E-ISSN: XXXX-XXXX

Volume 1 Issue 1 2026

Journal of Data Science Informatics and Computer Engineering





https://journals.pens.ac.id/index.php/jdsice

Politeknik Elektronika Negeri Surabaya

JOURNAL OF DATA SCIENCE INFORMATICS AND COMPUTER ENGINEERING

ONLINE ISSN: XXXX-YYYY
Volume 1, Issue 1, September 2026, Pages 1 - 40

EDITOR'S GREETING

All praise to Allah for the successful release of the Journal of Data Science Informatics and Computer Engineering (JDSICE), Volume 1, Issue 1, March–September 2026. The dedication of the editorial team, reviewers, and contributors has made it possible for this inaugural issue to be published on time. We sincerely hope that this edition will serve as a valuable reference and contribute to the advancement of data science, informatics, and computer engineering.

This first issue presents 5 selected articles, written by lecturers and students from Indonesian universities. Their contributions reflect the growing enthusiasm and commitment of the academic community in Indonesia to produce impactful research and foster collaboration between educators and learners.

Throughout the submission period, from March 2026 until September 2026, the journal **received 6 manuscripts**. After a thorough peer-review process, **5 manuscripts were accepted** while **1 was declined**, demonstrating our effort to uphold rigorous academic standards and ensure the quality of published works.

As a new journal, JDSICE is committed to becoming a consistent platform for disseminating scholarly work and encouraging innovation in the fields of data science, informatics, and computer engineering. We extend our deepest gratitude to all authors, reviewers, and the editorial team for their dedication. Special appreciation is also given to the researchers, lecturers, and students who entrusted their work to JDSICE.

We hope the articles published in this edition will enrich scientific knowledge, inspire further research, and provide practical insights for academic and professional communities. To continuously improve the quality of this journal, we warmly welcome constructive feedback and suggestions from readers and contributors.

Surabaya, September 28, 2026
Chair of the Editorial Board,

Prof. Dr. Setiawardhana, S.T., M.T.

JOURNAL OF DATA SCIENCE INFORMATICS AND COMPUTER ENGINEERING

ONLINE ISSN: XXXX-YYYY
Volume 1, Issue 1, September 2026, Pages 1 - 40

Editorial Team and Reviewer

Editor in Chief
Setiawardhana (Scopus ID: 55903265800) Politeknik Elektronika Negeri Surabaya, Indonesia

Managing Editor
Adnan Rachmat Anom Besari (Scoupus ID: 57201671258) Politeknik Elektronika Negeri Surabaya, Indonesia

Editorial Board
Khoironi (SINTA ID : 6794894) Politeknik Elektronika Negeri Surabaya, Indonesia
Prasetyo Wibowo (SINTA ID : 6789820) Politeknik Elektronika Negeri Surabaya, Indonesia
Nirwana Haidar Hari (Scopus ID: 57208469645) Politeknik Elektronika Negeri Surabaya, Indonesia
Renovita Edelani (SINTA ID : 6790001) Politeknik Elektronika Negeri Surabaya, Indonesia

Peer Reviewer
Riyanto Sigit (Scopus ID: 35811196100) Politeknik Elektronika Negeri Surabaya, Indonesia
Arna Fariza (Scopus ID:) Politeknik Elektronika Negeri Surabaya, Indonesia
M. Udin Harun Al Rasyid (Scopus ID: 55569020700) Politeknik Elektronika Negeri Surabaya, Indonesia
Selvia Ferdiana Kusuma (Scopus ID: 57189257534) Politeknik Elektronika Negeri Surabaya, Indonesia
Bima Sena Bayu Dewantara (Scopus ID: 56039019600) Politeknik Elektronika Negeri Surabaya, Indonesia

JOURNAL OF DATA SCIENCE INFORMATICS AND COMPUTER ENGINEERING

ONLINE ISSN: XXXX-YYYY
Volume 1, Issue 1, September 2026, Pages 1 - 40

TABLE OF CONTENTS

No	Title	Pages
1	**Design and Implementation of a Microservices-Based Primary Healthcare Information System with Mobile Telemedicine Support** *Author(s): Andhik Ampuh Yunanto, Mu'arifin, Adhitia Nur Riski*	1-10
2	**A Mobile-Based BISINDO Recognition and Communication Support System Using MediaPipe, MLP, and GRU** *Author(s): Adrian Yoga Chrisarianto, Entin Martiana Kusumaningtyas, Aliridho Barakbah, Ira Prasetyaningrum*	11-17
3	**Sentiment Classification and Category Mapping of Public Perception on Tax Service Quality at KPP Pratama Surabaya Gubeng Using Naive Bayes and Support Vector Machine** *Author(s): Puput Ayu Setiawati, Fitrah Maharani Humaira, Diah Kusumadewi Sugiarto*	18-24
4	**A Spatial-Temporal Framework for Hepatitis-A Vulnerability Analysis Using Hierarchical Clustering: Case Study of Pacitan Regency, Indonesia** *Author(s): Arna Fariza, Fanny Maulana Rizky, Arif Basofi*	25-33
5	**Explainable AI for Multi-Label Classification of Doctor Patient Communication Using Transformer-Based Models** *Author(s): Selvia Ferdiana Kusuma, Mohammad Farid Naufal, Rizal Furqan Ramadhan*	34-40

Publisher : Politeknik Elektronika Negeri Surabaya

Address : Jl. Raya ITS, Keputih, Kecamatan Sukolilo, Kota Surabaya, Jawa Timur 60111

Website : https://journals.pens.ac.id/jdsice/index

E-mail : jdsice@journals.pens.ac.id

Journal of Data Science Informatics and Computer Engineering (JDSICE) published since September 2026 by Pusat Penelitian dan Pengabdian Masyarakat
Politeknik Elektronika Negeri Surabaya

Journal of Data Science, Informatics and Computer Engineering (JDSICE),
Vol. 1, Issue 1
February 2026
ISSN: XXXX-YYYY

Design and Implementation of a Microservices-Based Primary Healthcare Information System with Mobile Telemedicine Support

Andhik Ampuh Yunanto*, Mu'arifin, Adhitia Nur Riski
Informatics and Computer Engineering Department, Politeknik Elektronika Negeri Surabaya, Indonesia
**Corresponding author: Andhik Ampuh Yunanto; email: andhik@pens.ac.id*

Manuscript received July 25, 2026, revised September 10, 2026, accepted September 21, 2026
doi: zz.yyyyy/xxxxx.pp134-1399

Abstract – *Primary healthcare centers require integrated information systems to reduce delays and errors in patient data processing, medical record management, and service reporting. This study presents the design and implementation of a microservices-based primary healthcare information system with mobile telemedicine support for Wonorejo Trisulo Community Health Center. The system was developed using JavaScript and Node.js, with MongoDB for data storage and REST APIs connected through an API Gateway. The web application supports patient registration, emergency and outpatient services, electronic medical records, cashier operations, pharmacy, inventory, laboratory, radiology, and reporting, while the mobile interface enables patients to submit and review telemedicine requests. The system was evaluated through 30 scenario-based black-box functional tests and an initial questionnaire-based user perception assessment involving six healthcare personnel. All functional test scenarios produced the expected outputs. The user assessment showed preliminary positive perceptions of application performance (80%), ease of use (70%), efficiency (63%), interface attractiveness (80%), usefulness (93%), and overall satisfaction (83%). These user-perception results are descriptive and should not be generalized as evidence of overall usability, satisfaction, or system adoption.*

CC BY SA

Keywords: *Electronic medical records, Health information system, Microservices, Primary healthcare, Telemedicine*

I. Introduction

Digital transformation has become an important strategy for improving the accessibility, continuity, and efficiency of primary healthcare services. The World Health Organization emphasizes that transforming paper-based services into person-centered digital systems requires systematic requirements analysis, workflow mapping, structured health data, interoperability, and the involvement of healthcare workers as system users [1]. In Indonesia, the digitalization of healthcare records is also supported by the Minister of Health Regulation Number 24 of 2022 concerning medical records, which provides a regulatory foundation for managing medical records electronically [2]. These developments highlight the need for healthcare facilities, including community health centers, to adopt integrated information systems that support both clinical and administrative activities.

Electronic medical records (EMRs) can improve the availability and timeliness of patient information, facilitate coordination among healthcare personnel, and support more efficient healthcare processes [3], [4]. Nevertheless, successful implementation is not determined solely by the availability of software. Infrastructure limitations, inadequate training, workflow incompatibility, insufficient technical support, and user readiness may influence the adoption and effective use of digital health systems [3]. Therefore, healthcare information systems should be designed according to actual operational workflows and the needs of their intended users.

Wonorejo Trisulo Community Health Center provides primary healthcare services to the surrounding community in Plosoklaten District, Kediri Regency, Indonesia. Before the development of the proposed system, several activities were still performed manually, including patient registration, medical record documentation, medicine inventory management, and service reporting. This condition created difficulties in retrieving historical patient information, maintaining consistent records, monitoring medicine availability, and preparing reports.

Furthermore, different service units required access to related patient and service information, while manually recorded documents were maintained separately. These operational conditions motivated the development of an integrated system connecting registration, clinical examination, emergency and outpatient care, pharmacy, laboratory, radiology, cashier, inventory, medical records, and reporting processes.

An integrated healthcare application consists of multiple business functions with different responsibilities and data requirements. Microservices architecture provides an approach for separating application functions into independently managed services that communicate through defined interfaces [5]. In healthcare environments, microservices have also been proposed to support modularity, service separation, and information exchange among independently managed healthcare components [6]. Although microservices can provide architectural flexibility, their implementation also introduces additional concerns related to service decomposition, communication, monitoring, security, and testing. Therefore, the architecture should be adapted to the operational requirements and scale of the target healthcare environment.

Digital healthcare systems can also extend patient access through mobile health and telemedicine services. Previous studies have reported generally positive experiences with telemedicine in primary care while identifying usability, technological skills, communication, infrastructure, and user preferences as important implementation factors [7]. Telemedicine implementation in resource-constrained and safety-net healthcare settings has similarly demonstrated potential benefits in accessibility while revealing challenges related to infrastructure, workflows, and patient preferences [8]. These findings indicate that telemedicine can complement conventional primary healthcare services when it is appropriately integrated with existing clinical and administrative processes.

Several recent studies have demonstrated different approaches to healthcare information system development. Arcila-Diaz and Valdivia [9] proposed a microservices-based architecture for dental health records using independent databases and an API Gateway. Lagakis et al. [10] developed ClinApp, a microservices-based platform supporting medical appointment scheduling and patient data management. Burka et al. [11] presented a user-centered digital health information system combining a mobile application with a web-based dashboard to support chronic disease management in primary care. In addition to system implementation, usability evaluation has been recognized as an important component of health information system assessment, as demonstrated by Østervang et al. [12]. These studies show that microservices, mobile applications, web dashboards, and usability assessment have increasingly been adopted in digital healthcare systems, although their functional scope and implementation contexts differ.

In the Indonesian context, studies have also investigated the implementation and utilization of digital health information systems. Muchlis [13] examined factors influencing successful electronic medical record implementation in community health centers in Jakarta. Aisyah et al. [14] investigated health information system utilization across public health centers in Indonesia, highlighting variations in implementation and operational conditions. Fu'ad et al. [15] reported the implementation of a cloud-based electronic medical record system at a community health center. Martono et al. [16] further addressed healthcare information interoperability through the development of a Satu Sehat Interoperability Agent, demonstrating the importance of integrating healthcare information systems with Indonesia's national health data ecosystem. These studies provide important evidence regarding EMR adoption, health information system utilization, interoperability, and digital patient-data management in Indonesian healthcare settings. More broadly, previous application-development research has demonstrated that user-centered design and interaction simplification can contribute to improved usability and user satisfaction in interactive applications [17].

Based on the reviewed literature, electronic medical records, microservices-based healthcare systems, primary-care information systems, and telemedicine have been investigated from different perspectives. However, among the selected related studies, implementations that combine administrative and clinical workflows, electronic medical records, pharmaceutical and diagnostic services, a microservices-based backend with an API Gateway, web-based access for healthcare personnel, and mobile telemedicine support within a single community health center remain limited. Accordingly, the contribution of this study is not the introduction of a new microservices protocol or software technology, but rather the practical integration of established technologies into a unified architecture tailored to the operational requirements of a community health center. A comparative analysis between the proposed system and selected related healthcare systems is presented in Section IV.

This study presents the design and implementation of a microservices-based primary healthcare information system with mobile telemedicine support at Wonorejo Trisulo Community Health Center. The system was implemented using JavaScript and Node.js, MongoDB for data storage, and REST application programming interfaces connected through an API Gateway. The web application integrates patient registration, electronic medical records, emergency and outpatient services, pharmacy and inventory management, cashier operations, laboratory examinations, radiology services, and reporting, while the mobile interface supports patient telemedicine requests. The main contributions of this study are threefold: first, it provides a practical integration of administrative, clinical, electronic medical record,

pharmaceutical, diagnostic, reporting, and telemedicine workflows within a community health center environment; second, it implements a microservices-based backend with an API Gateway to support both healthcare-personnel web access and patient-oriented mobile telemedicine services; and third, it provides implementation evidence through scenario-based functional testing and an initial user perception assessment. Therefore, the novelty of this work is primarily integrative and application-oriented rather than based on a new software architecture mechanism.

The remainder of this paper is organized as follows. Section II describes the research and system development methodology, including requirements analysis, system development, functional testing, and initial user perception evaluation. Section III presents the system architecture, service communication, database design, and implementation. Section IV reports the functional and user-evaluation results and compares the proposed system with related healthcare information systems. Finally, Section V concludes the study and outlines potential directions for future development.

II. Research Methodology

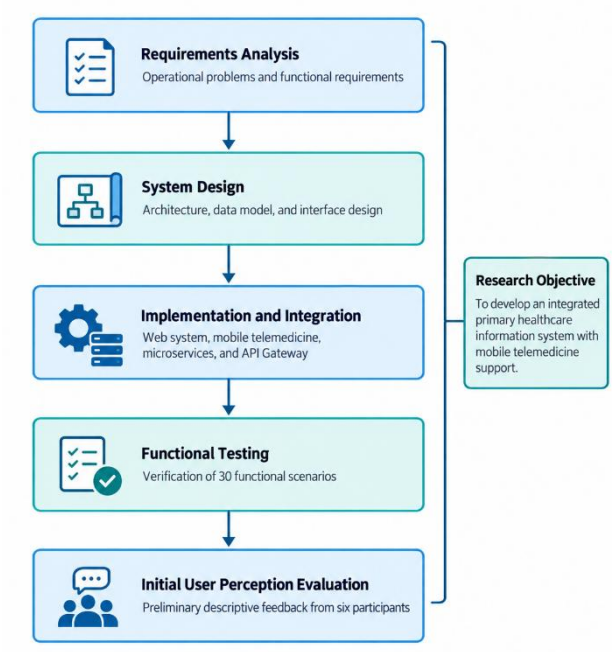


Fig. 1. Research and system development workflow

This study employed a system development and evaluation approach consisting of five principal stages: requirements analysis, system design, implementation and integration, functional testing, and initial user perception evaluation. Each stage produced an outcome directly related to the overall objective of developing an integrated primary healthcare information system with mobile telemedicine support. Requirements analysis identified operational problems and functional requirements; system design translated these requirements into application, data, and interface models; implementation and integration realized the web, mobile, microservices, and API components; functional testing verified the implemented functions against predefined scenarios; and the initial user perception evaluation provided preliminary feedback from personnel in the target deployment environment. The overall research and system development workflow is illustrated in Fig. 1.

A. Requirements Analysis

The requirements analysis was conducted at Wonorejo Trisulo Community Health Center, Plosoklaten District, Kediri Regency, Indonesia. Information regarding the existing healthcare service processes was obtained through observation of operational activities and communication with personnel involved in healthcare-center services. The analysis focused on identifying problems associated with manual data processing, including difficulties in retrieving patient records, documenting medical examinations, monitoring medicine inventory, and preparing healthcare service reports.

The results of this stage were translated into the functional requirements of the proposed system. The identified business processes included user management, patient registration, emergency services, outpatient services, electronic medical records, cashier operations, pharmaceutical inventory, pharmacy services, laboratory examinations, radiology services, report generation, and mobile telemedicine support.

The system was designed to accommodate different categories of users with different access requirements, including administrators, registration personnel, doctors, nurses, pharmacists, laboratory personnel, radiology personnel, cashier personnel, and patients. Healthcare personnel access functions according to their assigned responsibilities, while registered patients access the telemedicine component through the mobile-oriented interface. Thus, the main outcome of the requirements-analysis stage was a set of functional and user-access requirements used as the basis for subsequent system design and implementation.

B. System Development Procedure

The system design stage translated the identified requirements into application architecture, data structures, process models, and user interfaces. The design artifacts included use case diagrams, data flow diagrams, an entity relationship diagram, a conceptual data model, and a physical data model. User interfaces were also designed for the web-based healthcare information system and the mobile telemedicine component.

The implementation stage realized the designed system using JavaScript and Node.js for server-side application development and MongoDB for data storage. Express.js was used to handle HTTP requests, while Mongoose was used to manage access between the Node.js services and MongoDB. The backend components communicated through Representational State Transfer Application Programming Interfaces

(REST APIs), and an API Gateway was implemented as the intermediary between client applications and backend services.

The backend was separated into service components according to their responsibilities. The primary healthcare information service supports administrative, clinical, pharmaceutical, diagnostic, and reporting functions, while the telemedicine service supports patient-oriented remote service functions. Each service operates through an independent service port and manages its corresponding data storage. The web and mobile interfaces were subsequently integrated with these backend services through the API Gateway.

User authentication was implemented to control access to the system, and passwords were processed using bcrypt before storage. The outcome of this stage was an integrated web and mobile healthcare information system capable of supporting the operational functions identified during requirements analysis. The detailed architecture and implementation are described in Section III.

C. Functional Testing

Functional testing was conducted using a scenario-based black-box testing approach to verify whether the implemented functions produced the expected outputs for specified user actions and inputs. The testing focused on externally observable system behavior without examining the internal program structure.

The test scenarios covered the principal functions of the system, including authentication, user management, patient registration, medical record management, emergency and outpatient services, pharmacy and inventory management, cashier operations, laboratory and radiology services, and telemedicine functions.

Testing was conducted at Wonorejo Trisulo Community Health Center from May 31 to June 11, 2022. The application was operated on a Windows-based computer equipped with an AMD Ryzen 3 5300U processor and 8 GB of memory. A test scenario was classified as passed when the observed system behavior was consistent with the expected result. In total, 30 functional scenarios were evaluated. The outcome of this stage was used to verify the functional completeness of the implemented system before interpreting the initial user feedback.

D. Initial User Perception Evaluation

An initial user perception evaluation was conducted to obtain preliminary feedback regarding the implemented application. The evaluation involved six personnel from Wonorejo Trisulo Community Health Center who participated in the system trial and were involved in healthcare-center operational activities. The available evaluation records did not document the specific professional role of each individual respondent; therefore, balanced representation across all intended system-user roles could not be established.

The six participants were actual personnel from the target deployment environment and were familiar with healthcare service processes supported by the application. Nevertheless, considering the small sample size and the absence of complete role-specific respondent information, the evaluation was intended only to provide descriptive feedback on initial user perceptions. It was not designed to establish statistically generalizable conclusions regarding usability, satisfaction, or system adoption.

The questionnaire evaluated six aspects: application performance and stability, ease of use, efficiency and convenience, interface attractiveness, usefulness for users, and overall user satisfaction. Each statement was rated using a five-point Likert scale, with Strongly Agree = 5, Agree = 4, Neutral = 3, Disagree = 2, and Strongly Disagree = 1.

The percentage score for each evaluation aspect was calculated using:

$$P = \frac{\sum S_i}{N \times S_{\max}} \times 100\% \qquad (1)$$

where P represents the percentage score, $\sum S_i$ is the total score obtained from the respondents, N is the number of respondents, and $S_{\max}$ is the maximum Likert score of 5.

The resulting percentages were interpreted using five categories: 0–20% as Strongly Disagree, 21–40% as Disagree, 41–60% as Neutral, 61–80% as Agree, and 81–100% as Strongly Agree. Due to the small number of respondents, the results were analyzed descriptively and were not used to make statistically generalized claims.

III. System Design and Implementation

A. Overall System Architecture

The proposed system was designed to connect healthcare personnel and patients through web-based and mobile-oriented interfaces. The web application serves as the primary interface for healthcare personnel, while the mobile-oriented interface provides telemedicine functions for registered patients. Both interfaces communicate with backend services through Representational State Transfer Application Programming Interfaces (REST APIs).

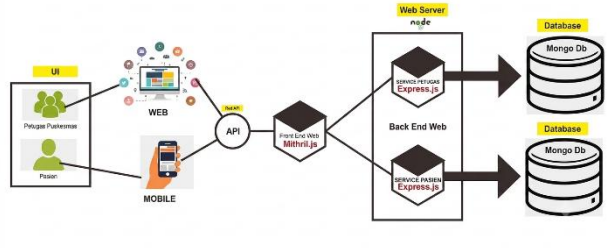


Fig. 2. Overall architecture of the web and mobile primary healthcare information system

Fig. 2 illustrates the overall architecture of the proposed system. Healthcare personnel access the web application to manage administrative, clinical, pharmaceutical, diagnostic, and reporting activities. Patients access the mobile telemedicine interface to submit remote healthcare requests. Requests from both interfaces are directed to the

backend through an API layer that connects the client applications to the corresponding services and data storage components.

