous emotional expressions. As a result, the updated model becomes more accurate and human-aligned without discarding previously acquired knowledge or incurring the high computational cost of full retraining.

Overall, the experimental findings validate that combining transformer-based language models with selective human validation and database-supported incremental learning provides an effective and scalable solution for emotion classification on YouTube comments. The proposed framework supports continuous performance improvement, mitigates catastrophic forgetting, and is suitable for long-term deployment in dynamic social media environments.

## REFERENCES


[1] H. Ahuja, N. Kaur, P. Kumar, and A. Hafiz, "Machine Learning based Sentiment Analysis of YouTube Video Comments," *2023 1st International Conference on Advances in Electrical, Electronics and Computational Intelligence, ICAEECI 2023*, 2023, doi: 10.1109/ICAEECI58247.2023.10370907.

[2] S. Yang, D. Brossard, D. A. Scheufele, and M. A. Xenos, "The science of YouTube: What factors influence user engagement with online science videos?," *PLoS One*, vol. 17, no. 5 May, May 2022, doi: 10.1371/JOURNAL.PONE.0267697.

[3] F. A. Acheampong, C. Wenyu, and H. Nunoo-Mensah, "Text-based emotion detection: Advances, challenges, and opportunities," *Engineering Reports*, vol. 2, no. 7, p. e12189, Jul. 2020, doi: 10.1002/ENG2.12189;PAGEGROUP:STRING:PUBLICATION.

[4] T. Padma, R. Visweshvar, K. Tamilarasan, and C. J. Bhadrinath, "Dynamic YouTube Comment Sentiment Analysis with Supervised Fine-Tuned BERT," *2024 International Conference on Cognitive Robotics and Intelligent Systems, ICC - ROBINS 2024*, pp. 663–669, 2024, doi: 10.1109/ICC-ROBINS60238.2024.10533926.

[5] U. Krishna, "YouTube Comments Sentiments Analysis," *Int J Res Appl Sci Eng Technol*, vol. 13, no. 1, pp. 875–880, Jan. 2025, doi: 10.22214/IJRASET.2025.66475.

[6] T. Lonkar, T. Katkar, M. Karajgar, G. Lonkar, and S. Shelar, "YouTube Comments Analyzer Using Natural Language Processing And Artificial Intelligence," *International Journal of Computer Sciences and Engineering*, vol. 12, no. 12, pp. 1–14, Dec. 2024, doi: 10.26438/IJCSE/V12I12.114.

[7] M. Z. Asghar, S. Ahmad, A. Marwat, and F. M. Kundi, "Sentiment Analysis on YouTube: A Brief Survey," Nov. 2015, Accessed: Jan. 04, 2026. [Online]. Available: https://arxiv.org/pdf/1511.09142

[8] O. El Azzouzy, T. Chanyour, and S. J. Andaloussi, "Transformer-based models for sentiment analysis of YouTube video comments," *Sci Afr*, vol. 29, p. e02836, Sep. 2025, doi: 10.1016/J.SCIAF.2025.E02836.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of the 2019 Conference of the North*, pp. 4171–4186, Jun. 2019, doi: 10.18653/V1/N19-1423.

[10] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," Association for Computational Linguistics (ACL), Nov. 2020, pp. 843–857. doi: 10.18653/V1/2020.AACL-MAIN.85.

[11] M. Yeza Baihaqi, E. Halawa, R. Asyahira, S. Syah, A. Nurrahma, and W. Wijaya, "Emotion Classification in Indonesian Language: A CNN Approach with Hyperband Tuning," *Jurnal Buana Informatika*,

vol. 14, no. 02, pp. 137–146, Oct. 2023, doi: 10.24002/JBI.V14I02.7558.

[12] N. Hilmiaji, K. M. Lhaksmana, and M. D. Purbolaksono, "Identifying Emotion on Indonesian Tweets using Convolutional Neural Networks," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 3, pp. 584–593, Jun. 2021, doi: 10.29207/RESTI.V5I3.3137.

[13] M. Masana, X. Liu, B. Twardowski, M. Menta, A. D. Bagdanov, and J. Van De Weijer, "Class-Incremental Learning: Survey and Performance Evaluation on Image Classification," *IEEE Trans Pattern Anal Mach Intell*, vol. 45, no. 5, pp. 5513–5533, May 2023, doi: 10.1109/TPAMI.2022.3213473;PAGE:STRING:ARTICLE/CHAPTER.

[14] S. (santi) thomas, Y. (Yuliana) Yuliana, and N. (Noviyanti) P, "Study Analisis Metode Analisis Sentimen pada YouTube," *Journal of Information Technology*, vol. 1, no. 1, pp. 1–7, Mar. 2021, doi: 10.46229/JIFOTECH.V1I1.201.

[15] M. H. Algifari and E. D. Nugroho, "Emotion Classification of Indonesian Tweets using BERT Embedding," *Journal of Applied Informatics and Computing*, vol. 7, no. 2, pp. 172–176, Nov. 2023, doi: 10.30871/JAIC.V7I2.6528.