B. Microservices and API Gateway

The backend was implemented using a microservices-based approach in which the primary healthcare information service and the telemedicine service operate as separate backend components. Each service is deployed through an independent service port and manages its corresponding MongoDB data storage. The separation of service processes and databases was intended to reduce direct dependency between application components and to allow each service to be accessed independently.

Fig. 3 provides a more detailed representation of the communication between client applications, the API Gateway, backend services, and databases. Both the web and mobile clients communicate with the backend through the API Gateway rather than directly accessing individual services or databases.

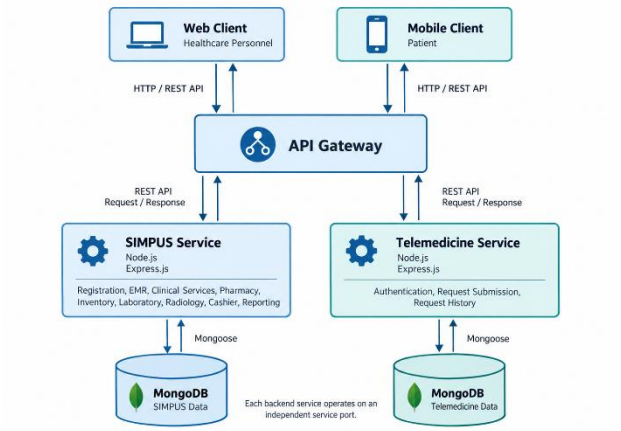


Fig. 3. Detailed service communication through the API Gateway in the proposed microservices-based architecture.

An incoming HTTP request from a client is first received by the API Gateway. The gateway identifies the requested function and forwards the request to the appropriate backend service through its designated service port. The selected service processes the request, accesses the corresponding MongoDB collections when data operations are required, and returns the response to the client through the API Gateway.

The implemented architecture contains two principal backend service groups. The SIMPUS service handles internal primary healthcare functions, including patient registration, electronic medical records, clinical services, pharmacy, inventory, laboratory, radiology, cashier operations, and reporting. In contrast, the telemedicine service handles patient-oriented remote service functions, including patient authentication, telemedicine request submission, and access to request history.

Communication between the API Gateway and backend services is performed using HTTP-based REST APIs. Standard create, retrieve, update, and delete operations are handled through application endpoints. Node.js provides the server-side JavaScript runtime, while Express.js handles HTTP routing and request processing. Database access from the Node.js services is managed using Mongoose.

The overall communication sequence can therefore be summarized as client request → API Gateway → selected service → database operation → service response → API Gateway → client. This communication pattern enables different client platforms to use the same backend environment while keeping database access encapsulated within the corresponding backend service. No direct access from the web or mobile client to MongoDB is required.

User authentication was also implemented to provide controlled access to protected system functions. Passwords are processed using bcrypt before being stored in the database. However, the present study did not include dedicated penetration testing or comprehensive security-performance evaluation. Therefore, authentication, password hashing, and service separation are reported as implemented security and access-control mechanisms rather than evidence of complete system security.

C. Web-Based Healthcare Services

The web application was developed for healthcare personnel whose responsibilities are distributed across several operational units. After authentication, each user is directed to the system dashboard and provided with functions corresponding to the assigned role. The implemented modules are summarized in Table I.

TABLE I
MAIN MODULES OF THE WEB-BASED HEALTHCARE INFORMATION SYSTEM

Module	Main functions
User management	Manages user accounts, access roles, reference data, and database backup or restoration
Patient registration	Registers new patients, searches existing patients, and records outpatient or emergency visits
Electronic medical records	Stores patient identity, visit history, examination results, diagnoses, treatments, and prescriptions
Emergency service	Records nursing and medical examinations for emergency patients
Outpatient service	Manages patient queues and examination records for outpatient clinics
Cashier	Calculates and confirms payments associated with healthcare services
Pharmaceutical inventory	Manages medicines, medical supplies, batches, stock, and distribution requests
Pharmacy	Processes prescriptions, medicine delivery, and non-prescription medicine transactions
Laboratory	Records patients, laboratory procedures, and examination results
Radiology	Manages radiology requests and examination information
Reporting	Generates healthcare and supporting-service reports

The user-management module allows administrators to

create accounts, assign access roles, maintain reference data, and perform database backup and restoration. Role separation restricts users to functions associated with their operational responsibilities. For example, registration personnel access patient-registration functions, whereas laboratory and radiology personnel access examination-related information relevant to their respective units.

The patient-registration module supports both new and returning patients. For a new patient, personnel record identity information and create an initial medical record. For a returning patient, an existing record can be retrieved using patient information or a medical-record number. The patient can subsequently be registered for outpatient or emergency services. This process helps maintain continuous visit information and reduces unnecessary duplication of patient records.

Clinical processes are supported through emergency and outpatient modules. Nurses can record initial examination information, including patient complaints and vital signs, while doctors can complete diagnostic and treatment information such as diagnoses, clinical actions, and prescribed medicines. These clinical records are associated with the corresponding patient medical record.

Supporting modules connect clinical services with cashier, pharmacy, pharmaceutical inventory, laboratory, and radiology activities. Information produced during clinical examination can therefore be used by the appropriate supporting unit without requiring repeated patient-data entry. The resulting implementation establishes an integrated workflow from patient registration and clinical examination to supporting healthcare services and reporting.

D. Mobile Telemedicine Support

The telemedicine component was implemented as a mobile-oriented interface for patients who had previously been registered by healthcare personnel. Patient access to telemedicine is initiated through the primary healthcare information system, where the corresponding patient record is associated with telemedicine access credentials.

The mobile interface provides patient authentication, a telemedicine dashboard, a healthcare request form, and access to previously submitted requests. After successful authentication, patients can select the relevant clinic and doctor, enter a health-related message, indicate whether the condition is urgent, and submit the request to the healthcare center.

The implemented telemedicine workflow consists of five principal activities:

1. healthcare personnel register an eligible patient for telemedicine access;
2. the system assigns or provides patient authentication credentials;
3. the patient logs in through the mobile-oriented interface;
4. the patient submits a telemedicine request; and
5. the patient can review the history of previously submitted requests.

The telemedicine component therefore extends selected healthcare functions beyond the internal web application by providing patients with remote access to request-based healthcare services. However, the implemented feature should be regarded as **mobile telemedicine request and remote-monitoring support**, rather than a complete synchronous teleconsultation platform. The current implementation does not include real-time video consultation, automated diagnosis, integrated electronic prescribing, or remote medical-device connectivity.

E. Database Design, Data Management, and Access Control

The database design was developed through three modeling stages: an Entity Relationship Diagram (ERD), a Conceptual Data Model (CDM), and a Physical Data Model (PDM). These models were used to represent the main healthcare data entities and their relationships before the data structures were implemented in MongoDB. The principal entities include Login, Employee, Patient, Registration, Examination, Cost Type, Prescription, Medicine, Doctor, Practice Schedule, and Polyclinic.

The ERD shown in Fig. 4 illustrates the logical relationships among the principal entities. Patient data are associated with registration records, which subsequently connect to examination and healthcare-service information. Examination records contain clinical information such as complaints, diagnoses, treatments, and physiological measurements and can be associated with prescription data. Prescription records are connected to medicine information, while doctor records are associated with practice schedules and polyclinic services. Login information is also related to healthcare personnel to support controlled system access.

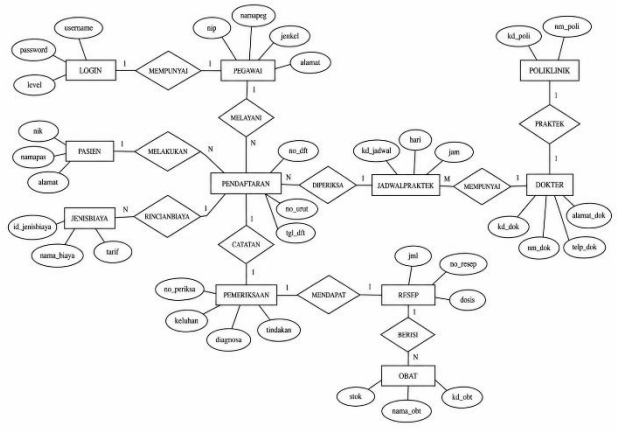


Fig. 4. Entity relationship diagram of the primary healthcare information system.

The conceptual and physical data models further refine these relationships by defining data attributes, keys, and implementation-oriented structures. Due to manuscript space limitations, the complete CDM and PDM are provided as Supplementary Material S1. These models

were used as design references before the corresponding data structures were mapped into MongoDB collections during system implementation.

In the implemented system, MongoDB was used for data storage, while Mongoose was used by the Node.js backend services to access and manage the corresponding collections. Data operations are performed through backend services rather than directly from the web or mobile clients. Client requests are routed through the API Gateway to the appropriate service, which performs the required database operation and returns the response to the client.

The application also implements role-based access control to distinguish functions available to different healthcare personnel. Users authenticate using registered credentials, and passwords are processed using bcrypt before storage. After authentication, users are provided with system functions according to their assigned access roles. The SIMPUS and telemedicine services maintain separate backend processes and corresponding data storage according to their functional responsibilities.

IV. Results and Discussion

A. System Implementation Results

The proposed system was successfully implemented as an integrated web-based primary healthcare information system with mobile telemedicine support. The web application provides role-based access to administrative, clinical, pharmaceutical, diagnostic, financial, and supporting-service functions, while the mobile-oriented interface provides selected telemedicine functions for registered patients.

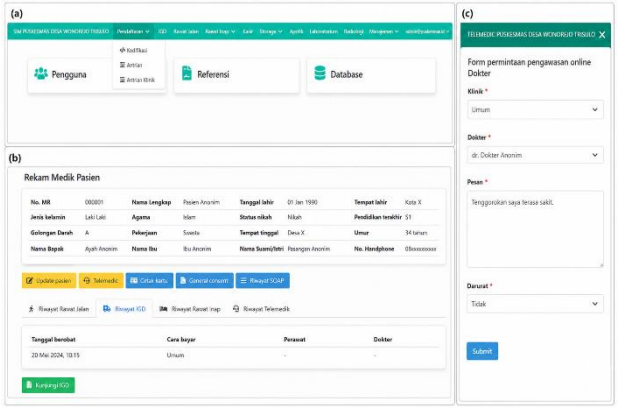


Fig. 5. Selected interfaces of the implemented system: (a) web-based healthcare management dashboard, (b) patient medical record and visit history, and (c) mobile telemedicine request form

Fig. 5 presents selected interfaces of the implemented system. The web dashboard provides access to the healthcare modules according to the assigned user role. The patient medical-record interface presents patient information and previous healthcare activities, while the mobile telemedicine interface enables patients to select a clinic and doctor, provide a health-related message, indicate the urgency of the request, and submit the request to the healthcare center.

The implementation results demonstrate that the requirements identified during the initial analysis stage were translated into operational system functions. Patient information can be reused through registration, clinical examination, pharmacy, laboratory, radiology, cashier, and other authorized workflows. In addition, the mobile telemedicine component extends selected services to patients through the same backend environment and API Gateway structure.

B. Functional Testing Results

Functional testing was conducted using 30 scenario-based black-box test cases. Each test examined whether an implemented function produced the expected response after a specified user action or input. The scenarios covered authentication and system management, patient registration, clinical services, cashier and pharmaceutical functions, diagnostic services, and telemedicine.

Table II summarizes the functional testing results according to the main functional areas.

TABLE II
SUMMARY OF FUNCTIONAL TESTING RESULTS

Functional area	No. of scenarios	Examples of functions evaluated	Result
Authentication and system management	6	Login, dashboard, user management, reference data, database management	Passed
Patient registration	5	Patient search, new and existing patient registration, outpatient and emergency registration	Passed
Clinical services	2	Emergency and outpatient examination workflows	Passed
Cashier, inventory, and pharmacy	9	Payment processing, additional charges, medicine inventory, stock batches, distribution, pharmacy services	Passed
Diagnostic services	4	Radiology and laboratory requests and processing	Passed
Telemedicine	4	Patient telemedicine registration, request handling, mobile access, request submission	Passed
Total	**30**		**30/30 Passed**

All 30 scenarios produced results consistent with the expected system behavior. The tests confirmed that users could authenticate and access appropriate functions, patient information could be processed through registration and healthcare services, and supporting functions such as pharmacy, inventory, laboratory, radiology, and cashier operations could be executed according to the defined workflows. The telemedicine scenarios also confirmed that registered patients could access the mobile-oriented interface and submit remote healthcare requests.

These findings demonstrate functional correctness for the scenarios evaluated in this study. However, the black-box tests were designed to verify system functionality rather than production-scale performance. The evaluation did not measure concurrent-user capacity, response time, throughput, database performance, service recovery, or fault tolerance. Therefore, the results should not be interpreted as evidence that the implemented microservices architecture provides better performance or scalability than other architectural approaches.

C. Initial User Perception Results

The initial user perception evaluation was based on six completed questionnaire responses from personnel at Wonorejo Trisulo Community Health Center who participated in the system trial. The respondents were actual personnel involved in healthcare-center operations. However, the available evaluation records did not document the specific professional role of each participant; therefore, balanced representation across all intended user roles could not be established.

The participants evaluated six aspects using a five-point Likert scale: application performance and stability, ease of use, efficiency and convenience, interface attractiveness, usefulness, and overall satisfaction. Because of the small sample size, the results are presented descriptively and should be interpreted as preliminary perceptions of the participating users rather than as representative measures of usability, satisfaction, or system adoption.

Table III summarizes the percentage score obtained for each aspect.

TABLE III
RESULTS OF THE INITIAL USER PERCEPTION EVALUATION

No.	Evaluation aspect	Score	Interpretation
1	Application performance and stability	80%	Agree
2	Ease of use	70%	Agree
3	Efficiency and convenience	63%	Agree
4	Interface attractiveness	80%	Agree
5	Usefulness for healthcare personnel	93%	Strongly Agree
6	Overall user satisfaction	83%	Strongly Agree
	Average	**78.2%**	**Agree**

The highest score was obtained for usefulness, at 93%, indicating a positive perception of the system's potential usefulness among the participating personnel. Overall satisfaction received 83%, while application performance and interface attractiveness both received 80%. Ease of use obtained 70%, whereas efficiency and convenience received the lowest score at 63%.

The relatively lower score for efficiency and convenience suggests that some participants may not yet have experienced an optimal interaction workflow during the trial. Possible factors include adaptation to a new digital workflow, the number of available modules, and limited familiarity with the application. However, these explanations should be treated as possible interpretations because the questionnaire did not collect qualitative feedback regarding the reasons behind individual ratings.

The average score across the six evaluated aspects was 78.2%, corresponding to the "Agree" category used in this study. Nevertheless, this result represents only preliminary descriptive feedback from the six participants and does not provide sufficient evidence to generalize usability, satisfaction, or system adoption to a broader population of healthcare personnel.

D. Comparative Analysis and Discussion

The results demonstrate that the proposed system integrates multiple primary healthcare workflows within a single web and mobile environment. Patient information recorded during registration can subsequently support clinical examination, pharmacy, inventory, laboratory, radiology, cashier, and reporting processes. The mobile telemedicine component further extends selected services by enabling registered patients to submit remote healthcare requests and review their request history.

The technologies applied in this study, including REST APIs, an API Gateway, Node.js, MongoDB, and microservices, are well-established technologies. Therefore, the contribution of this work is primarily derived from their integration and application within the operational context of a community health center rather than from the introduction of a new architectural mechanism. Table IV compares the proposed system with selected related digital healthcare systems.

TABLE IV
COMPARISON WITH RELATED DIGITAL HEALTHCARE SYSTEMS

Study	Architecture / Platform	Main Scope	Mobile / Telemedicine
Osei-Tutu and Song [6]	Microservices architecture	Healthcare information exchange	Not reported
Arcila-Diaz and Valdivia [9]	Microservices and API Gateway	Dental health records	Not reported
Lagakis et al. [10]	Microservices platform	Appointment and patient data management	Patient-oriented functions
Burka et al. [11]	Mobile app and web dashboard	Primary-care chronic disease management	Mobile application
Fu'ad et al. [15]	Cloud-based EMR	Community health center records	Not reported
Proposed system	Microservices and API Gateway	Integrated primary healthcare services and EMR	Mobile telemedicine requests

Table IV shows that the selected related systems address different aspects of digital healthcare. Previous microservices-based studies mainly focused on healthcare information exchange, dental health records, or appointment and patient data management [6], [9], [10]. Other studies emphasized mobile support for primary-care programs or electronic medical record management at community health centers [11], [15]. In comparison, the proposed system combines a broader set of administrative, clinical, pharmaceutical, diagnostic, financial, and electronic medical record functions with a microservices-based backend and mobile telemedicine support. Thus, its

contribution is primarily integrative and application-oriented, rather than based on a new microservices technology.

The functional testing results support this integration by showing that all 30 evaluated scenarios produced the expected system behavior. However, these results demonstrate functional correctness only and do not establish performance superiority over other healthcare information systems or architectural approaches. Similarly, the initial user perception assessment should be interpreted cautiously because it involved only six participants. The relatively high usefulness score suggests positive preliminary perceptions of the system, whereas the lower efficiency and convenience score indicates that further improvement of interaction workflows may still be required.

The microservices implementation separates the primary healthcare information service from the telemedicine service and uses the API Gateway to route requests to the corresponding backend service. This structure enables both web and mobile clients to access backend functions without direct database access. Nevertheless, the implementation contains a limited number of principal services and should be regarded as an initial microservices-based implementation rather than a large-scale distributed healthcare platform.

The mobile telemedicine component also remains limited to asynchronous request submission and request-history access. It does not currently provide real-time video consultation, integrated messaging, automated diagnosis, electronic prescribing, or remote medical-device integration. Future development and evaluation are therefore required to determine how the system performs under broader operational conditions and whether additional telemedicine functions can improve its practical use.

E. Study Limitations

Several limitations should be acknowledged. First, the initial user perception evaluation involved only six personnel from a single community health center. The available evaluation records also did not provide the specific professional role of each respondent. Consequently, the questionnaire results should be regarded as descriptive and exploratory and cannot be generalized to all intended healthcare-user groups.

Second, the questionnaire was locally prepared and was not based on a standardized usability instrument such as the System Usability Scale. Future evaluation should involve a larger and more diverse participant group representing different operational roles and should employ a validated usability instrument.

Third, the study did not quantitatively compare the implemented system with the previous manual workflow in terms of service duration, data-retrieval time, error rate, or reporting time. The microservices architecture was also not evaluated through performance, scalability, availability, or fault-tolerance measurements. Therefore, no claim is made regarding performance superiority over monolithic or other architectural approaches.

Finally, security evaluation was limited to the implemented authentication, role-based access, and password-hashing mechanisms. Penetration testing, security auditing, encryption evaluation, and regulatory-compliance assessment were outside the scope of this study. The telemedicine component was also limited to asynchronous request support rather than a complete real-time teleconsultation service.

V. Conclusion

This study presented the design and implementation of a microservices-based primary healthcare information system with mobile telemedicine support for Wonorejo Trisulo Community Health Center. The developed system integrates administrative, clinical, electronic medical record, pharmaceutical, diagnostic, financial, and reporting functions through a web-based application, while the mobile-oriented interface enables registered patients to submit and review telemedicine requests. The backend architecture separates the SIMPUS and telemedicine services, which communicate with client applications through REST APIs and an API Gateway and manage their corresponding MongoDB data storage.

Functional testing confirmed that all 30 evaluated scenarios produced the expected system behavior, covering authentication, patient registration, clinical services, cashier and pharmaceutical processes, diagnostic services, and telemedicine functions. The initial user perception assessment involving six healthcare personnel produced an average score of 78.2%, with usefulness receiving the highest score of 93% and overall satisfaction reaching 83%. These results indicate positive preliminary perceptions among the participating users; however, they should not be generalized as evidence of overall usability, satisfaction, or system adoption because of the limited sample size.

The main contribution of this study is the practical integration of multiple primary healthcare workflows, electronic medical records, a microservices-based backend with an API Gateway, and mobile telemedicine support within a single community health center environment. Future work should involve a larger and more diverse group of users, standardized usability instruments, quantitative comparison with previous manual workflows, and performance, scalability, and security testing. The telemedicine component may also be extended with real-time consultation, secure messaging, electronic prescription support, and integration with broader digital healthcare services.

Acknowledgements

The authors would like to express their gratitude to the management and healthcare personnel of Wonorejo Trisulo Community Health Center, Kediri, Indonesia, for

providing the necessary information and supporting the system testing and user evaluation. The authors also acknowledge Politeknik Elektronika Negeri Surabaya for the academic support provided during the development of this study.

References

[1] World Health Organization, Digital Transformation Handbook for Primary Health Care: Optimizing Person-Centred Point of Service Systems. Geneva, Switzerland: World Health Organization, 2024.

[2] Ministry of Health of the Republic of Indonesia, Regulation of the Minister of Health of the Republic of Indonesia Number 24 of 2022 concerning Medical Records. Jakarta, Indonesia, 2022.

[3] C. H. Tsai, A. Eghdam, N. Davoody, G. Wright, S. Flowerday, and S. Koch, "Effects of electronic health record implementation and barriers to adoption and use: A scoping review and qualitative analysis of the content," Life, vol. 10, no. 12, Art. no. 327, 2020, doi: 10.3390/life10120327.

[4] A. Uslu and J. Stausberg, "Value of the electronic medical record for hospital care: Update from the literature," Journal of Medical Internet Research, vol. 23, no. 12, Art. no. e26323, 2021, doi: 10.2196/26323.

[5] M. Waseem, P. Liang, M. Shahin, A. Di Salle, and G. Márquez, "Design, monitoring, and testing of microservices systems: The practitioners' perspective," Journal of Systems and Software, vol. 182, Art. no. 111061, 2021, doi: 10.1016/j.jss.2021.111061.

[6] K. Osei-Tutu and Y.-T. Song, "A microservices enterprise architecture for healthcare information exchange (HIE) in developing countries," in Proc. 10th Int. Congr. Advanced Applied Informatics (IIAI-AAI), 2021, pp. 762–767, doi: 10.1109/IIAI-AAI53430.2021.00134.

[7] M. Ilali, M. Le Berre, I. Vedel, and V. Khanassov, "Telemedicine in the primary care of older adults: A systematic mixed studies review," BMC Primary Care, vol. 24, Art. no. 152, 2023, doi: 10.1186/s12875-023-02085-7.

[8] A. A. Tierney, M. Mosqueda, G. Cesena, J. L. Frehn, D. D. Payán, and H. P. Rodriguez, "Telemedicine implementation for safety net populations: A systematic review," Telemedicine and e-Health, vol. 30, no. 3, pp. 622–641, 2024, doi: 10.1089/tmj.2023.0260.

[9] J. C. Arcila-Diaz and C. Valdivia, "A microservice-based software architecture for improving the availability of dental health records," International Journal of Computing, vol. 21, no. 4, pp. 475–481, 2022, doi: 10.47839/ijc.21.4.2783.

[10] P. Lagakis, E. Logaras, A. Billis, E. Stamkopoulos, I. Dimitriadis, A. Vakali, and P. D. Bamidis, "ClinApp: A microservices-based platform for efficient medical visit scheduling," Studies in Health Technology and Informatics, vol. 305, pp. 339–340, 2023, doi: 10.3233/SHTI230499.

[11] D. Burka, R. Gupta, A. E. Moran, J. Cohn, S. R. Choudhury, T. Cheadle, R. Mullick, and T. R. Frieden, "Keep it simple: Designing a user-centred digital information system to support chronic disease management in low/middle-income countries," BMJ Health & Care Informatics, vol. 30, no. 1, Art. no. e100641, 2023, doi: 10.1136/bmjhci-2022-100641.

[12] C. Østervang, C. M. Jensen, E. Coyne, K. B. Dieperink, and A. Lassen, "Usability and evaluation of a health information system in the emergency department: Mixed methods study," JMIR Human Factors, vol. 11, Art. no. e48445, 2024, doi: 10.2196/48445.

[13] H. A. Muchlis, "Model sukses implementasi rekam medis elektronik di Puskesmas DKI Jakarta," Indonesian of Health Information Management Journal (INOHIM), vol. 12, no. 1, pp. 31–41, 2024, doi: 10.47007/inohim.v12i01.561.

[14] D. N. Aisyah, A. H. Setiawan, C. A. Mayadewi, A. F. Lokopessy, Z. Kozlakidis, and L. Manikam, "Understanding health information systems utilization across public health centers in Indonesia: Cross-sectional study," JMIR Medical Informatics, vol. 13, Art. no. e68613, 2025, doi: 10.2196/68613.

[15] R. N. Fu'ad, S. Riyadi, D. Pratama, and R. Wicaksono, "Implementation of an integrated cloud-based electronic medical record system at community health center," Sinkron: Jurnal dan Penelitian Teknik Informatika, vol. 9, no. 4, pp. 3269–3277, 2025, doi: 10.33395/sinkron.v9i4.15206.

[16] H. Y. Martono, A. A. Yunanto, and D. S. Santoso, "Satu Sehat Interoperability Agent," in Proc. 2024 International Conference on Information Technology Systems and Innovation (ICITSI), 2024, pp. 236–241, doi: 10.1109/ICITSI65188.2024.10929376.

[17] A. A. Yunanto, F. F. Putri, D. I. Permatasari, Nailussa'ada, F. F. Hardiansyah, U. Sa'adah, and A. S. Aziz, "Design and Implementation the Prayer Reminder Application using KISS Principle based on User Centered Design," Procedia Computer Science, vol. 234, pp. 1484–1491, 2024, doi: 10.1016/j.procs.2024.03.149.

Journal of Data Science, Informatics and Computer Engineering (JDSICE),
Vol. 1, Issue 1
September 2026
ISSN: XXXX-YYYY

A Mobile-Based BISINDO Recognition and Communication Support System Using MediaPipe, MLP, and GRU

Adrian Yoga Chrisarianto[1]*, Entin Martiana Kusumaningtyas[2], Aliridho Barakbah[3], Ira Prasetyaningrum[4]
[1]*Associate Degree Program in Computer Science, Electronic Engineering Polytechnic Institute of Surabaya, Surabaya, Indonesia*
**Corresponding author: Adrian Yoga Chrisarianto; email: adrianyogachrisarianto@email.com*

Manuscript received August 14, 2026, revised September 11, 2026, accepted September 22, 2026
doi: zz.yyyyy/xxxxx.pp134-1399

Abstract – *Communication between Indonesian Sign Language users and the general public remains limited because public familiarity with the language is low, while many existing recognition depend on dedicated sensors, run on desktop or web platforms, or recognize only static alphabets. This study develops a landmark-based mobile BISINDO recognition system that supports introductory communication through alphabet recognition, dynamic word recognition, and text-to-speech output. The system provides a mobile-oriented translation workflow using a Streamlit-based interface, camera streaming, prediction stabilization, and text-to-speech output through an external mobile integration bridge. The alphabet dataset contains 13,152 landmark samples, and the word dataset contains 3,600 sequences. In live-camera testing, the alphabet model correctly classified 233 of 260 trials (89.62%), whereas the word model correctly classified 282 of 360 trials (78.33%). Across three users, alphabet accuracy ranged from 84.62% to 100%, and word accuracy ranged from 77.78% to 83.33%. Performance was strongest at 50–80 cm and declined at 110 cm. Sentence construction capability was evaluated by combining recognized alphabet and word outputs into introductory communication sequences. The findings demonstrate the feasibility of an integrated BISINDO recognition workflow for introductory communication scenarios, while future improvements should focus on vocabulary expansion, broader signer-independent evaluation, and mobile deployment optimization.*

CC BY SA

Keywords: *Dynamic gesture recognition, Indonesian Sign Language, landmark-based hand recognition, mobile assistive technology, multilayer perceptron*

I. Introduction

Communication barriers remain a practical problem for Deaf and speech-impaired people when communicating with members of the public who do not understand sign language. The World Health Organization estimates that approximately 1.5 billion people worldwide live with some degree of hearing loss [1]. A survey cited in the accompanying thesis also reported limited public familiarity with Indonesian Sign Language, with only 15% of 100 respondents indicating strong familiarity [2]. Indonesian Sign Language, commonly known as BISINDO, is a visual language used by the Indonesian Deaf community through hand gestures, facial expressions, and body movement. Word signs may vary across regions, while alphabetic finger spelling is comparatively consistent [3]. Mobile devices provide a practical channel for assistive communication. Indonesian telecommunications statistics reported that 68.65% of the population owned a mobile phone in 2024 [4]. A camera-based recognition that does not require specialized gloves or external sensors can therefore reduce adoption barriers, provided that the recognition model is sufficiently accurate, computationally manageable, and robust to signer and environmental variation. Previous studies address only parts of this problem. SETARA used a camera-based Android approach for BISINDO words but emphasized qualitative effectiveness and vocabulary expansion [5]. A flex-sensor glove using k-nearest neighbors achieved 65.38% accuracy but depended on dedicated hardware [6]. U-SIBI provided a web learning environment for SIBI alphabets and vocabulary [7]. A random-forest BISINDO alphabet recognition reported 91.28% accuracy but focused on static alphabet recognition and desktop use [8].

A convolutional neural network web translator recognized SIBI and BISINDO alphabets but did not integrate dynamic BISINDO words into a mobile communication workflow [9].

TABLE I
SYSTEMATIC COMPARISON WITH CLOSELY RELATED SYSTEMS

Study	Scope	Platform and method	Main Difference/ Limitation or distinction
SETAR A [5]	BISIND O words	Android; camera; convolutional network and MediaPipe	Focuses on BISINDO word recognition, but vocabulary coverage and quantitative evaluation remain limited.
Flex glove [6]	SIBI letters and numbers	Sensor glove; k-nearest neighbors	Achieves 65.38% accuracy, but requires additional hardware for gesture acquisition..
U-SIBI [7]	SIBI alphabet and vocabul ary	Web; convolutional network and Flask	Designed mainly as a learning platform and does not provide an integrated BISINDO mobile communication workflow.
Rando m forest [8]	BISIND O alphabet	Desktop; random forest	Provides alphabet classification performance, but only handles static gestures without dynamic words.
Web CNN [9]	SIBI and BISIND O alphabet s	Web; VGG16 convolutional network	Focuses on image-based alphabet recognition without integrated mobile word recognition.
This study	26 BISIND O letters and 18 phrases	MediaPipe, multilayer perceptron, gated recurrent unit, Flutter	Integrates static and dynamic recognition, mobile camera interaction, sentence construction, text-to-speech output, although vocabulary size and signer generalization remain limited.

Study	Scope	Platform and method	Main Difference/ Limitation or distinction
This study	26 BISIND O letters and 18 phrases	MediaPipe, multilayer perceptron, gated recurrent unit, Flutter	Integrates static and dynamic recognition, mobile camera interaction, sentence construction, text-to-speech output, although vocabulary size and signer generalization remain limited.

Recent work has introduced word-level BISINDO datasets and temporal baselines [10], [11], demonstrating that dynamic recognition is an active research direction. However, the supplied literature does not show a single system that combines static finger spelling, dynamic introductory phrases, live mobile access, stabilized output, local history, and speech output. The gap addressed here is therefore primarily a system-integration and practical-evaluation gap rather than a claim of a new learning algorithm.

Unlike previous studies that mainly focus on alphabet recognition or isolated word recognition, this study contributes an integrated BISINDO communication support workflow combining static alphabet recognition, dynamic word recognition, mobile camera interaction, sentence construction, and text-to-speech output. The novelty of this work does not lie in proposing a new recognition algorithm, but in integrating established techniques into a practical BISINDO-oriented assistive communication prototype.

The contributions are threefold. First, the study integrates landmark-based static and temporal recognition in one recognition and communication support system: a multilayer perceptron classifies 26 alphabet gestures, while a gated recurrent unit processes 18 dynamic words and phrases. Second, it evaluates the system through internal validation, live-camera trials, three-user tests, camera-distance tests, and predefined sentence construction. Third, it implements a one-way mobile communication workflow using Flutter, WebView, local SQLite history, and text-to-speech output. Section II describes the proposed system and experiment, Section III presents the results and scientific discussion, and Section IV concludes the paper.

II. PROPOSED SYSTEM AND EXPERIMENTAL METHOD

A. Design and Method Selection

The system avoids direct classification of full image pixels. MediaPipe Hands extracts 21 three-dimensional landmarks from each detected hand [12]. Each detected hand is represented as a set of 21 three-dimensional landmarks, where each landmark consists of normalized coordinates (x_i, y_i, z_i):

$$L = \{ (x_i, y_i, z_i) \mid i = 0, 1, \ldots, 20 \} \quad (1)$$

Therefore, each hand produces 63 coordinate features calculated as:

$$F_\{landmark\} = 21 \times 3 = 63 \quad (2)$$

Landmark-based representations reduce dependence on background pixels and allow spatial relationships to be represented through normalized coordinates, inter-point distances, and orientation features [13]. For alphabet recognition, the landmark representation is combined with geometric descriptors to construct the input vector for the multilayer perceptron. The final feature representation is formulated as:

$$F_\{MLP\} = 63 + (5 + 10 + 5 + 2) = 85 \quad (3)$$

where the additional features consist of fingertip-to-wrist distances, fingertip pair distances, joint angles, and handedness indicators. The multilayer perceptron is selected for static alphabet gestures because its input is a fixed numerical feature vector; previous work has demonstrated the suitability of MediaPipe features with multilayer perceptrons for hand-gesture classification [14].

$$X_GRU \in R\text{^}(30 \times 127) \quad (4)$$

where each sequence consists of 30 frames and each frame contains 127 features, calculated from one hand-count feature and two hand landmark representations:

$$127 = 1 + 63 + 63 \quad (5)$$

The gated recurrent unit is selected for dynamic words because it can model temporal dependencies across landmark sequences while using fewer gates than a long short-term memory network [15]. A majority voting mechanism using a six-frame sliding buffer with a three-vote threshold is applied to reduce frame-level fluctuations [16].

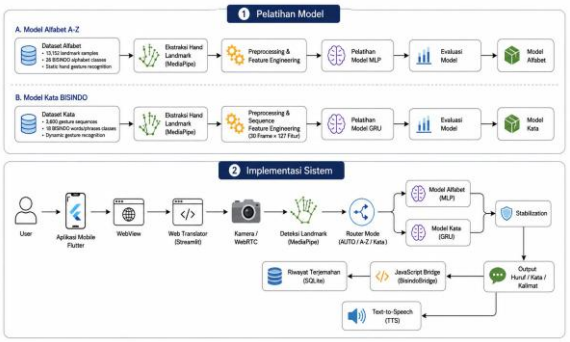


Fig. 1. Overall architecture of the proposed BISINDO recognition and communication support system using MediaPipe, MLP, and GRU.

B. Dataset Acquisition and Representation

The alphabet dataset contains 13,152 landmark samples distributed across 26 classes, with 501–542 samples per class. Each sample represents one static BISINDO alphabet gesture and is converted into an 85-dimensional feature vector consisting of normalized landmark coordinates and geometric descriptors. The dynamic dataset contains 3,600 sequences covering 18 introductory words and phrases, with 200 sequences per class. Each dynamic sequence consists of 30 frames, where each frame is represented by 127 temporal landmark features.

The datasets were collected from two voluntary signers for both alphabet and dynamic word recognition. Participants A and B contributed to the dataset collection, and both participants were also included in the real-time evaluation process, while Participant C was an additional unseen user used for generalization testing. Before data collection, participants were informed about the purpose of the study and agreed to participate voluntarily.

The recording process was conducted under several practical environmental conditions, including a laboratory environment with stable lighting, a boarding-room environment with moderate lighting, and a coffee-shop environment with less stable lighting. The camera was positioned at several distances representing near, medium, and relatively far usage conditions. Each recorded sample was manually checked to ensure consistency between the performed gesture and its corresponding BISINDO label before being used for training. The dynamic word dataset was divided using an 80/20 stratified split, resulting in 2,880 training sequences and 720 validation sequences, with approximately 40 validation sequences allocated for each class.

The internal dataset partition was used for model training and validation, while generalization capability was further examined through live-camera evaluation involving an additional unseen participant. However, the current study does not represent a fully signer-independent training-validation protocol because two evaluation participants were also involved in dataset collection. Therefore, the remaining performance gap between validation accuracy and real-world testing may

be affected by signer variation, gesture execution style, and environmental conditions. Future work should adopt a larger subject-independent evaluation protocol with separated training and testing signers.

Although Participant C represents an unseen signer evaluation, the current experiment does not represent a fully signer-independent protocol because Participants A and B contributed to dataset collection and were also included in the evaluation process. Therefore, the obtained results should be interpreted as prototype-level validation under the tested conditions rather than as a general representation of BISINDO recognition performance across diverse users. Future work should involve a larger number of independent signers with separated training and testing participants to evaluate broader generalization capability.

C. Training Configuration and Reproducibility

Both training curves in the thesis extend to 100 epochs. The word pipeline uses a 30-frame by 127-feature input, while the alphabet pipeline uses a dual-hand feature vector. The final training configuration used in this study is summarized in Table II to support reproducibility.

TABLE II
MODEL AND TRAINING CONFIGURATION

Parameter	Alphabet Model	Word Model
Input	85-dimensional engineered feature vector	30×127 landmark sequence
Feature extraction	MediaPipe + geometric features	MediaPipe temporal landmarks
Classes	26 alphabets	18 words/phrases
Model	MLP	GRU
Architecture	MLP with Dense layers (256, 128, 64) and 26-class Softmax output.	GRU(64) → Dense(32) → Dense(18 Softmax)
Activation	ReLU + Softmax	tanh + ReLU + Softmax
Dropout	0.35/0.30/0.20	0.20
Optimizer	Adam	Adam
Learning rate	0.001	0.001
Epoch	100	100
Batch size	32	32
Loss	Categorical CE + LS 0.05	Categorical CE + LS 0.05
Split	80/20 stratified	80/20 stratified sequence split
Evaluation	Accuracy, F1, confusion matrix	Accuracy from validation and real-time testing

Camera distance evaluation was conducted at 50 cm, 80 cm, and 110 cm. Ten predefined introductory sentences were tested until the target sentence was successfully produced. The proposed system was implemented as a mobile-based application; however, Detailed analysis of application-level performance, such as execution latency, frame rate stability, and hardware resource consumption, is outside the scope of this study. Therefore, the current evaluation focuses on recognition accuracy and functional camera-based interaction rather than detailed mobile hardware performance analysis.

D. Mobile Integration and Evaluation Protocol

The recognition system is executed through a Streamlit web interface displayed inside a Flutter WebView. Camera frames are acquired through the web recognition, processed by MediaPipe, routed to alphabet, word, or automatic mode, stabilized, and returned as letters, words, or sentences. The system provides a JavaScript communication bridge for transferring recognized text to a mobile host application, where additional storage and speech services can be implemented [17]–[19]. This is a one-way recognition workflow that converts recognized signs into text and speech output; it does not support reverse translation from speech or text into signs.

Internal evaluation of the alphabet model used accuracy, loss, macro F1-score, and confusion matrix analysis, while the word model evaluation was based on validation accuracy obtained during model evaluation and real-time camera testing. Live evaluation was conducted using a smartphone equipped with a Snapdragon 6s Gen 3 processor, 8 GB RAM, Android 16 operating system, and dual cameras (50 MP main camera and 8 MP front camera). Alphabet testing consisted of 10 trials for each of 26 classes (260 trials), while word testing consisted of 20 trials for each of 18 classes (360 trials). User variation was evaluated using three participants, including two participants involved in dataset collection and one additional unseen participant. Camera distance experiments were conducted at 50 cm,
80 cm, and 110 cm. Ten predefined introductory sentences were tested to evaluate sentence construction capability, where each sentence was attempted until the expected output was produced. Although the system was evaluated through real-time camera interaction,

comprehensive device-level benchmarking, including application latency, frame rate, and resource consumption, was not conducted.

III. RESULTS AND DISCUSSION

The testing was conducted using a smartphone equipped with a Snapdragon 6s Gen 3 chipset (8 GB of RAM).

A. Internal Model Behaviour

For both models, the training history reports final training accuracy of approximately 0.86–0.87 and validation accuracy of approximately 0.90–0.91. Training loss decreased from about 1.60 to 0.76, and validation loss decreased from about 1.07 to 0.63. The alphabet confusion matrix is predominantly diagonal, with errors concentrated among visually similar hand configurations. The thesis identifies confusions involving C and L and additional pairs such as U–C, P–K, H–D, B–F, and T–K.

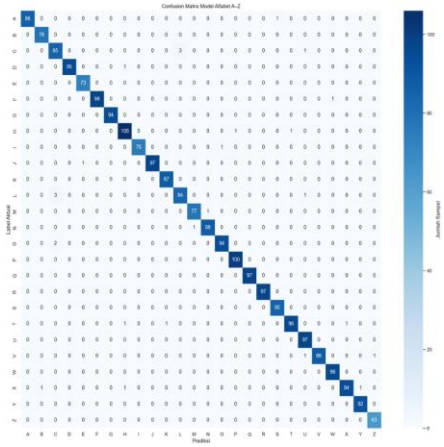


Fig. 2. Confusion matrix from the validation test on the A–Z Alphabet (MLP) model.

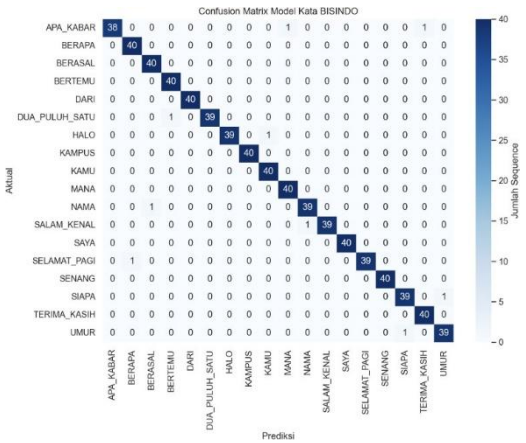


Fig. 3.Validation confusion matrix for the 18-word BISINDO GRU model

The word confusion matrix shows a strong diagonal pattern, indicating that most BISINDO word classes were correctly classified by the GRU model. The validation results demonstrate that the model successfully learned temporal patterns from the gesture sequences under the prepared dataset conditions. However, these internal validation results should be interpreted carefully because they represent performance on the collected dataset and do not fully represent recognition performance for unseen users and uncontrolled environments. The narrative reports approximately 0.90–0.91 validation accuracy, while another passage states that, and the displayed matrix contains counts close to 40 per class rather than the 20 stated in the text. Consequently, the validation results should be interpreted with careful consideration because consistency between reported metrics and exported evaluation artifacts is required to ensure reproducibility. The present discussion therefore treats the live-camera results as the more relevant deployment evidence rather than presenting the internal matrix as proof of signer-independent generalization.