[16] S. Jadhav and V. Milosavljevic, "Sentiment Analysis of User comments for a YouTube Educational videos MSc Research Project Masters of Science in Data Analytics (MSCDAD_A_JAN24I)".

[17] S. Shetty, S. Shetty, P. Kamath B, and D. Shetty, "SENTIMENT PATTERNS IN YOUTUBE COMMENTS: A COMPREHENSIVE ANALYSIS," *Computer Science & Engineering: An International Journal (CSEIJ)*, vol. 15, no. 1, 2025, doi: 10.5121/cseij.2025.15119.

[18] Y. Shao *et al.*, "An Improved BGE-Adam Optimization Algorithm Based on Entropy Weighting and Adaptive Gradient Strategy," *Symmetry (Basel)*, vol. 16, no. 5, p. 623, May 2024, doi: 10.3390/sym16050623.

[19] A. Mao, M. Mohri, and Y. Zhong, "Cross-Entropy Loss Functions: Theoretical Analysis and Applications," 2023.

[20] Y. O. Sihombing, R. Fuad Rachmadi, S. Sumpeno, and M. J. Mubarok, "Optimizing IndoRoBERTa Model for Multi-Class Classification of Sentiment & Emotion on Indonesian Twitter," *Proceeding - IEEE 10th Information Technology International Seminar, ITIS 2024*, pp. 12–17, 2024, doi: 10.1109/ITIS64716.2024.10845566.

[21] I. Markoulidakis and G. Markoulidakis, "Probabilistic Confusion Matrix: A Novel Method for Machine Learning Algorithm Generalized Performance Analysis," *Technologies (Basel)*, vol. 12, no. 7, p. 113, Jul. 2024, doi: 10.3390/technologies12070113.

[22] J. H. Cabot and E. G. Ross, "Evaluating prediction model performance," *Surgery (United States)*, vol. 174, no. 3, pp. 723–726, Sep. 2023, doi: 10.1016/j.surg.2023.05.023.

[23] H. A. Mumtahana, "Optimization of Transaction Database Design with MySQL and MongoDB," *SinkrOn*, vol. 7, no. 3, pp. 883–890, Jul. 2022, doi: 10.33395/sinkron.v7i3.11528.

[24] R. Wang, J. Fei, R. Zhang, M. Guo, Z. Qi, and X. Li, "DRnet: Dynamic Retraining for Malicious Traffic Small-Sample Incremental Learning," *Electronics 2023, Vol. 12, Page 2668*, vol. 12, no. 12, p. 2668, Jun. 2023, doi: 10.3390/ELECTRONICS12122668.

[25] M. Masana, X. Liu, B. Twardowski, M. Menta, A. D. Bagdanov, and J. Van De Weijer, "Class-incremental learning: survey and performance evaluation on image classification," *IEEE Trans Pattern Anal Mach Intell*, vol. 45, no. 5, pp. 5513–5533, Oct. 2020, doi: 10.1109/TPAMI.2022.3213473

# 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.

- The subscript for the permeability of vacuum $\mu_0$, and other common scientific constants, is zero with subscript formatting, not a lowercase letter “o.”
- In American English, commas, semi-/colons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
- A graph within a graph is an “inset,” not an “insert.” The word alternatively is preferred to the word “alternately” (unless you really mean something that alternates).
- Do not use the word “essentially” to mean “approximately” or “effectively.”
- In your paper title, if the words “that uses” can accurately replace the word using, capitalize the “u”; if not, keep using lower-cased.
- Be aware of the different meanings of the homophones “affect” and “effect,” “complement” and “compliment,” “discreet” and “discrete,” “principal” and “principle.”
- Do not confuse “imply” and “infer.”
- The prefix “non” is not a word; it should be joined to the word it modifies, usually without a hyphen.
- There is no period after the “et” in the Latin abbreviation “et al.”
- The abbreviation “i.e.” means “that is,” and the abbreviation “e.g.” means “for example.”

## IV. Using the Template

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention as below

ULTIMATICS_firstAuthorName_paperTitle.

In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper. Please take note on the following items.

### A. Authors and Affiliations

The template is designed so that author affiliations are not repeated each time for multiple authors of the same affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

### 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 | | |

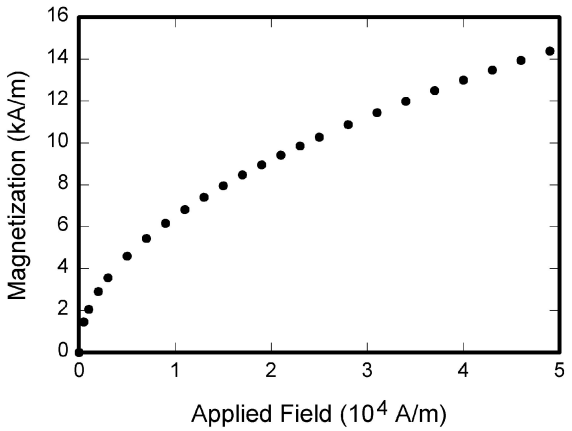



Fig. 1. Example of a figure caption

## 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