B. Live-Camera Performance

TABLE III
LIVE-CAMERA Results

Test	Alphabet model	Word model
Overall manual test	233/260 (89.62%)	282/360 (78.33%)
User A	100.00%	83.33%
User B	84.62%	83.33%
User C	84.62%	77.78%
50 cm	100.00%	80.00%
80 cm	100.00%	80.00%
110 cm	84.61%	60.00%

The alphabet model achieved 89.62% across 260 live trials, while the dynamic word model achieved 78.33% across 360 trials. The lower word accuracy is consistent with the greater sensitivity of temporal gestures to movement speed, repetition timing, hand trajectory, and landmark interruption. Multi-user results also show that performance changes across signers. These tests are useful practical checks, but three participants are insufficient for a strong subject-independent claim.

Camera distance affected both models. Alphabet accuracy remained 100% at 50 and 80 cm and fell to 84.61% at 110 cm. Word accuracy was 80% at 50 and 80 cm and fell to 60% at 110 cm. As the hand occupies fewer pixels at longer distances, MediaPipe landmark localization becomes less stable. The system should therefore communicate a recommended operating range of 50–80 cm and should not present the current results as

distance-invariant performance.

C. Conversational Function and Practical Contribution

All ten predefined introductory sentences were eventually produced in 17 total attempts, corresponding to 1.7 attempts per sentence. Six sentences succeeded on the first attempt, giving a first-attempt sentence success rate of 60%. The remaining sentences required two to four attempts, particularly when switching between word mode and alphabetic finger spelling. The appropriate interpretation is therefore 100% eventual completion on a closed set of ten target sentences, not 100% open-conversation accuracy.

The practical contribution is the integration of alphabet and dynamic-word recognition with saved history and speech output. Nevertheless, the current 18-word vocabulary corresponds to only 36%–56.25% of the class counts in three external word-level BISINDO resources discussed in the thesis [10], [11]. These percentages describe class-count coverage only and do not establish accuracy superiority. Direct comparison with the 91.28% random-forest alphabet result [8] is also inappropriate because the datasets, users, and evaluation protocols differ. The present system should therefore be positioned as an integrated prototype rather than a state-of-the-art accuracy claim.

D. Explanation of the Validation-to-Deployment Gap

The decrease from approximately 0.90–0.91 internal validation accuracy to 78.33% live word accuracy indicates a domain shift between curated sequences and uncontrolled camera input. The thesis identifies lighting, camera angle, user anatomy, movement speed, gesture consistency, and landmark stability as plausible contributors. A further concern is that the dataset split is not documented as signer-independent. A rigorous revision should use a subject-independent protocol, such as leave-one-signer-out validation or a held-out signer test set, and report macro precision, recall, F1-score, confidence intervals, and per-class live errors. An ablation comparing raw landmarks, engineered features, and prediction stabilization would also clarify which components produce the observed gains.

IV. CONCLUSION

This study presents a landmark-based mobile BISINDO recognition system that combines a multilayer perceptron for 26 static alphabet classes and a gated recurrent unit for 18 dynamic introductory words and phrases. Live-camera evaluation achieved 89.62% accuracy for the alphabet model and 78.33% accuracy for the word model. The system achieved its best operating performance at camera distances of 50–80 cm, and ten predefined introductory sentences were successfully constructed by combining recognized alphabet and word outputs. These findings demonstrate the preliminary feasibility of an integrated mobile-oriented BISINDO recognition workflow for limited introductory communication scenarios. However, the current system should be considered as a communication support prototype rather than a complete BISINDO translation system, since vocabulary coverage and signer-independent evaluation remain limited.

Appendix

A video documenting the results of the sentence construction and real-time detection tests can be accessed via the additional repository link at: https://drive.google.com/drive/folders/19BiXZH5LSocU3FtSEIainsH8W7FdCyp0:

Acknowledgements

The author would like to thank the Department of Information and Computer Engineering at the Surabaya State Electronics Polytechnic (PENS) for providing research facilities. Special thanks are also extended to Ms. Entin Martiana K., Mr. Ali Ridho Barakbah, and Ms. Ira Prasetyaningrum, the academic advisors who provided comprehensive guidance throughout the process, from the design to the evaluation of this system.

References

[1] World Health Organization, World Report on Hearing. Geneva, Switzerland: World Health Organization, 2021.
[2] "Perancangan Media Pembelajaran Interaktif Pengenalan Dasar-Dasar BISINDO untuk Masyarakat Non-Tunarungu," Telkom University Repository, 2025.
[3] Direktorat SMP Kemendikbudristek, "Mengenal Bahasa Isyarat di Indonesia: SIBI dan BISINDO," 2023.
[4] Badan Pusat Statistik, Statistik Telekomunikasi Indonesia 2024. Jakarta, Indonesia: BPS, 2025.
[5] J. T. Sihotang and A. N. A. Adena, "SETARA (Aplikasi Penerjemah Bahasa Isyarat) sebagai Solusi Disekuilibrium Komunikasi dan Pendidikan di Kota Semarang," Journal for Energetic Youngsters, vol. 2, no. 1, pp. 27–45, 2024.
[6] R. A. Zakaria, A. Firmansyah, and Y. A. Prabowo, "Rancang Bangun Flex Sensor Gloves untuk Penerjemah Bahasa Isyarat Menggunakan K-Nearest Neighbors," in Seminar Nasional Sains dan Teknologi Terapan VII, Surabaya, Indonesia, 2019.
[7] N. A. Dzikri and D. A. Rosdiana, "Aplikasi Website Penerjemah Bahasa Isyarat sebagai Media Pembelajaran SLB Negeri Cicendo Kota Bandung," Jurnal Ilmiah Rekayasa dan Manajemen Sistem Informasi, vol. 10, no. 1, pp. 42–50, 2025.
[8] A. Hadi and L. L. Van FC, Rancang Bangun Aplikasi Penerjemah BISINDO Menggunakan Metode Random Forest Classifier, undergraduate thesis, Universitas Lancang Kuning, Pekanbaru, Indonesia, 2023.
[9] J. A. Jong, Aplikasi Penerjemah Bahasa Isyarat Menggunakan Convolutional Neural Network Berbasis Web, internship report, Universitas Katolik De La Salle Manado, Manado, Indonesia, 2024.
[10] G. O. Kindy, G. Leonali, and H. Lucky, "Word-Level BISINDO: A Novel Video Indonesian Sign Language Dataset

and Baseline Methods," Procedia Computer Science, vol. 269, pp. 249–258, 2025, doi: 10.1016/j.procs.2025.08.277.
[11] M. D. Kautsar, A. M. Hariono, and R. Akmal, "Attention vs LSTM: Improving Word-Level BISINDO Recognition," arXiv:2409.01975, 2024.
[12] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, and M. Grundmann, "MediaPipe Hands: On-Device Real-Time Hand Tracking," arXiv:2006.10214, 2020.
[13] J. Shin, A. S. M. Miah, M. H. Kabir, M. A. Rahim, and A. A. Shiam, "A Methodological and Structural Review of Hand Gesture Recognition Across Diverse Data Modalities," IEEE Access, vol. 12, pp. 142606–142611, 2024.
[14] S. Barman and S. Majumdar, "Person-Independent Hand Gesture Recognition Using MediaPipe and Multi-Layer Perceptron," in Smart Systems and Wireless Communication, vol. 433. Singapore: Springer, 2025, doi: 10.1007/978-981-96-1348-9_1.
[15] B. Subramanian, B. Olimov, S. M. Naik, S. Kim, K.-H. Park, and J. Kim, "An Integrated MediaPipe-Optimized GRU Model for Indian Sign Language Recognition," Scientific Reports, vol. 12, Art. no. 11964, 2022.
[16] M. F. Wahid, R. Tafreshi, and R. Langari, "A Multi-Window Majority Voting Strategy to Improve Hand Gesture Recognition Accuracies Using Electromyography Signal," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 28, no. 2, pp. 427–436, 2020, doi: 10.1109/TNSRE.2019.2961706.
[17] Flutter, "Flutter Documentation," 2023.
[18] flutter_tts, "Text to Speech Plugin for Flutter," 2023.
[19] SQLite, "SQLite Documentation," 2023.

Journal of Data Science, Informatics and Computer Engineering (JDSICE),
Vol. 1, Issue 1
February 2026
ISSN: XXXX-YYYY

Sentiment Classification and Category Mapping of Public Perception on Tax Service Quality at KPP Pratama Surabaya Gubeng Using Naive Bayes and Support Vector Machine

Puput Ayu Setiawati[1], Fitrah Maharani Humaira[2], Diah Kusumadewi Sugiarto[3]
[1]*Electronic Engineering Polytechnic Institute of Surabaya, Indonesia*
[2]*Electronic Engineering Polytechnic Institute of Surabaya, Indonesia*
[3]*KPP Pratama Surabaya Gubeng, Indonesia*
**Corresponding author: Third Author; email: puputayusetiawati2@gmail.com*

Manuscript received April 9, 2026, revised September 17, 2026, accepted September 21, 2026
doi: zz.yyyyy/xxxxx.pp134-1399

Abstract – *Public feedback on tax services provides valuable information for evaluating service quality, but the large volume of textual opinions requires systematic analysis to identify sentiment and service-related concerns. This study aims to classify public sentiment toward the services of KPP Pratama Surabaya Gubeng, map reviews into service-related categories, and compare the performance of Naive Bayes and Support Vector Machine (SVM). Data were collected from Google Maps reviews and Google Forms, resulting in 465 records. After preprocessing, 365 valid reviews remained and were mapped into three categories is service, place, and overall. Sentiment labeling produced 293 positive and 72 negative reviews. Undersampling was then applied to obtain a balanced dataset of 120 reviews, consisting of 60 positive and 60 negative samples. The balanced dataset was evaluated using 70:30, 80:20, and 90:10 training-testing ratios, corresponding to 84/36, 96/24, and 108/12 training/testing samples, respectively. The best reported performance was achieved by Naive Bayes, with 75.00% accuracy, 77.14% precision, 75.00% recall, and 75.56% F1-score, while SVM achieved 72.22% across the reported metrics. These findings indicate that Naive Bayes provides better classification performance for the characteristics of the collected review data and that category mapping provides additional information about the service aspects discussed by the public.*

CC BY SA

Keywords: *Tax, Sentiment, Naive Bayes, Support Vector Machine, Public opinion*

I. Introduction

The Tax Service Office (KPP) Pratama Surabaya Gubeng is a government institution responsible for providing taxation services to the public. The quality of tax services is important because public experiences and perceptions can provide information for evaluating service delivery. In the digital era, public feedback can be collected through online platforms, including Google Maps and online questionnaires such as Google Forms [1]. Such textual feedback can be analyzed using text mining techniques to identify information and patterns within large collections of data [2].

The increasing availability of textual feedback requires a systematic approach to process and interpret public opinions. Reviews may express different sentiments and address various aspects of service delivery. Sentiment analysis can be used to classify opinions according to their polarity [3], while category mapping can provide additional information about the aspects most frequently discussed. In tax services, this approach can distinguish opinions related to service quality, physical facilities, and overall evaluations.

Machine learning methods are widely used for text classification, including Naive Bayes and Support Vector Machine (SVM) [4]. Naive Bayes uses a probabilistic approach that is relatively simple and efficient for text classification [5], whereas SVM constructs a decision boundary and is suitable for high-dimensional feature spaces commonly found in text data [6]. Previous studies have applied Naive Bayes to public service sentiment analysis [7] and compared Naive Bayes and SVM in review classification, showing that performance may vary according to dataset characteristics [8]. Other research has highlighted the importance of preprocessing and data

balancing in sentiment classification [9]. These findings demonstrate the applicability of machine learning for digital opinion analysis, while research on taxation services in Indonesia using multi-source public feedback remains limited.

Based on this research gap, this study analyzes public perceptions of KPP Pratama Surabaya Gubeng by integrating Google Maps reviews and Google Forms responses. The data are processed through preprocessing, topic mapping, sentiment labeling, data balancing, and classification using Naive Bayes and SVM. The reviews are mapped into three categories, namely Service to "pelayanan", Place to "tempat", and Overall, to identify the aspects most frequently discussed by the public. Sentiment labeling combines numerical ratings with textual polarity analysis [10]. The performance of Naive Bayes and SVM is compared using accuracy, precision, recall, and F1-score. The contribution of this study is the integration of multi-source public opinion data with sentiment classification and category mapping in the context of tax service quality at KPP Pratama Surabaya Gubeng. The study also provides an empirical comparison of Naive Bayes and SVM on the collected review data. The findings are expected to provide a data-driven view of public perceptions and identify service-related issues that can support the evaluation and improvement of tax service quality.

II. Methodology

The research methodology consists of a series of systematic stages for collecting, processing, classifying, and evaluating public opinion regarding the tax services provided by KPP Pratama Surabaya Gubeng. The research process consists of data collection, data exploration, text preprocessing, topic mapping, sentiment labeling, data balancing, classification modeling, and model evaluation. The overall research procedure is presented in Fig. 1.

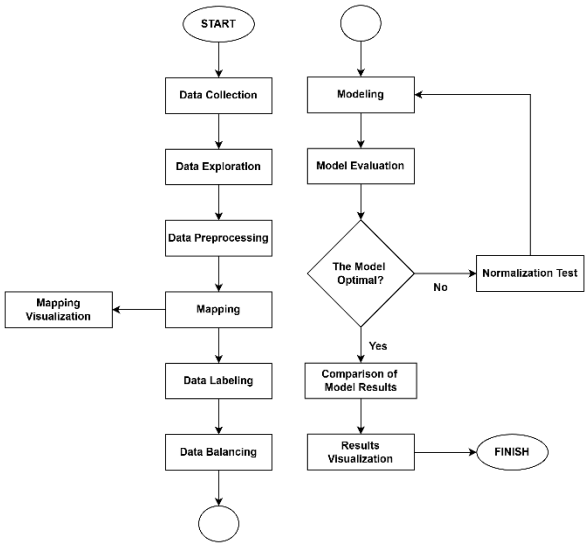


Fig. 1. Research Procedure Flowchart

A. Data Collection

The data collection stage aims to obtain public opinions regarding the services of KPP Pratama Surabaya Gubeng. Data were collected from two primary sources: user reviews on Google Maps and responses to an online questionnaire distributed through Google Forms [1], [11]. Google Maps reviews were collected using a web scraping technique [12], [13], while Google Forms was used to obtain direct responses from participants.

The Google Maps data cover reviews from 2016 to 2025, while the Google Forms questionnaire was distributed voluntarily through platform X from January 20, 2025, to June 1, 2025. A total of 465 records were collected, consisting of 311 Google Maps reviews and 154 Google Forms responses. Each record contains several attributes, including username, date, rating, and review text. These attributes were used as the primary data for subsequent analysis and classification.

TABLE I
DATA INFO

No.	Variabel	Description
1	*Username*	The name of the user who provided the review
2	*Date*	The date when the review was submitted
3	*Rating*	Numerical rating on a scale of 1 to 5 stars
4	*Review*	Textual content of the user's opinion

B. Data Exploration

The data exploration stage begins by integrating two primary data sources, namely Google Maps reviews and responses from Google Forms questionnaires, into a unified dataset. This process involves schema harmonization, including renaming columns such as 'Tanggapan' to 'Review', standardizing 'Timestamp' into 'Date', and consolidating identity-related fields into 'Username'. Irrelevant columns that do not contribute to sentiment analysis are removed to maintain data consistency and quality.

Following the integration process, an initial analysis is conducted to examine the distribution of ratings. In addition, temporal analysis is performed to observe the trend of review counts per year from 2016 to 2025, providing insights into fluctuations in data volume over time. Visualization techniques are employed to present an overview of the dataset characteristics before proceeding to further data processing stages [2].

The data exploration indicate that the dataset is highly imbalanced. The majority of reviews are dominated by 5-star ratings (68.6%), followed by 4-star ratings (15.5%). In contrast, lower ratings (1 to 3 stars) account for less than 10% of the dataset. This distribution suggests that users tend to provide feedback primarily when they are either highly satisfied or dissatisfied. The distribution of reviews

by rating and year is presented in Fig. 2 and Fig. 3, respectively.

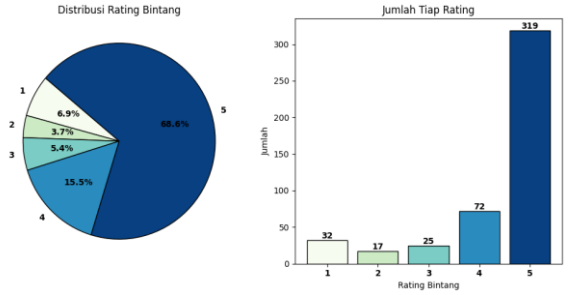


Fig. 2. Visualization of Star Rating Distribution

In addition, temporal analysis shows fluctuations in the number of reviews over time, with significant increases observed in 2022 and 2025. These trends may indicate the influence of external factors such as service improvements or policy changes.

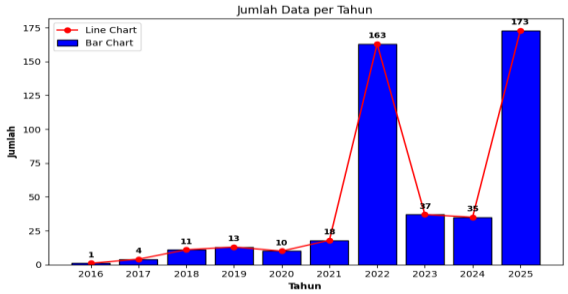


Fig. 3. Number of Review Data per Year

C. Data Preprocessing

The data preprocessing stage is conducted to clean and prepare textual data before the classification process. This step is essential to ensure that the input data are consistent, meaningful, and free from noise. Proper preprocessing can significantly improve the performance of text classification models [14], [15], especially for public opinion data, which often contain variations in writing styles. In this study, preprocessing is performed through six main steps as summarized in Table II.

TABLE II
DATA PREPROCESSING STAGES

No.	Step	Description
1	*Text Cleaning*	Removing numbers, punctuation, and irrelevant symbols.
2	*Case Folding*	Converting all text to lowercase.
3	*Normalization*	Standardizing informal or non-standard words.
4	*Tokenization*	Splitting sentences into individual tokens
5	*Stopword Removal*	Removing common words with low semantic value
6	*Stemming*	Converting words into their root forms.

This process refers to common practices applied in various text-based sentiment analysis studies in Indonesia. Proper application of preprocessing steps can significantly improve model performance in Indonesian language sentiment classification [16].

After preprocessing, 365 valid reviews remained. The reduction from 465 to 365 records was caused by the removal of empty, duplicate, or irrelevant records. The dataset was subsequently used for topic mapping and sentiment labeling. The preprocessing stage improves data quality by ensuring consistency and reducing noise, which is essential for accurate sentiment classification. An example of the preprocessing result is presented in Fig. 4.

Review	Year	Normalization	Tokenizing	Stopword	Stemming
pelayanan di kpp pratama surabaya gubeng sanga...	2025	pelayanan di kpp pratama surabaya gubeng sanga...	pelayanan, di, kpp, pratama, surabaya, gubeng,...	pelayanan, kpp, pratama, surabaya, gubeng, mem...	layan, kpp, pratama, surabaya, gubeng, muas, t...
pelayanan di kpp pratama surabaya gubeng sanga...	2025	pelayanan di kpp pratama surabaya gubeng sanga...	pelayanan, di, kpp, pratama, surabaya, gubeng,...	pelayanan, kpp, pratama, surabaya, gubeng, ram...	layan, kpp, pratama, surabaya, gubeng, ramah, ...
saya sangat mengapresiasi pelayanan yang diber...	2025	saya sangat mengapresiasi pelayanan yang diber...	saya, sangat, mengapresiasi, pelayanan, yang, ...	mengapresiasi, pelayanan, kpp, pratama, suraba...	apresiasi, layan, kpp, pratama, surabaya, gube...
saya merasa puas dengan pelayanan di kpp prata...	2025	saya merasa puas dengan pelayanan di kpp prata...	saya, merasa, puas, dengan, pelayanan, di, kpp...	puas, pelayanan, kpp, pratama, surabaya, guben...	puas, layan, kpp, pratama, surabaya, gubeng, a...

Fig. 4. Data Preprocessing Result

D. Mapping

After completing the preprocessing stage, the next step is data mapping to categorize each review based on the discussed topic. This process aims to identify the aspects most frequently highlighted in public opinions. The mapping is performed using a rule-based approach by detecting specific keywords within each preprocessed and stemmed review. This approach is suitable for relatively small datasets and enables fast and interpretable initial classification [17]. In this study, three main categories are used for topic mapping, as summarized in Table III.

TABLE III
TOPIC MAPPING CATEGORIES

No.	Variabel	Description
1	*Service*	Reviews containing keywords related to service quality or staff interaction (e.g., ‘layan’, ‘cepat’, 'ramah').
2	*Place*	Reviews referring to physical conditions or facilities (e.g., ‘nyaman’, ‘parkir’, ‘bersih’)
3	*Overall*	Reviews that do not contain keywords from the above categories and represent general opinions

E. Data Labeling

Sentiment labeling was performed to classify each review into two sentiment classes Positive and Negative. A hybrid labeling approach was applied by combining numerical rating information with textual polarity analysis using TextBlob [10].

TextBlob provides a polarity score ranging from -1 to +1, where a higher score indicates a more positive textual orientation and a lower score indicates a more negative orientation. This approach was applied particularly to reviews with neutral ratings to determine their sentiment based on the textual content.

The labeling rules applied in this study are as follows:

1. Reviews with a rating of 4 or 5 were labeled Positive.
2. Reviews with a rating of 1 or 2 were labeled Negative.
3. Reviews with a rating of 3 were analyzed using TextBlob to obtain the polarity of the review text.

F. *Data Balancing*

Class imbalance in a dataset can cause classification models to become biased toward the majority class, leading to reduced performance, especially in identifying patterns from the minority class. To address this issue, a data balancing process is applied to ensure a more balanced class distribution. In this study, an undersampling technique is used by reducing the number of samples from the majority class (Positive) to match the number of samples in the minority class (Negative). Undersampling is chosen because the initial dataset is relatively small, making it more efficient while minimizing the risk of introducing noise compared to oversampling techniques [18], [19].

G. *Modelling*

The modeling stage in this study employs two widely used classification algorithms in sentiment analysis, namely Naive Bayes and Support Vector Machine (SVM). Before modeling, the dataset is split into training and testing sets using three different ratios: 70:30, 80:20, and 90:10. Both algorithms are selected due to their effectiveness in handling text data, particularly short and direct reviews commonly found in digital opinion datasets [4].

1. Naive Bayes

   Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem with the assumption of feature independence [5]. In this study, Naive Bayes is used to classify textual reviews from Google Maps and Google Forms into sentiment categories such as positive and negative. The model calculates the probability of a text belonging to a particular class based on the frequency of words appearing in the text.
2. Support Vector Machine (SVM)

   Support Vector Machine (SVM) is a machine learning algorithm widely used for classification and regression tasks. SVM works by identifying an optimal hyperplane in a high-dimensional space that separates data into different classes with maximum margin. The margin represents the distance between the hyperplane and the closest data points from each class. A larger margin indicates better generalization performance when applied to unseen data [8].

H. *Model Evaluation*

Model evaluation is a crucial stage in the development of machine learning systems, particularly in sentiment analysis. The main objective is to assess how well the model can classify data into the correct sentiment categories, such as positive and negative. In this study, model performance is evaluated using standard metrics, including accuracy, precision, recall, and F1-score. These evaluation metrics were used to compare the performance of Naive Bayes and SVM across the three training-testing ratios. The comparison provides a basis for determining the classification model that performs better on the balanced public opinion dataset [20], [21].

III. Results and Discussion

A. *Topic Mapping Results*

The topic mapping results show that the "Pelayanan" category is the most dominant, with a total of 292 reviews. This finding indicates that service quality is the primary concern for taxpayers when providing feedback. Other categories such as "Tempat" and "Overall" appear less frequently, suggesting that physical infrastructure is not the main focus of public evaluation. The dominance of the Service category indicates that service delivery is the primary aspect discussed in the collected public feedback. This result provides additional information beyond sentiment polarity because it identifies the specific aspect of the tax service that receives the most attention from the public.

The distribution of reviews across the three categories is visualized in Fig. 5.

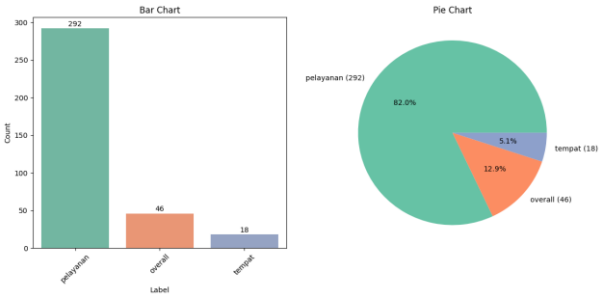


Fig. 5. Topic Mapping Chart

B. *Sentiment Labeling Results*

The sentiment labeling process reveals an imbalanced class distribution. Out of 365 reviews, 293 are classified

as positive and 72 as negative. This imbalance reflects a strong tendency toward positive sentiment in public perception of the services. The sentiment distribution is presented in Fig. 6.

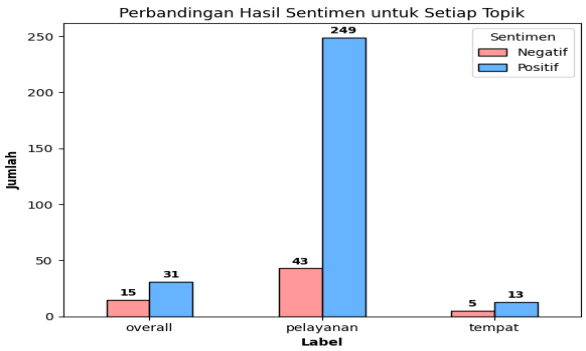


Fig. 6. Sentiment Distribution Chart

The Examples of sentiments from the collected reviews are presented in Table IV.

TABLE IV
EXAMPLE OF SENTIMENT

No.	Review	Rating	Sentiment
1	saya mengurus npwp baru dan semua prosesnya berjalan lancar petugasnya responsif dan ramah	4	Positive
2	saya agak kecewa karena harus bolak balik cuma untuk satu urusan	3	Negative

C. Data Balancing Results

To address class imbalance, an undersampling technique was applied. The majority class (Positive) is reduced to match the minority class (Negative), resulting in a balanced dataset of 120 reviews, consisting of 60 Positive and 60 Negative reviews. This balanced dataset was then used for model training and testing.

D. Model Classification Results

The performance of Naive Bayes and SVM was evaluated using accuracy, precision, recall, and F1-score. The results indicate that Naive Bayes slightly outperformed SVM across all evaluation metrics using the specified training and testing sample ratios, in Table V.

TABLE V
TRAINING AND TESTING DATA DISTRIBUTION

Ratio	Train	Test
70:30	84	36
80:20	96	24
90:10	108	12

1. Classification Using the Naive Bayes (NB) Model
The Naive Bayes model was applied to a balanced dataset and evaluated using an 80:20 train-test split. The model achieved 63.64% accuracy for the negative class and 61.54% for the positive class. To assess performance consistency, additional experiments were conducted using 70%, 80%, and 90% training splits. Among these, the 90:10 ratio yielded the best overall performance and was selected as the optimal configuration.

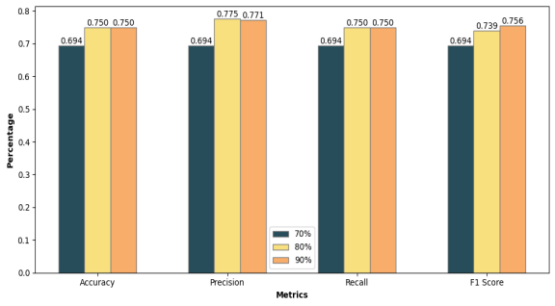


Fig. 7. Naive Bayes Model Performance Evaluation

2. Classification Using Support Vector Machine (SVM)
The second model, Support Vector Machine (SVM), followed the same processing pipeline as Naive Bayes and was evaluated on a balanced dataset using an 80:20 train-test split. The model achieved similar accuracy to Naive Bayes, with 63.64% for the negative class and 61.54% for the positive class. Further experiments with 70:30, 80:20, and 90:10 splits showed that the best performance was obtained at 70:30 with 72.22% accuracy, while performance declined at higher training proportions, indicating potential overfitting and sensitivity to data distribution.

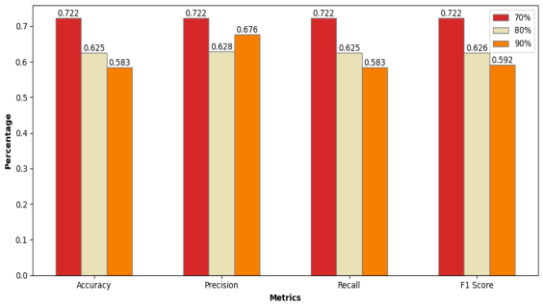


Fig. 8. SVM Model Performance Evaluation

E. Model Performance Evaluation

The performance of the two classification models, Naive Bayes and Support Vector Machine (SVM), was evaluated to compare their effectiveness in classifying sentiment. The evaluation was conducted using standard metrics, including accuracy, precision, recall, and F1-score. In addition, further analysis was performed to assess the models' ability to predict sentiment across different topics. The overall performance comparison is presented in Table VI.

TABLE VI
PERFORMANCE METRICS COMPARISON

Metric	Naive Bayes	SVM
Accuracy	0.7500	0.7222
Precision	0.7714	0.7222
Recall	0.7500	0.7222
F1-Score	0.7556	0.7222

The results indicate that Naive Bayes slightly outperforms SVM across all evaluation metrics. Naive Bayes achieves an accuracy of 75% and a precision of 77.14%, indicating that most of its positive predictions are correct. Meanwhile, SVM shows consistent performance with values of 72.22% across all metrics, demonstrating stable but slightly lower performance compared to Naive Bayes. Overall, these findings suggest that Naive Bayes is more effective for this dataset.

F. Visualization

Visualization is used to identify the main focus of public opinion regarding the services of KPP Pratama Surabaya Gubeng. One of the methods applied is a word cloud, which represents word frequency based on size. The results show that terms such as “tax”, “service”, “friendly”, and “fast” are the most dominant.

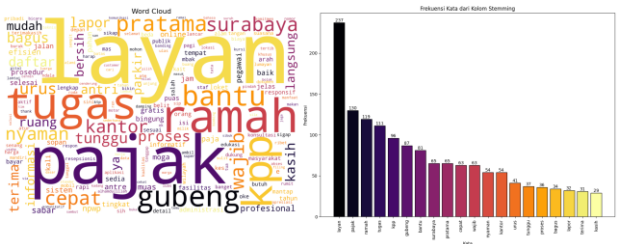


Fig. 9. Word Cloud and Word Frequency

For a more quantitative analysis, word frequency is also calculated and visualized using a bar chart. The term “service” appears most frequently (237 occurrences), followed by “tax” (130), “friendly” (119), and “staff” (111). The presence of the term “staff” indicates references to service personnel in public reviews. Overall, these visualizations confirm that service quality and interaction are the primary concerns of users. The dominance of positively connoted words such as “friendly”, “helpful”, “fast”, and “comfortable” also reflects a tendency toward positive sentiment in the dataset. The wordcloud provides an additional visualization of frequently occurring terms in the processed reviews.



Fig. 10. Word Cloud for Each Topic

IV. Conclusion

This study classified public sentiment toward the services of KPP Pratama Surabaya Gubeng using Naive Bayes and Support Vector Machine (SVM). The analysis of public feedback collected from Google Maps and Google Forms resulted in 365 valid reviews after preprocessing, consisting of 293 positive and 72 negative reviews. After undersampling, a balanced dataset of 120 reviews was obtained, consisting of 60 positive and 60 negative samples. In addition to sentiment classification, the reviews were mapped into three categories: Service, Place, and Overall. The Service category was the most frequently discussed aspect, indicating that service quality and staff interaction were prominent aspects in public feedback. This category mapping complements sentiment classification by providing information about the service aspects discussed by the public.

The classification results show that Naive Bayes outperformed SVM based on the reported evaluation metrics. Naive Bayes achieved 75.00% accuracy, 77.14% precision, 75.00% recall, and 75.56% F1-score, while SVM achieved 72.22% for the reported metrics. These results indicate that Naive Bayes was more suitable for the characteristics of the collected short-text review data.

Overall, the integration of multi-source public feedback, sentiment classification, and category mapping provides a useful approach for analyzing public perceptions of tax service quality. The findings can support the identification of service-related concerns and provide information for the evaluation and improvement of tax services at KPP Pratama Surabaya Gubeng.

Appendix

Program Repository:
https://github.com/puputayus/KP_KPP-Pratama-Surabaya-Gubeng

Acknowledgements

The authors would like to express their sincere gratitude to all parties who contributed to this research, including the respondents to the data collection and examiner for their continuous support, guidance, and encouragement throughout the research process.

References

[1] A. Ubarok, N. T. Aprilia, and S. Susanti, “Analisis kepuasan pengguna layanan Google Forms sebagai media survei online menggunakan model DeLone & McLean,” Jurnal Informatika, vol. 7, no. 2, 2020.

[2] V. Gupta and G. S. Lehal, “A survey of text mining techniques and applications,” IEEE Access, vol. 7, pp. 21687–21712, 2019.

[3] S. Poria, D. Hazarika, N. Majumder, and R. Mihalcea, “Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research,” arXiv preprint arXiv:2005.00357, 2020.

[4] H. S. Bhatti, B. Raza, S. Khalid, and J. Iqbal, “A comparative analysis of machine learning algorithms for sentiment analysis using Python,” IEEE Access, vol. 9, pp. 123612–123625, 2021.

[5] “Naive Bayes classifier,” Wikipedia, 2024. [Online]. Available: https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[6] S. Tripathy, A. Samal, and D. P. Dogra, “Sentiment analysis using rule-based classification and Min–Max normalization for social media text,” IEEE Trans. Comput. Social Systems, vol. 8, no. 2, pp. 315–326, 2021.

[7] A. Syafira, E. Prasetyo, and D. Lestari, “Sentiment analysis on public services using Naive Bayes classifier,” Journal of Data Science Applications, vol. 4, no. 2, pp. 101–108, 2022.

[8] R. Oktaviani, H. Widodo, and S. Anwar, “Comparative study of sentiment classification using SVM and Naive Bayes on e-commerce reviews,” Procedia Computer Science, vol. 179, pp. 876–882, 2021.

[9] N. Rachmawati and A. Santoso, “Optimizing sentiment classification with balanced datasets in public feedback analysis,” Indonesian Journal of Artificial Intelligence, vol. 6, no. 1, pp. 33–40, 2023.

[10] F. Lestari, R. Aditya, and A. Ramadhan, “Hybrid sentiment labeling based on rating and textual polarity for review classification,” Jurnal Teknologi dan Sistem Informasi, vol. 11, no. 1, pp. 45–53, 2023.

[11] “Google Maps,” Wikipedia, 2024. [Online]. Available: https://en.wikipedia.org/wiki/Google_Maps

[12] “Web scraping,” Wikipedia, 2024. [Online]. Available: https://en.wikipedia.org/wiki/Web_scraping

[13] A. Brenning and S. Henn, “Web scraping: A promising tool for geographic data acquisition,” arXiv preprint arXiv:2305.19893, 2023.

[14] A. F. Aufar et al., “Mengoptimalkan preprocessing teks untuk analisis sentimen yang akurat pada ulasan e-wallet,” Journal of Information and Computer Technology Education (JICTE), vol. 7, no. 2, 2023.

[15] M. Arief and M. B. M. Deris, “Text preprocessing impact for sentiment classification in product review,” in Proc. Sixth Int. Conf. Informatics and Computing (ICIC), 2021, pp. 1–6.

[16] D. P. Putra and A. Wibowo, “Text preprocessing optimization for Indonesian sentiment analysis,” Jurnal Teknologi Informasi dan Ilmu Komputer, vol. 9, no. 1, pp. 35–42, 2022.

[17] A. Kusumawardani, D. Maulidina, and D. Fitria, “Topic classification in customer reviews using rule-based and keyword matching approach,” Jurnal Teknologi Informasi dan Komputer, vol. 7, no. 2, pp. 112–119, 2022.

[18] D. Maulidya, T. Rahman, and M. Hidayatullah, “Handling class imbalance in Indonesian sentiment analysis: A comparative study,” Jurnal Informatika dan Data Sains, vol. 7, no. 1, pp. 55–64, 2023.

[19] B. Yousefimehr et al., “Data balancing strategies: A survey of resampling and augmentation methods,” arXiv preprint arXiv:2505.13518, 2025.

[20] F. Wulandari, U. Hasanah, and B. Pramudyo, “Evaluation metrics for sentiment classification in short texts,” Jurnal Ilmu Komputer dan Aplikasi, vol. 11, no. 2, pp. 65–72, 2023.

[21] D. M. W. Powers, “Evaluation: From precision, recall and F-score to ROC, informedness, markedness & correlation,” arXiv preprint arXiv:2010.16061, 2020.

Journal of Data Science, Informatics and Computer Engineering (JDSICE),
Vol. 1, Issue 1
February 2026
ISSN: XXXX-YYYY

A Spatial-Temporal Framework for Hepatitis-A Vulnerability Analysis Using Hierarchical Clustering: Case Study of Pacitan Regency, Indonesia

Arna Fariza*, Fanny Maulana Rizky, Arif Basofi
Department of Informatics and Computer Engineering, Politeknik Elektronika Negeri Surabaya, Indonesia
**Corresponding author: Arna Fariza; email: arna@pens.ac.id*

Manuscript received July 24, 2026, revised September 24, 2026, accepted September 25, 2026
doi: zz.yyyyy/xxxxx.pp134-1399

Abstract – *Hepatitis A remains a significant public health concern in Indonesia, particularly in rural and coastal regions such as Pacitan Regency, where inadequate sanitation and contaminated water sources contribute to recurring outbreaks. Traditional mitigation strategies, including public hygiene campaigns and local sanitation initiatives, often lack the spatial precision and temporal depth required to identify and prioritize high-risk areas effectively. This study proposes a spatial-temporal vulnerability mapping framework using agglomerative hierarchical clustering to classify 12 sub-districts in Pacitan Regency based on four key indicators: population density, sanitation rate, clean water availability, and health indicators. The hierarchical clustering evaluation showed that Ward consistently achieved the highest Silhouette and Calinski–Harabasz scores across the three years, while Average obtained the lowest mean Davies–Bouldin Index of 0.61, demonstrating complementary strengths across the clustering evaluation metrics. The digital interface supports informed decision-making by health authorities and raises public awareness by delivering accessible insights to residents regarding the vulnerability status of their respective districts.*

Keywords: *Agglomerative hierarchical clustering, Hepatitis-A, Pacitan Regency, spatial-temporal visualization.*

I. Introduction

Hepatitis A is an infectious disease caused by the hepatitis A virus (HAV), which primarily affects the liver and is commonly transmitted through the fecal-oral route, particularly in areas with inadequate sanitation and contaminated water supplies [1]. In Indonesia, hepatitis A remains a public health concern, especially in rural and coastal regions, where periodic outbreaks have been reported in schools and communities [2].

Pacitan Regency in East Java, Indonesia, has experienced recurring Hepatitis A outbreaks, often escalating into public health emergencies requiring immediate intervention [3]. The 2019 outbreak was particularly severe, affecting 1,310 individuals across multiple subdistricts in Pacitan [4]. Studies identified the Kaligoro River as a primary source of contamination, worsened by communal food-sharing practices and inadequate sanitation in densely populated areas [3].

Identifying vulnerable regions over time and space is crucial for effective surveillance and intervention by local health authorities, including the Public Health Services of the Pacitan Regency. Traditional efforts to mitigate these outbreaks include promoting Clean and Healthy Living Behaviors (PHBS) and localized water sanitation initiatives, which have increased public awareness but lack the precision to identify high-risk areas effectively [5]. These methods often fail to provide actionable insights for targeted interventions, limiting their long-term effectiveness [6]. However, conventional epidemiological mapping often fails to capture the dynamic nature of disease spread, as it lacks the integration of both spatial and temporal dimensions [7].

Recent advances in geospatial analysis and machine learning offer new opportunities for health vulnerability assessment. In particular, hierarchical clustering has proven effective for uncovering patterns in complex health datasets and grouping areas with similar characteristics [8][9]. When integrated with spatial-temporal visualization techniques, this approach can provide actionable insights for targeted public health responses [10].

The study employs agglomerative hierarchical

clustering, a form of unsupervised machine learning, to group 12 sub-districts based on their similarity across the four criteria, consist of population density, sanitation rate, clean water availability rate, and health indicators. These factors represent a combination of environmental, demographic, and healthcare-related variables that collectively affect transmission potential and public health resilience. By calculating Euclidean distances and applying various linkage method, we uncover natural groupings that reflect varying degrees of vulnerability. This clustering is performed on a temporal basis, allowing us to observe changes in sub-district classification across multiple years. The results are then visualized spatial-temporal maps of hepatitis A vulnerability across the regency.

The contributions of this paper include: (1) the development of a vulnerability mapping approach tailored to hepatitis A, (2) the application of agglomerative hierarchical clustering to group 12 sub-district with similar risk profiles, and (3) the spatial-temporal visualization of clustering results to aid Public Health Services of the Pacitan Regency decision-making.

II. Related Works

Understanding the dynamics of disease outbreaks and identifying regions at varying levels of risk requires robust methodologies for analyzing spatial-temporal data. Clustering techniques are widely used in epidemiology to group geographical areas based on shared characteristics, enabling policymakers to target interventions more effectively. Among these techniques, hierarchical clustering has emerged as a versatile tool due to its ability to uncover relationships within complex, multidimensional datasets [11].

The application of hierarchical clustering in public health has expanded significantly over the past decade. It is particularly suitable for epidemiological analysis because it can form nested clusters, allowing researchers to conduct multi-level examinations of disease patterns [12]. Unlike partition-based methods such as k-means, hierarchical clustering does not require predefining the number of clusters, which is beneficial for exploratory data analysis [13].

Hierarchical clustering operates by iteratively merging (agglomerative) or splitting (divisive) clusters, generating a tree-like structure known as a dendrogram, which visually represents data relationships. In the context of Hepatitis A vulnerability mapping, this technique provides a useful framework for visualizing relative risk levels across districts, making it easier to identify high-priority intervention areas. Agglomerative clustering, in particular, is favored in public health studies. It begins with each data point as an individual cluster and successively merges the closest pairs until all points belong to a single cluster [14].

The linkage criterion plays a vital role in determining how clusters are formed. Ward's method, which minimizes the variance within clusters, is widely used for its robustness and is effective in spatial epidemiology applications [15]. Average linkage, on the other hand, emphasizes internal cohesion of clusters [13], while complete and single linkage methods are useful in identifying extreme clustering scenarios—either compact or widely dispersed data groups [14].

This research adopts agglomerative hierarchical clustering with various linkage methods to comprehensively assess vulnerability patterns. Ward's linkage is utilized to produce compact clusters, average linkage for balanced grouping, and complete/single linkage to explore sensitivity to spatial outliers. Each method has demonstrated effectiveness in previous disease clustering studies, including mapping dengue fever and cholera outbreaks [16][17].

Feature selection is also essential to clustering success. Epidemiological variables such as population density, sanitation coverage, clean water availability, and health indicators are selected as clustering inputs. These factors align with prior studies emphasizing the importance of combining socio-environmental and health-related indicators for accurate disease vulnerability modelling [18][19]. To ensure compatibility across variables, standardization and normalization techniques are applied—a best practice in hierarchical clustering [20].

Finally, integrating hierarchical clustering with spatial-temporal visualization enhances the ability to detect and communicate patterns of disease vulnerability. Studies show that clustering paired with GIS-based visualization improves understanding of geographical disparities in health risks and supports evidence-based public health planning [21].

III. Proposed Methodology

This research aims to assess the risk of Hepatitis A across various districts in Pacitan by analyzing key health and environmental indicators. The system developed for this purpose employs a hierarchical clustering algorithm to classify districts based on the risk levels associated with Hepatitis A, considering factors such as population density, sanitation, water availability, and existing health conditions. The system design can be seen in Figure 1.

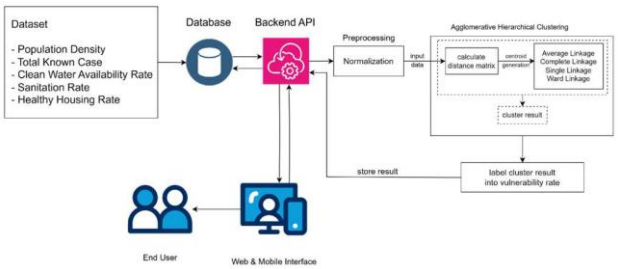


Fig. 1. System Design

The system dynamically updates as new data is entered, ensuring that the classifications remain current. These risk

classifications are visually represented in a web-based dashboard, allowing users to interact with and explore the data. The general workflow begins with the collection of relevant data from official sources, followed by preprocessing and clustering using the hierarchical clustering method. The resulting clusters categorize the districts into different risk levels—such as low, medium, or high—and these categories are represented using visualizations on a web interface for ease of interpretation.

A. Dataset Formation

The dataset used in this study was obtained from the Public Health Services of the Pacitan Regency. It comprises records from the years 2017, 2018, and 2019, detailing key indicators for 21 sub-districts. These indicators include:

- Population density, representing the number of individuals residing in a specific area;
- Sanitation rate, indicating the proportion of households with adequate sanitation facilities;
- Clean water availability rate, showing the percentage of households with access to clean water; and
- Health indicators, specifically including the number of Hepatitis A cases reported during the years in question.

This dataset is comprehensive and allows for an annual analysis of risk levels, reflecting changes in environmental and health conditions across the district.

To better understand the relationships between the features used in the clustering process, we conducted a Pearson correlation analysis. Pearson's correlation coefficient measures the linear relationship between two variables, with values ranging from -1 to 1. A coefficient of 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no linear relationship.

Pearson's correlation is particularly useful in this context as it helps identify potential multicollinearity issues, where highly correlated features might distort the clustering results, particularly for algorithms like K-means that rely on distance-based metrics. By understanding the feature correlation, we can determine whether dimensionality reduction techniques, such as Principal Component Analysis (PCA), should be applied to reduce redundancy and improve the clustering process. The Pearson correlation coefficients between the features were calculated, and a heatmap matrix was generated to visually inspect the strength and direction of relationships. This allowed us to identify pairs of features that exhibit strong correlations and to evaluate the overall structure of the feature set.

Figure 2 presents the correlation matrix among the features used in the analysis. Several key observations can be identified. The Sanitation Level and Clean Water features exhibit a strong positive correlation (r = 0.54), indicating that areas with better sanitation levels tend to have better access to clean water. Similarly, the Population Density and Healthy Housing features show a significant positive correlation (r = 0.60), suggesting that areas with higher population densities tend to have better housing conditions. In contrast, the Total Case and Population Density features show a very weak correlation (r = -0.053), as do the Clean Water and Population Density features (r = 0.0091), indicating a minimal relationship between these variables.

These observations suggest that some features are highly correlated, which could potentially affect the quality of the clustering. For instance, the strong correlation between Sanitation Level and Clean Water suggests that these features may convey similar information, potentially leading to redundancy in the data and influencing the performance of distance-based clustering methods. In contrast, weakly correlated features such as Total Case and Population Density could provide unique information for the clustering process.

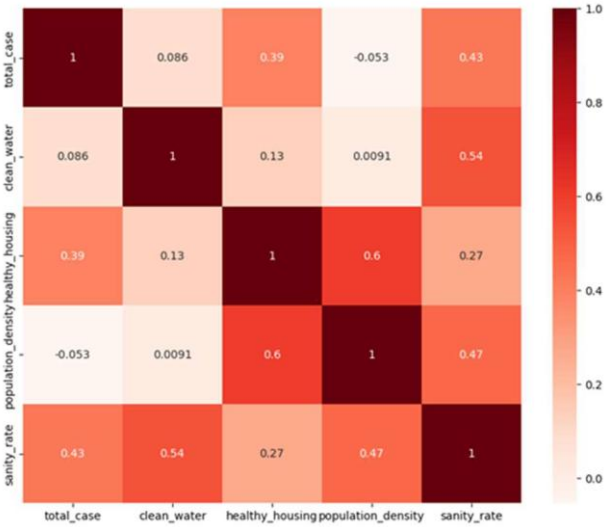


Fig. 2. Heatmap of Feature Similarity Correlation Matrix

B. Pre-processing

Before clustering, the dataset undergoes a preprocessing stage to ensure that all variables are standardized and suitable for analysis. This involves applying min-max normalization, which scales each variable to a uniform range, typically [0, 1]. Normalization is essential for balancing the influence of variables with differing units or scales, ensuring that no single indicator disproportionately affects the clustering results. The normalization process is performed using the formula:

$$X_{normalized} = \frac{X - X_{min}}{X - X_{min}} \quad (1)$$

where X represents the original value, and X_{min} and X_{max} are the minimum and maximum values of the variable in the dataset, respectively. After normalization, the dataset is ready for clustering, with all variables contributing equally to the risk classification.

C. Agglomerative Hierarchical Clustering for Hepatitis-A Vulnerability Assessment

The clustering process in this study utilizes a hierarchical clustering algorithm to group data based on similarities. The objective is to identify patterns or clusters within the dataset that can inform decision-making processes for stakeholders. This process is carried out on the preprocessed dataset, which has been normalized to ensure uniformity in data scale. The workflow of agglomerative hierarchical clustering for assessing Hepatitis-A vulnerability is as follows:

1) Calculates the distance matrix

Initially, the algorithm calculates the distance matrix, a pairwise comparison of data points based on their features. The Euclidean distance is employed as the metric to measure similarities or dissimilarities between the data points. For two data xi and xj, the Euclidean distance d is calculated as:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n} (x_{i,k} - x_{j,k})^2} \quad (2)$$

where n is the number of features, and $x_{i,k}$ and $x_{j,k}$ are the values of the k-th feature for data points x_i and x_j, respectively.

2) Iterative merging of clusters

Following the computation of the distance matrix, the hierarchical clustering algorithm is applied. This study employs the agglomerative method, where clusters are iteratively merged based on their proximity. The Various linkage criteria are explored in this research to assess their impact on cluster formation. These criteria include:

- Ward's: Minimizes the variance within clusters, leading to compact and homogeneous groupings.
- Single Linkage: Considers the shortest distance between data points in different clusters, resulting in elongated and chain-like clusters.
- Complete Linkage: Uses the largest distance between data points in different clusters, producing compact and spherical clusters.
- Average Linkage: Computes the average distance between all pairs of points from two clusters, balancing the characteristics of single and complete linkage.

Among these criteria, Ward's method is selected as the primary linkage criterion for the following reasons. First, Ward's method minimizes the total within-cluster variance at each merging step, which tends to produce clusters of relatively equal size and high internal homogeneity. This characteristic is particularly suitable for the present study, where the objective is to classify sub-districts into distinct vulnerability levels (high, medium, and low) that are internally consistent and clearly separable. Second, unlike single linkage, which is prone to chaining effects, and complete linkage, which is sensitive to outliers, Ward's method provides a more balanced and stable clustering structure, especially when the features have been normalized. Third, the variance-minimization principle of Ward's method aligns well with the composite nature of the vulnerability indicators used in this study, which combine multiple normalized variables (e.g., sanitation, clean water access, population density, and healthy housing) into a single grouping framework. To ensure robustness, however, the other linkage criteria are still computed and compared, and their resulting dendrograms are evaluated alongside the Ward-based solution

3) Dendrogram Construction

Each linkage criterion provides a unique perspective on the dataset's structure, and their application helps evaluate the robustness of clustering results. To identify the optimal number of clusters, a dendrogram is generated for each linkage method. The dendrogram visually represents the hierarchical relationships between clusters, and a predefined threshold is used to determine the final cluster count. This threshold is chosen based on domain-specific knowledge and the overall goals of the analysis. In addition to the visual inspection of the dendrogram, the optimal number of clusters is validated using the silhouette score, which measures the cohesion within clusters and the separation between clusters. The combination of dendrogram inspection and silhouette analysis ensures that the selected cluster count is both statistically and contextually justified.

The clustering results are visualized using a dendrogram, which graphically represents the merging process and provides insights into the hierarchical relationships between clusters. By setting a predefined threshold or "cut-off" on the dendrogram, the optimal number of clusters is determined. This threshold is selected based on the dataset's characteristics and the specific requirements of the research.

4) Cluster Interpretation and Visualization

Once the clusters are identified, they are labeled according to their characteristics and further analyzed to extract meaningful patterns. Assign each sub-district to a cluster representing high, medium, and low vulnerability sub-districts. Visualize the clusters on a map to show spatial distribution.

IV. Result and Discussion

In this section, evaluation of the clustering results, as well as visualization of the user interface, is performed.

A. Spatial-Temporal Hepatitis-A Vulnerability

The results of the clustering process using agglomerative hierarchical clustering produced the

vulnerability levels of Hepatitis A for each sub-district in Pacitan Regency for the years 2017–2019, as shown in Table I. Table I presents a comparison of clustering outcomes based on four linkage criteria: Ward, single, average, and complete.

The table shows that some sub-districts, such as Tulakan, Pacitan, Bandar, and Donorojo, consistently appeared in the high vulnerability cluster in earlier years but shifted to lower vulnerability levels by 2019. Conversely, sub-districts like Ngadirojo and Sudimoro demonstrated an increase in vulnerability levels over time.

The comparison across different linkage methods reveals a high degree of consistency in cluster assignments, particularly under the Ward and Average linkage methods. This consistency suggests that the clustering results are robust and reliable for identifying spatial-temporal patterns of disease vulnerability across multiple years. These findings offer valuable insight into the dynamic distribution of Hepatitis A risk, providing a foundation for more targeted interventions and resource allocation by the local health authorities.

B. Hierarchical Clustering Evaluation

Three clustering evaluation metrics—Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index—were used to assess the quality of clustering for the dataset.

1) Silhouette Coefficient

The Silhouette Coefficient was used to assess the quality of clustering by evaluating both the cohesion and separation of clusters. The score ranges from -1 to 1, where a higher value indicates that the data points are well clustered, with a clear separation between different clusters. A score near 1 indicates that the data points are well matched to their own cluster and poorly matched to neighboring clusters, while a negative value suggests that the points may have been assigned to the wrong cluster. Table II presents the Silhouette Coefficient scores for the clustering results across the years 2017, 2018, and 2019 using different clustering methods.

The Silhouette Coefficient results indicate that the Ward method consistently achieved the highest clustering quality across all three years, with an average score of 0.34, suggesting relatively cohesive and moderately well-separated clusters. This suggests that the Ward method is particularly effective in creating distinct clusters. In contrast, the Single method showed consistently lower scores, especially in 2017 and 2019, reflecting poor separation and cohesion between clusters. The Average and Complete methods exhibited moderate performance, with scores generally falling between 0.4 and 0.6. While the Complete method saw a slight improvement in 2019, it still did not outperform the Ward method in terms of clustering quality. Overall, these observations highlight that the Ward method is the most reliable for achieving well-defined clusters, while the Single method should be avoided due to its poor performance. The Average and Complete methods showed more moderate results, with the Complete method showing some improvement over time, though not enough to surpass the Ward method.

TABLE I
HIERARCHICAL CLUSTER RESULT OF HEPATITIS-A VULNERABILITY

Sub District	2017				2018				2019			
	Ward	***Single***	***Average***	***Complete***	***Ward***	***Single***	***Average***	***Complete***	***Ward***	***Single***	***Average***	***Complete***
Arjosari	Medium	Medium	Medium	Medium	Medium	Medium	Medium	Medium	Low	Low	Low	Low
Bandar	Medium	Medium	Medium	Medium	Medium	Medium	Medium	Medium	**High**	**High**	**High**	**High**
Donorojo	Medium	Medium	Medium	Medium	Medium	Medium	Medium	Medium	**High**	**High**	**High**	**High**
Kebonagung	Medium	Medium	Medium	Medium	Low	Medium	Low	Low	Medium	Medium	Medium	Medium
Nawangan	Medium	Medium	Medium	Medium	Medium	Medium	Medium	Medium	Medium	Medium	Medium	Medium
Ngadirojo	Low	Low	Low	Low	Medium	Medium	Medium	Medium	Medium	Medium	Medium	Medium
Pacitan	**High**	**High**	**High**	**High**	**High**	**High**	**High**	**High**	Low	Low	Low	Low
Pringkuku	Medium	Medium	Medium	Medium	Low	Medium	Low	Low	Medium	Medium	Medium	Medium
Punung	Medium	Medium	Medium	Medium	Medium	Medium	Medium	Medium	Medium	Medium	Medium	Medium
Sudimoro	Low	Low	Low	Low	Low	Low	Low	Low	Medium	Medium	Medium	Medium
Tegalombo	Medium	Medium	Medium	Medium	Medium	Medium	Medium	Medium	Medium	Medium	Medium	Medium
Tulakan	**High**	**High**	**High**	**High**	**High**	**High**	**High**	**High**	Low	Low	Low	Low

TABLE II
SILHOUETTE COEFFICIENT EVALUATION

Linkage Criteria	Silhoutte Coefficient Score		
	2017	*2018*	*2019*
Ward	**0.46**	0.24	0.33
Single	0.30	**0.27**	0.32
Average	0.45	**0.27**	0.24
Complete	0.45	0.21	**0.29**

2) Davies-Bouldin Index

The Davies-Bouldin Index (DBI) is a metric used to evaluate the clustering quality by calculating the ratio of within-cluster scatter to between-cluster separation. A lower DBI value indicates better clustering, as it implies that the clusters are compact and well-separated from each other. Conversely, higher values suggest poor separation and compactness, meaning the clusters are less distinct and more dispersed. Table III presents the Davies-Bouldin Index scores for the clustering results across the years 2017, 2018, and 2019 using different clustering methods.

TABLE III
DAVIES-BOULDIN INDEX EVALUATION

Linkage Criteria	Davies-Bouldin Index Score		
	2017	*2018*	*2019*
Ward	0.95	1.24	0.86
Single	**0.45**	**0.79**	0.70
Average	0.53	**0.79**	**0.51**
Complete	0.53	1.29	0.97

The Davies–Bouldin Index (DBI) results indicate that the Average linkage method achieved the best overall clustering performance across the three years, with the lowest mean DBI score of 0.61. The Average method also achieved the lowest DBI score in 2019 (0.51) and tied with the Single method in 2018 (0.79). In 2017, the Single linkage method obtained the lowest DBI score (0.45). In contrast, the Ward method yielded the highest mean DBI score of 1.02, indicating relatively less compact and less well-separated clusters compared with the other linkage methods. The Complete method produced an intermediate overall performance, with a mean DBI score of 0.93. Overall, the results suggest that the Average linkage method provided the most favorable clustering quality based on the DBI criterion.

3) Calinski-Harabasz Index

The Calinski-Harabasz Index (CHI) evaluates clustering quality by measuring the ratio of the sum of between-cluster dispersion to within-cluster dispersion. A higher CHI value indicates better clustering, as it suggests that the clusters are well-separated and distinct. In contrast, a lower CHI value implies less clear separation between the clusters, suggesting that the clustering is not as effective. Table IV presents the Calinski-Harabasz Index scores for the clustering results across the years 2017, 2018, and 2019 using different clustering methods.

TABLE IV
CALINSKI-HARABASZ INDEX EVALUATION

Linkage Criteria	Calinski-Harabasz Index Score		
	2017	*2018*	*2019*
Ward	**7.61**	3.99	**4.68**
Single	3.31	3.69	4.01
Average	7.06	3.69	2.58
Complete	7.06	**4.20**	4.35

The Calinski-Harabasz Index results indicate that the Ward method achieved the highest CHI scores across all three years, demonstrating the best clustering performance in terms of separation and compactness (with scores of 7.61 in 2017, 3.99 in 2018, and 4.681 in 2019). The Single method had the lowest CHI scores, suggesting weaker clustering quality due to poor separation between clusters (with scores of 3.31 in 2017, 3.69 in 2018, and 4.01 in 2019). The Complete and Average methods performed similarly, with the Complete method slightly outperforming the Average method in 2018 and 2019 (scores of 4.20 in 2018 vs 3.69, and 4.35 in 2019 vs 2.58). Nevertheless, none of the methods surpassed the Ward method, which consistently delivered the best results. Based on these findings, the Ward method is the most effective for generating well-separated and compact clusters, while the Single method should be avoided due to its inferior clustering results.

Overall, the hierarchical clustering evaluation reveals that the clustering performance varied across linkage criteria and evaluation metrics. The Ward method consistently achieved the highest Silhouette Coefficient and Calinski–Harabasz Index scores across the three years, indicating relatively stronger cluster cohesion and separation according to these two criteria. However, the Average linkage method obtained the lowest mean Davies–Bouldin Index score (0.61), indicating more favorable cluster compactness and separation according to the DBI criterion. The Single method generally exhibited lower Silhouette and Calinski–Harabasz scores, although it achieved the lowest DBI value in 2017 and was tied with the Average method in 2018. These findings indicate that no single linkage criterion consistently outperformed the others across all evaluation metrics. Overall, the Ward and Average methods demonstrated complementary strengths, with Ward showing stronger performance in terms of cluster separation and cohesion based on the Silhouette and Calinski–Harabasz indices, while Average provided more favorable results according to the Davies–Bouldin Index.

C. User Interface

The system provides two user-facing interfaces: a web-based dashboard for administrators and a mobile

application for residents of Pacitan Regency. Both interfaces are built upon a three-tier architecture consisting of a data layer, a processing layer, and a presentation layer. The data layer manages the preprocessed dataset and spatial boundary files, the processing layer executes the analytical pipeline (correlation analysis, agglomerative hierarchical clustering with Ward's linkage, and cluster labeling), and the presentation layer delivers the visualization and interaction components to end users.

The web-based interface is designed for administrators to manage the dataset, perform re-clustering, and inspect the clustering results. Its main functions include: (i) uploading and editing datasets, (ii) triggering the re-clustering process, (iii) visualizing clustering results on an interactive map, and (iv) exporting the resulting vulnerability classification. As shown in Figure 3, administrators can create and edit datasets through a structured form, while Figure 4 presents the dashboard that displays clustering outcomes and vulnerability levels. Figure 5 further illustrates the spatial visualization of clustering results, highlighting areas with different vulnerability levels, and Figure 6 presents a graph showing clustering results for each district across all years simultaneously, providing a comprehensive temporal view of vulnerability.

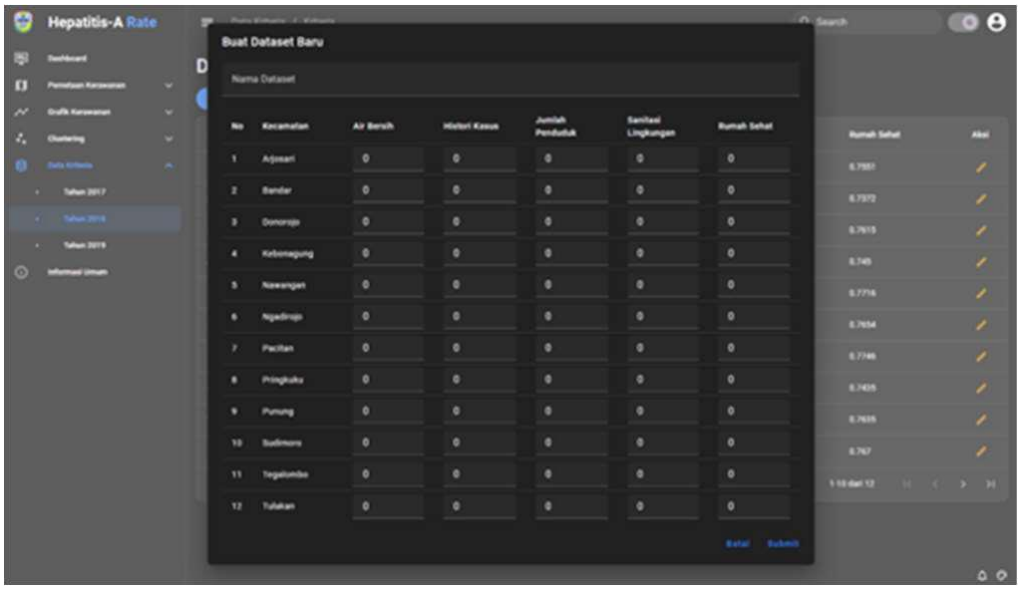


Fig. 3. Web interface for creating and editing datasets

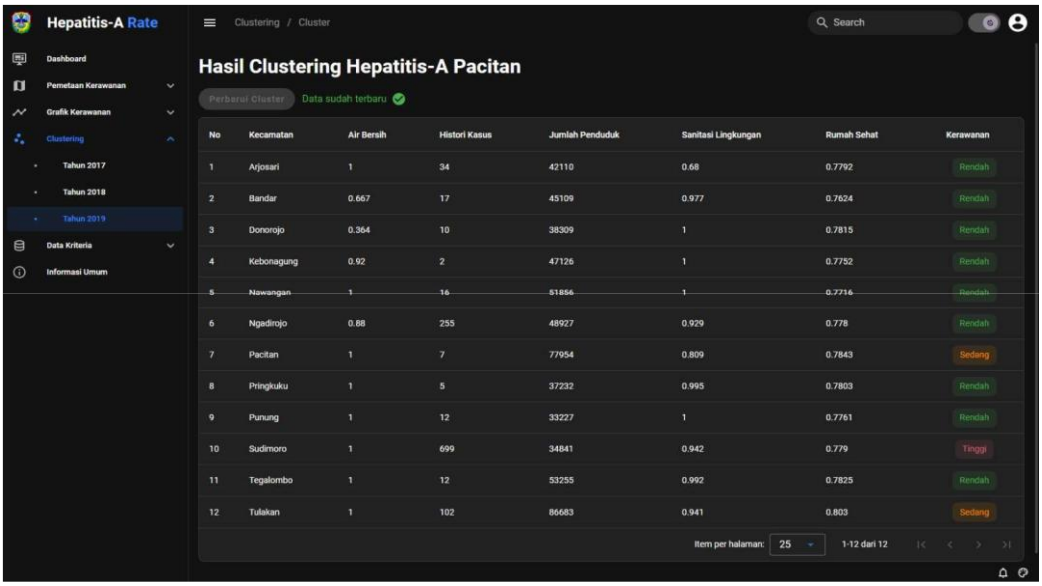


Fig. 4. Web dashboard displaying clustering results and vulnerability levels

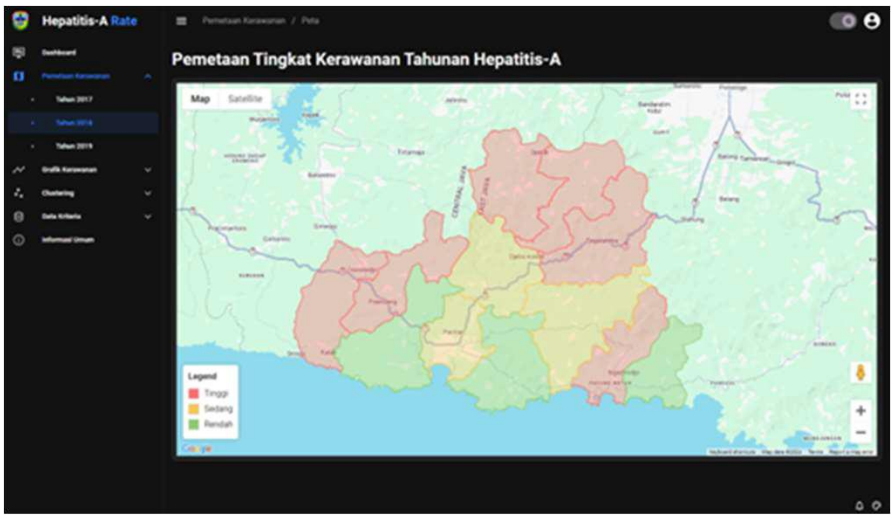


Fig. 5. Web interface showing the clustering results with vulnerability levels in each area

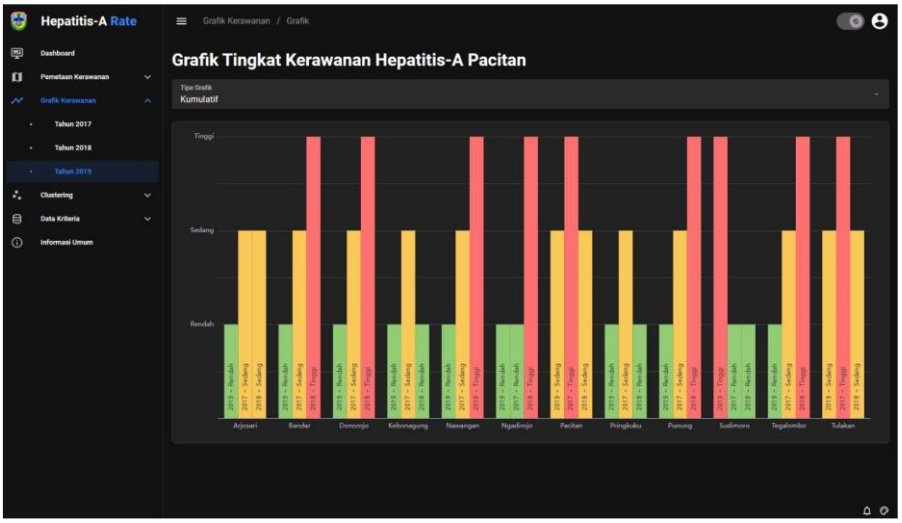


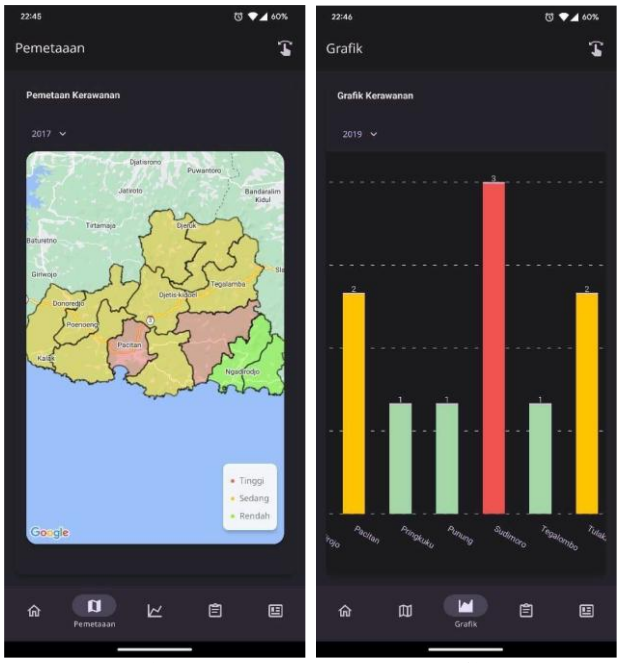
Fig. 6. Graph displaying the clustering results for each district across all years

The mobile application targets residents of Pacitan Regency and is designed to communicate vulnerability status in an accessible format. Its main functions include: (i) displaying the vulnerability level of the user's area based on the latest clustering results, (ii) providing temporal graphs of clustering outcomes per district per year, and (iii) delivering real-time updates through a simplified, non-technical interface. Figure 7(a) shows the mobile interface displaying area vulnerability data, while Figure 7(b) presents the per-year clustering graph for each district, allowing residents to track changes in vulnerability over time.

V. Conclusion

This study developed a spatial-temporal framework for assessing Hepatitis A vulnerability across 12 sub-districts in Pacitan Regency using agglomerative hierarchical clustering over three consecutive years (2017–2019). The results demonstrate that vulnerability patterns varied across sub-districts and years, indicating the dynamic nature of Hepatitis A vulnerability and the importance of continuous spatial-temporal surveillance. Four linkage criteria—Ward, Single, Average, and Complete—were evaluated using the Silhouette Coefficient, Davies–Bouldin Index, and Calinski–Harabasz Index. The results indicate that Ward generally provided stronger cluster cohesion and separation based on the Silhouette and Calinski–Harabasz indices, whereas Average achieved a more favorable mean Davies–Bouldin Index of 0.61. These findings demonstrate that clustering performance depends on the evaluation criterion and that no single linkage method consistently outperformed the others across all metrics. The study further developed a web-based administrative dashboard and mobile application interface to provide accessible visualizations of Hepatitis A vulnerability for health administrators and the public. These digital tools can facilitate the interpretation and dissemination of spatial-temporal vulnerability information to support epidemiological surveillance and public awareness. Future research should incorporate additional real-time and contextual risk factors, investigate predictive models for vulnerability

forecasting, and evaluate the framework across other regions to assess its generalizability and applicability.



(a) (b)

Fig. 7. (a) Mobile app interface displaying area vulnerability data; (b) Mobile app graph showing the clustering results for each district per year

Acknowledgements

This work was supported by Politeknik Elektronika Negeri Surabaya.

References

[1] M. Migueres, S. Lhomme, J. Izopet, "Hepatitis A: epidemiology, high-risk groups, prevention and research on antiviral treatment," Viruses, vol. 13 no. 10, pp. 1900, 2021.

[2] Indonesian Ministry of Health, "Profil Kesehatan Indonesia Tahun 2022," Jakarta, Indonesia: Kemenkes RI, 2023. [Online]. Available: https://pusdatin.kemkes.go.id

[3] S. Retyono, S. S. Rahardjo, B. Murti, "Multilevel analysis: Village does not have contextual effect on hepatitis A outbreak in Pacitan, East Java, Indonesia," Journal of Epidemiology and Public Health, vol. 5 no. 2, pp. 195-206, 2020.

[4] R. Lukmitarani, M. Ardiyah, R. Azizah, "The outbreak of hepatitis A: Evidence in Pacitan District," Journal Kesehatan Lingkungan, vol. 13 no. 1, pp. 35, 2021.

[5] J. Rosenthal, R. E. Arku, J. Baumgartner, J. Brown, T. Clasen, J. N. Eisenberg, P. Hovmand, P. Jagger, D.A. Luke, A Quinn, G.N. Yadama, "Systems science approaches for global environmental health research: enhancing intervention design and implementation for household air pollution (HAP) and water, sanitation, and hygiene (WASH) program,". Environmental Health Perspectives, vol. 128 no. 10, pp. 105001, 2020.

[6] E. Hamunyela, M. Hipondoka, F. Persendt, H. S. Nghiyalwa, C. Thomas, K. Matengu, "Spatio-temporal characterization of surface water dynamics with Landsat in endorheic Cuvelai-Etosha Basin (1990–2021)," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 191, pp. 68-84, 2022.

[7] Y. Wang, X. Chen, X., F. Xue, "A review of Bayesian spatiotemporal models in spatial epidemiology," ISPRS International Journal of Geo-Information, vol. 13 no. 3, pp. 97, 2024.

[8] T. J. Loftus, B. Shickel, J. A. Balch, P. J. Tighe, K. L. B. Abbott, Fazzone, E.M. Anderson, J. Rozowsky, T. Ozrazgat-Baslanti, Y. Ren, S.A. Berceli, "Phenotype clustering in health care: a narrative review for clinicians," Frontiers in artificial intelligence, vol. 5, pp. 842306, 2022.

[9] Q. An, S. Rahman, J. Zhou, J. J. Kang, "A comprehensive review on machine learning in healthcare industry: classification, restrictions, opportunities and challenges," Sensors, vol. 23 no. 9, pp. 4178, 2023.

[10] D. Kim, B. Cánovas-Segura, M. Campos, J. M. Juarez, "Visualization of Spatial–Temporal Epidemiological Data: A Scoping Review," Technologies, vol. 12 no. 3, pp. 31, 2024.

[11] T. Alasalı, Y. Ortakcı, "Clustering Techniques in Data Mining: A Survey of Methods, Challenges, and Applications," Computer Science, vol. 9 no. 1, pp. 32-50, 2024.

[12] Z. Khorrami, M. Rezapour, K. Etemad, S. Yarahmadi, S. Khodakarim, A. Mahdavi Hezaveh, M. Kameli, N. Khanjani, "The patterns of non-communicable disease multimorbidity in Iran: a multilevel analysis," Scientific Reports, vol. 10 no. 1, pp. 3034, 2020.

[13] X. Ran, Y. Xi, Y. Lu, X. Wang, Z. Lu, "Comprehensive survey on hierarchical clustering algorithms and the recent developments," Artificial Intelligence Review, vol. 56 no. 8, pp. 8219-8264, 2023.

[14] S. Modak, "Finding groups in data: an introduction to cluster analysis: authored by Leonard Kaufman and Peter J. Rousseeuw," John Wiley and Sons, 2005, ISBN: 0-47-1-73578-7, 2024.

[15] N. Randriamihamison, N. Vialaneix, P. Neuvial, "Applicability and interpretability of Ward's hierarchical agglomerative clustering with or without contiguity constraints," Journal of Classification, vol 38 no. 2, pp. 363-389, 2021.

[16] H. Al-Nefaie, A. Alsultan, R. Abusaris, "Temporal and spatial patterns of dengue geographical distribution in Jeddah, Saudi Arabia," Journal of Infection and Public Health, vol. 15 no. 9, pp. 1025-1035, 2022.

[17] L. M. Mageto, N. Mutono, G. Aboge, P. Gathura, E. Okunga, A. Muange, C. Mbae, S.M. Thumbi, S. Kariuki, "Spatial and temporal analysis of cholera cases in high-risk areas of Kenya. "medRxiv, pp. 2023-10, 2023.

[18] D. S. S. Rejeki, N. Nurhayati, B. Aji, "A spatiotemporal analysis of dengue hemorrhagic fever in Banyumas, Indonesia," Int J Public Health Sci, vol. 10 no. 2, pp. 231, 2021.

[19] M. Widawati, P. W. Dhewantara, R. Anasi, T. Wahono, R. Marina, I. P. Pertiwi, A.A. Wibowo, A. Ruliansyah, M.U. Riandi, D. Widiastuti, E.P. Astuti, "An investigation of geographical clusters of leptospirosis during the outbreak in Pangandaran, West Java, Indonesia," Geospatial health, vol. 18 no. 2, 2023.

[20] G. James, D. Witten, T. Hastie, R. Tibshirani, J. Taylor, "Statistical learning. In An introduction to statistical learning: With applications in Python (pp. 15-67)," Cham: Springer International Publishing, 2023.

[21] B. Oinam, V. Anand, R. K. Kajal, "A spatiotemporal geographic information system-based assessment of human immunodeficiency virus/acquired immune deficiency syndrome distribution in Manipur, India," Indian Journal of Public Health, vol. 65 no. 4, pp. 362-368, 2021..

Journal of Data Science, Informatics and Computer Engineering (JDSICE),
Vol. 1, Issue 1
February 2026
ISSN: XXXX-YYYY

Explainable AI for Multi-Label Classification of Doctor–Patient Communication Using Transformer-Based Models

Selvia Ferdiana Kusuma[1]*, Mohammad Farid Naufal[2], Rizal Furqan Ramadhan[3]
[1]*Politeknik Elektronika Negeri Surabaya, Indonesia*
[2]*Universitas Surabaya, Indonesia*
[3]*Universitas Islam Negeri Sayyid Ali Rahmatullah, Indonesia*
**Corresponding author: First Author; email: selvia@pens.ac.id*

Manuscript received April 19, 2026, revised September 24, 2026, accepted September 25, 2026
doi: zz.yyyyy/xxxxx.pp134-1399

Abstract – *Effective doctor–patient communication plays a critical role in clinical outcomes, yet analyzing and interpreting such communication remains a challenging task. This study presents an explainable artificial intelligence (XAI) approach for the multi-label classification of doctor responses into six categories: Fostering the Relationship (FR), Gathering Information (GI), Providing Information (PI), Decision Making (DM), Enabling Disease- and Treatment-Related Behaviour (EDTRB), and Responding to Emotions (RE). A transformer-based model, specifically IndoBERT, is employed to perform text classification on a labeled dataset of clinical responses. To improve transparency and interpretability, explainable AI techniques such as SHAP and LIME are employed to identify the words that most strongly influence each prediction. Experimental results show that the proposed model achieves an overall accuracy of 85% in classifying doctor responses across multiple categories, while the XAI analysis provides meaningful insights into linguistic patterns associated with each communication type. The findings highlight the potential of combining deep learning and XAI to support the analysis of clinical communication, offering both high predictive accuracy and interpretable outputs that can assist healthcare professionals in evaluating and improving patient interaction quality.*

Keywords: *classification, explainable AI, patient communication, transformer-based model*

I. Introduction

Effective communication between doctors and patients is a fundamental component of high-quality healthcare services. It influences diagnostic accuracy, treatment adherence, and patient satisfaction. Different types of doctor responses, such as providing information, gathering patient data, or addressing emotional concerns, play distinct roles in clinical interactions [1]. Therefore, analyzing and categorizing these communication patterns is essential for understanding and improving healthcare delivery.

However, analyzing doctor–patient communication remains a challenging task [2]. Traditional approaches rely on manual annotation and qualitative analysis, which are time-consuming and prone to subjectivity. Recently, machine learning and deep learning techniques have been applied to automatically classify medical text [2]. Despite their high performance, these models often function as black boxes, providing little insight into how predictions are made. This lack of interpretability limits their adoption in sensitive domains such as healthcare.

Several studies have explored text classification in medical contexts and applied explainable artificial intelligence (XAI) techniques to improve model transparency [3]. However, most existing works focus on disease classification or clinical notes rather than structured communication categories in doctor responses. Additionally, limited research has addressed explainability in the context of Indonesian medical text, particularly using transformer-based models.

To address these challenges, this study proposes an explainable AI framework for classifying doctor responses into six communication categories: Fostering the Relationship (FR), Gathering Information (GI), Providing Information (PI), Decision Making (DM), Enabling Disease- and Treatment-Related Behaviour (EDTRB), and Responding to Emotions (RE). A transformer-based model, IndoBERT, is utilized to capture rich contextual and semantic representations of doctor responses. To improve transparency and interpretability, explainable AI techniques such as Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic

Explanations (LIME), are applied. SHAP provides a global and consistent explanation by quantifying the contribution of each word to the prediction based on Shapley values, while LIME offers local, instance-level explanations by highlighting the most influential words for a specific prediction. The combination of these methods enables a more comprehensive understanding of how the model makes decisions.

The main contributions of this study are as follows: (1) the development of a multi-label classification model for doctor–patient communication in the Indonesian context, (2) the integration of XAI techniques to enhance model interpretability, and (3) the extraction of meaningful linguistic patterns that characterize different types of clinical communication. These contributions aim to support both automated analysis and practical understanding of doctor–patient interactions.

II. Literature Study

This section reviews relevant studies related to doctor–patient communication classification, transformer-based language models, and XAI, which form the foundation of this research.

A. Doctor–Patient Communication Classification

Doctor–patient communication is a critical component of healthcare delivery, as it directly affects diagnostic accuracy, patient adherence, and overall satisfaction with medical services. In clinical practice, doctor responses can be categorized into several communication types, including Fostering the Relationship (FR), Gathering Information (GI), Providing Information (PI), Decision Making (DM), Enabling Disease- and Treatment-Related Behaviour (EDTRB), and Responding to Emotions (RE) [4]. Early approaches to analyzing clinical communication primarily relied on manual observation and expert annotation, which are time-consuming and subject to human bias. To overcome these limitations, recent studies have adopted Natural Language Processing (NLP) techniques to automate the classification of medical communication text.

Several studies have applied traditional machine learning methods such as Support Vector Machine (SVM) and Naïve Bayes, as well as deep learning approaches like BERT, to classify medical text [5] [6]. However, these models often struggle to capture complex contextual relationships within conversational data, particularly in doctor–patient interactions where context plays a crucial role.

B. Transformer-Based Models and IndoBERT

The development of transformer-based models, particularly BERT, has significantly advanced the field of NLP. BERT captures bidirectional context, allowing it to understand the meaning of words based on both preceding and following text, making it more effective than earlier models. For the Indonesian language, IndoBERT has been developed as a domain-adapted version of BERT trained on large-scale Indonesian corpora [7]. IndoBERT has demonstrated strong performance across various NLP tasks, including text classification, sentiment analysis, and medical text processing [8].

The key advantage of IndoBERT lies in its ability to better capture the linguistic characteristics of Indonesian compared to multilingual models [9][10]. Therefore, it is well-suited for classifying doctor responses in this study, where the data consists of context-rich Indonesian clinical communication.

C. Explainable Artificial Intelligence (XAI)

Although deep learning models such as BERT and IndoBERT achieve high performance, they are often considered black-box models due to their lack of interpretability. In the healthcare domain, model transparency is essential to ensure trust and support clinical decision-making. XAI aims to address this limitation by providing insights into how models make predictions [11] . Two widely used XAI techniques are SHAP and LIME [12] [13]. SHAP computes feature importance using Shapley values for both overall and instance-level insights, while LIME approximates the model locally with an interpretable surrogate [14].

Applying XAI to medical text analysis enables researchers to identify the most influential words or phrases contributing to classification outcomes [15], thereby providing deeper insights into communication patterns in clinical interactions.

The integration of deep learning and XAI has become an important trend in recent research, particularly in healthcare applications [16]. Several studies have combined transformer-based models with XAI techniques to improve interpretability without sacrificing performance. In medical text analysis, this integration allows not only accurate automated classification but also a better understanding of the reasoning behind model predictions. This is especially important in clinical contexts, where interpretability plays a key role in gaining acceptance among healthcare professionals.

However, studies specifically focusing on the classification of doctor communication using XAI remain limited, particularly for Indonesian-language data. Therefore, this study contributes by integrating IndoBERT with XAI techniques to develop a model that is both accurate and interpretable for analyzing clinical communication.

III. Method

This study proposes an explainable artificial intelligence (XAI) framework to interpret and understand the outcomes of doctor response classification into six

communication categories. The XAI pipeline is illustrated in Figure 1. The raw text is first preprocessed and tokenized using the IndoBERT tokenizer, then encoded into numerical representations with padding and attention masks. These inputs are processed by a fine-tuned IndoBERT model to classify the text into six communication categories. To improve interpretability, SHAP and LIME are applied to identify the most influential words behind each prediction. The results are further analyzed to highlight important tokens and provide insights into doctor communication patterns.

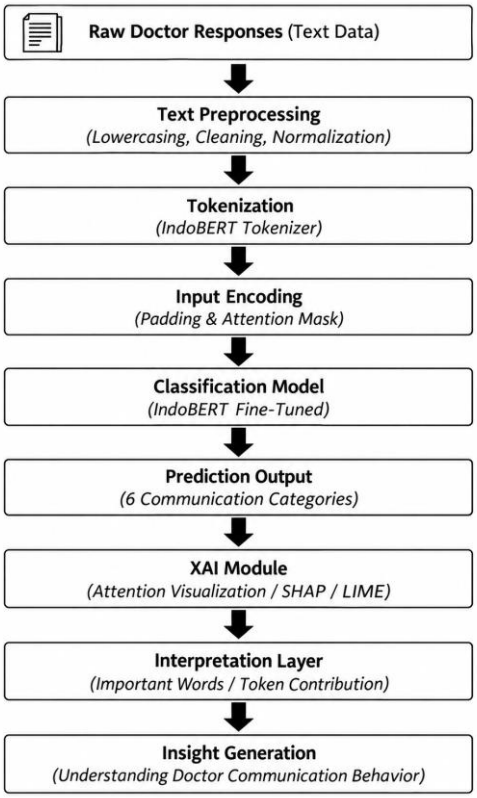


Fig. 1. XAI Pipeline for Interpreting Doctor Response Classification

A. Dataset Collection

The dataset used in this study consists of doctor responses collected from http://dx.doi.org/10.17632/p8d5bynh3m.1 . Each response is manually annotated into one of six predefined communication categories, namely Fostering the Relationship (FR), Gathering Information (GI), Providing Information (PI), Decision Making (DM), Enabling Disease- and Treatment-Related Behaviour (EDTRB), and Responding to Emotions (RE). Table 1 illustrates a sample of the annotated dataset, where each response has been classified into the predefined communication categories.

TABLE I
ANNOTATED DATASET

Sentence	Classification
Halo Rizal,Radang tenggorokan umunya disebabkan karena infeksi. Infeksi tersebut dapat oleh virus ataupun bakteri. Jika karena bakteri umumnya memerlukan antibiotika. Jika karena virus maka tidak perlu antibiotika dan hanya memerlukan pengobatan suportif untuk mennjang kekebalan tubuh.Penanganan secara umum : Pastikan anda cukup istirahat Jangan makan makanan berminyak, gorengan, pedas Makan makanan bergizi Wajib minum banyak air putih Berkumurlah dengan air garam (cukup 1/3-1/2 sdk teh untuk 200 cc air) tiap 3-4 jam sekali Nyeri dan demam dapat dikurangi dengan parasetamol atau ibuprofen Jika diperlukan dokter dapat meresepkan anda dengan antibiotika dan antiradang Obat yang anda sampaikan sudah merupakan obat yang umum dipakai pada radang tenggorokan. Amoksisilin merupakan antibiotika, hanya bermanfaat pada kasus yang disebabkan oleh infeksi bakteri, penambahannya umumnya tidak berbahaya kecuali jika nada alergi dengan amoksisilin tetapi sebaiknya konsultasikan dulu dengan dokter anda karena penggunaan antibiotik sembarangan akan memicu terjadinya resistensi/kekebalan kuman terhadap antibiotik. sakit tenggorokan demikian yang dapat saya sampaikan ,semoga bermanfaat.	3-PI 4-DM 5-EDTRB
Hai, Terimakasih telah bertanya ke Alodokter. Sebelumnya, perlu diperjelas, berapa besar ukuran benjolan yang Anda maksud? Apakah konsistensinya sangat keras? Sudah berapa lama munculnya? Apakah mudah atau sulit digerakkan? Benjolan yang keras di perut bagian bawah mungkin saja merupakan hal yang normal, misalnya karena tonjolan tulang simfisis (tulang panggul bagian depan). Benjolan ini bisa lebih jelas teraba pada wanita, orang yang memiliki panggul besar, atau postur tubuh yang kurus. Benjolan ini bukanlah sesuatu yang berbahaya yang perlu diobati. Selama tidak menimbulkan keluhan, seperti nyeri, gatal, gangguan menstruasi, gangguan buang air kecil atau besar, maka sebaiknya Anda tidak perlu khawatir ya.. Selain itu, beberapa kemungkinan lain penyebab benjolan di perut bagian bawah: Infeksi, misalnya abses Hernia Retensi urin Tumor kandung kemih Tumor saluran cerna Tumor kulit, misalnya lipoma, fibroma, sarkoma Tumor tulang , dan sebagainya Jika benjolan yang muncul terasa sangat mengganggu Anda, sebaiknya Anda periksakan langsung keluhan Anda ke dokter ya.. Umumnya, dokter akan melakukan pemeriksaan fisik dan dibantu dengan beberapa pemeriksaan sebagai penunjang, misalnya rontgen atau USG (ultrasonografi). Selanjutnya, barulah penanganan yang tepat bisa ditentukan. Sebagai langkah awal: Hindari memencet, mengurut, atau memanipulasi benjolan secara paksa Gunakan pakaian yang nyaman dan tidak terlalu ketat Jaga sanitasi lingkungan Jaga berat badan ideal Semoga membantu ya.. dr. Nadia Nurotul Fuadah	1-FR 2-GI 3-PI 4-DM 5-EDTRB 6-RE

The dataset comprises 500 responses with a multi-label annotation scheme, in which a single response can be assigned to multiple communication categories. As a result, the total number of labels exceeds the number of responses. The distribution of labels across the six communication categories is as follows: Fostering the Relationship (FR) = 500, Gathering Information (GI) = 41, Providing Information (PI) = 500, Decision Making (DM)

= 439, Enabling Disease- and Treatment-Related Behaviour (EDTRB) = 495, and Responding to Emotions (RE) = 24. The dataset is randomly split into 80% training data and 20% testing data, with random_state=42 ensuring reproducible results.

B. Text Processing

Before feeding the data into the model, several preprocessing steps were performed to improve text quality [17]. All text was first converted to lowercase to ensure consistency across samples. Unnecessary symbols and punctuation were removed to reduce noise in the data. Informal words were normalized where applicable to standardize variations in language usage. The cleaned text was then tokenized using the IndoBERT tokenizer, which converts the input into a format suitable for the transformer model. These preprocessing steps ensure that the input text is standardized and compatible with the transformer-based architecture.

C. Classification Model Using IndoBERT

The classification task is performed using IndoBERT, a pre-trained transformer model specifically designed for Indonesian language processing [18]. IndoBERT processes tokenized input and produces contextual representations for each token. After obtaining the text representation from the [CLS] token, the model applies a dropout layer to reduce overfitting. The resulting representation is then passed to a linear classification layer that produces six logits, representing the predicted probability scores for each communication category. Because a response may belong to multiple categories simultaneously, the model uses a sigmoid-based multi-label approach with BCEWithLogitsLoss. This loss function compares the predicted logits with the true label vector and calculates the training error for each category independently. Performance is assessed using standard metrics such as accuracy, precision, recall, and F1-score [18]. The evaluation metrics used in this study defined in Equations (1)-(4). True Positive (TP) indicates correctly identified positive cases, while True Negative (TN) refers to correctly identified negative cases. False Positive (FP) occurs when negative instances are wrongly predicted as positive, whereas False Negative (FN) arises when positive instances are incorrectly predicted as negative.

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

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

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

$$F1 = \{2TP\}/\{2TP + FP + FN\} \quad (4)$$

D. Explainable Artificial Intelligence (XAI)

To enhance model interpretability, explainable artificial intelligence (XAI) techniques are applied to analyze the classifications produced by the IndoBERT model. In this study, both Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) are utilized to provide complementary insights into the model's decision-making process.

LIME is employed to generate local explanations by approximating the model behavior around individual predictions, highlighting the most influential words that contribute to a specific classification outcome. Meanwhile, SHAP is used to quantify the contribution of each token to the prediction by computing feature importance values based on Shapley theory. SHAP enables both local and global interpretability, where local explanations identify important words for individual predictions, and global explanations reveal overall patterns by aggregating feature importance across the dataset [19].

Through this combined approach, it becomes possible to better understand the linguistic features associated with each communication category. For instance, expressions of empathy may strongly contribute to the Fostering the Relationship (FR) category, while directive or instructional phrases may be more dominant in Enabling Disease- and Treatment-Related Behaviour (EDTRB) responses.

E. Evaluation XAI

In addition to qualitative analysis, the explainability method was quantitatively evaluated using a faithfulness test [20]. The most important tokens identified by both SHAP and LIME were progressively removed or masked from the input text, and the corresponding changes in prediction confidence were observed. A substantial decrease in prediction confidence or a shift in the predicted class indicates a high level of faithfulness, confirming that the explanations accurately capture the model's decision-making process. In contrast, negligible changes suggest limited explanatory reliability.

IV. Result and Discussion

The following section presents the results and evaluates the model's performance and interpretability using both numerical metrics and XAI-driven qualitative analysis.

A. Classification Performance

The effectiveness of the IndoBERT-based approach was measured using established multi-label classification metrics including accuracy, precision, recall, and F1-score. The classification results are presented in Figure 2.

```
Micro F1   : 0.9784537389100126
Macro F1   : 0.8203915293413062
Samples F1 : 0.9764285714285714
Accuracy   : 0.85
HammingLoss: 0.028333333333333332

Classification Report:

              precision    recall  f1-score   support

        1-FR       1.00      1.00      1.00       100
        2-GI       1.00      1.00      1.00         4
        3-PI       1.00      1.00      1.00       100
        4-DM       0.89      0.97      0.93        86
     5-EDTRB       0.99      1.00      0.99        99
        6-RE       0.00      0.00      0.00         3

   micro avg       0.97      0.98      0.98       392
   macro avg       0.81      0.83      0.82       392
weighted avg       0.97      0.98      0.98       392
 samples avg       0.97      0.98      0.98       392
```

Fig. 2. Classification Result

The model achieved an accuracy of 0.85, with a micro F1-score of 0.978, macro F1-score of 0.820, and samples F1-score of 0.976. Additionally, a low Hamming loss of 0.028 indicates a relatively small proportion of misclassified labels. These results demonstrate that the model performs well overall, particularly in capturing label-wise predictions in a multi-label setting.

The detailed classification report shows that the model achieved perfect performance (F1-score = 1.00) for the Fostering the Relationship (FR), Gathering Information (GI), and Providing Information (PI) categories, indicating that these classes have clear and distinguishable linguistic patterns. Similarly, the Enabling Disease- and Treatment-Related Behaviour (EDTRB) class achieved a very high F1-score of 0.99, reflecting strong model capability in identifying directive and instructional language.

However, relatively lower performance was observed in the Decision Making (DM) category, with an F1-score of 0.93, suggesting some degree of overlap with other categories such as Providing Information (PI). The most challenging category is Responding to Emotions (RE), which achieved an F1-score of 0.00. This result is likely due to the extremely limited number of samples in this class, leading to difficulties in model generalization.

Overall, the results indicate that the transformer-based architecture effectively captures contextual and semantic information in doctor–patient communication. However, performance variations across categories highlight the impact of class imbalance and semantic overlap, particularly in low-resource classes such as RE and context-dependent categories such as DM.

B. XAI Analysis

To enhance model interpretability, explainable artificial intelligence (XAI) techniques were applied to analyze the contribution of individual words to the model's predictions. In this study, SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) were employed to provide complementary perspectives on model behavior.

Figure 3 presents the SHAP force plot for a sample prediction classified as Fostering the Relationship (FR) with a probability of 0.9925. The visualization illustrates the contribution of individual tokens to the model's prediction. Tokens such as Saya, Anda, and kekhawatiran strongly push the prediction toward the FR class, indicating that the model effectively captures interpersonal and empathetic expressions. Interestingly, clinically oriented tokens such as dokter, penyebab, and pastinya also contribute positively, suggesting that the model incorporates both relational and informational cues in its decision-making process.

Furthermore, the high base value indicates that the model has a prior tendency toward the FR class, which may be influenced by class distribution in the dataset. This result highlights the overlapping nature of communication categories, where a single response may simultaneously express empathy and provide medical guidance. Overall, the SHAP analysis provides deeper insight into how the model integrates linguistic and contextual features when classifying doctor responses.



Fig. 3. Shap XAI Result

Figure 4 presents the local explanation generated by LIME for a sample prediction classified as Fostering the Relationship (FR) with a high confidence score of 0.9925. The explanation highlights several influential tokens that contribute positively to the classification decision, including "*Terima, saya, mengerti, kekhawatiran, anda*". These tokens reflect empathetic and interpersonal communication patterns, indicating that the model successfully captures linguistic features associated with building rapport and acknowledging patient concerns.

In particular, phrases such as "*Terima kasih"* and "*saya mengerti kekhawatiran anda*" represent strong indicators of empathetic engagement, which are characteristic of the FR category. This suggests that the model is capable of identifying subtle relational cues in doctor–patient communication, aligning well with the intended definition of the category."

However, the explanation also reveals that tokens such as *"dokter, penyebab, sebaiknya"* and *"diketahui"* contribute significantly to the prediction, despite being more closely associated with informational or directive communication. These terms are typically indicative of categories such as Providing Information (PI) or Enabling Disease and Treatment Related Behaviour (EDTRB). Their strong contribution to the FR prediction suggests that the model does not rely solely on explicit empathetic expressions but instead considers the overall contextual structure of the response.

This observation highlights the inherently overlapping nature of doctor communication, where a single response

often serves multiple communicative functions, such as expressing empathy while simultaneously providing medical advice. As a result, certain tokens may contribute to multiple categories depending on context, which explains their influence in the FR classification.

Furthermore, the prominence of clinically oriented terms such as "*dokter"* and "*penyebab"* may indicate that the model has learned statistical associations from the training data, rather than purely semantic distinctions between categories. This suggests a potential bias toward frequently co-occurring terms, which may affect the interpretability of the model in cases where category boundaries are ambiguous.

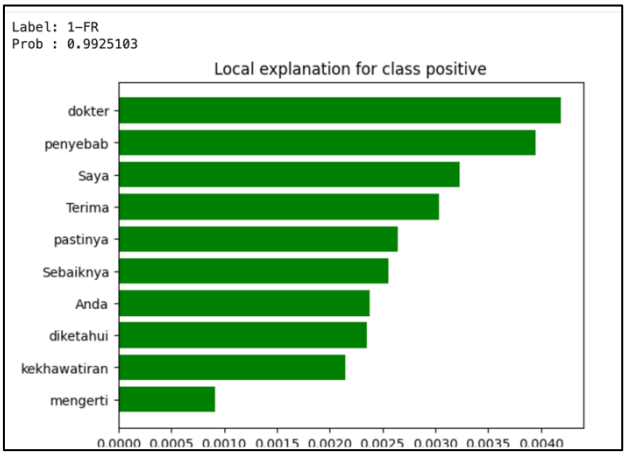


Fig. 4. Lime XAI Result

In terms of interpretability, LIME provides more intuitive and straightforward visualizations, making it easier to identify the most influential tokens contributing to a specific prediction. In contrast, SHAP offers a more comprehensive explanation by incorporating both the magnitude and direction of feature contributions relative to a baseline value, although it is relatively more complex to interpret. Therefore, LIME is more suitable for rapid and user-friendly interpretation, while SHAP provides a more theoretically grounded and consistent explanation of the model's behavior.

To assess the performance of the XAI methods, a faithfulness evaluation was conducted to determine whether the identified important features truly reflect the model's decision-making process. The faithfulness evaluation was conducted by masking the top-k important tokens identified by SHAP and LIME and measuring the resulting decrease in prediction confidence. The results indicate that removing these tokens leads to a significant drop in confidence and, in some cases, changes the predicted label, confirming that the identified features are truly influential. SHAP provides more consistent and distributed explanations, while LIME produces more focused local explanations, often resulting in sharper confidence drops. These findings demonstrate that both methods effectively capture the model's decision-making process and offer complementary interpretability.

V. Conclusion

This study demonstrates that the proposed IndoBERT-based model effectively classifies doctor responses in a multi-label setting, achieving strong overall performance, particularly for dominant categories such as Fostering the Relationship (FR), Gathering Information (GI), and Providing Information (PI). However, challenges remain due to class imbalance and semantic overlap, as reflected in lower performance for Decision Making (DM) and Responding to Emotions (RE).

The integration of explainable artificial intelligence (XAI) methods provides deeper insights into the model's decision-making process. Both SHAP and LIME reveal that the model captures not only empathetic expressions but also contextual and clinical information, highlighting the overlapping nature of doctor–patient communication. The results also suggest potential bias due to statistical co-occurrence patterns in the dataset. In terms of interpretability, LIME offers more intuitive explanations, while SHAP provides a more comprehensive and theoretically grounded analysis. The combination of both methods enhances the transparency and reliability of the model.

This study has several limitations. First, the dataset may not fully represent the diversity of real-world doctor–patient interactions, including variations in clinical settings, specialties, language styles, and patient characteristics. Second, class imbalance and semantic overlap between communication categories may affect the model's ability to accurately identify less frequent labels, particularly DM and RE. Finally, although SHAP and LIME improve interpretability, their explanations reflect associations learned from the dataset and do not necessarily indicate true clinical reasoning or causal relationships.

Future work should focus on addressing class imbalance, expanding dataset diversity, and exploring more advanced architectures to better capture the multi-functional nature of doctor responses.

References

[1] S. Juanita, D. Purwitasari, I. K. E. Purnama, A. F. Abdillah, and M. H. Purnomo, "Multi-Label Classification for Doctor's Behavioral Pattern Matching During Online Medical Interview using Machine Learning," *International Journal on Electrical Engineering and Informatics*, vol. 15, no. 3, pp. 435–452, Sep. 2023, doi: 10.15676/ijeei.2023.15.3.5.

[2] R. N. Hanami, R. Mahendra, and A. F. Wicaksono, "Semantic classification of Indonesian consumer health questions," *J. Biomed. Semantics*, vol. 16, no. 1, Dec. 2025, doi: 10.1186/s13326-025-00334-5.

[3] I. A. E. Madi, A. Redjdal, J. Bouaud, and B. Seroussi, "Exploring Explainable AI Techniques for Text Classification in Healthcare: A Scoping Review," in *Studies in Health Technology and Informatics*, IOS Press BV, Aug. 2024, pp. 846–850. doi: 10.3233/SHTI240544.

[4] S. Juanita, D. Purwitasari, I. K. E. Purnama, M. Raihan, and M. H. Purnomo, "Multi-Label Classification of Bilingual Doctor Responses in Online Medical Consultations Using Deep Learning," *Jurnal Nasional Pendidikan Teknik*

Informatika (JANAPATI), vol. 14, no. 2, pp. 432–445, Jul. 2025, doi: 10.23887/janapati.v14i2.96980.

[5] Y. Wang, Y. Wang, Z. Peng, F. Zhang, L. Zhou, and F. Yang, "Medical text classification based on the discriminative pre-training model and prompt-tuning," *Digit. Health*, vol. 9, Jan. 2023, doi: 10.1177/20552076231193213.

[6] X. Chen and Y. Du, "Enhancing medical text classification with GAN-based data augmentation and multi-task learning in BERT," *Sci. Rep.*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-98281-9.

[7] Y. Setiawan and L. A. Wulandhari, "Comparative Analysis of IndoBERT and LSTM for Multi-Label Text Classification of Indonesian Motivation Letter," *Jurnal Online Informatika*, vol. 10, no. 2, pp. 260–269, Aug. 2025, doi: 10.15575/join.v10i2.1499.

[8] N. K. Nissa and E. Yulianti, "Multi-label text classification of Indonesian customer reviews using bidirectional encoder representations from transformers language model," *International Journal of Electrical and Computer Engineering*, vol. 13, no. 5, pp. 5641–5652, Oct. 2023, doi: 10.11591/ijece.v13i5.pp5641-5652.

[9] H. Ahmadian, T. F. Abidin, H. Riza, and K. Muchtar, "Transformer-Based Indonesian Language Model for Emotion Classification and Sentiment Analysis," in *Proceeding - International Conference on Information Technology and Computing 2023, ICITCOM 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 209–214. doi: 10.1109/ICITCOM60176.2023.10442970.

[10] H. F. Karim and A. P. Wibowo, "BULLETIN OF COMPUTER SCIENCE RESEARCH Kinerja Metode Fine-Tuning IndoBERT untuk Klasifikasi Emosi Multi-Kelas pada Teks Informal Bahasa Indonesia," *Media Online)*, vol. 6, no. 1, pp. 63–74, 2025, doi: 10.47065/bulletincsr.v6i1.850.

[11] S. S. Band *et al.*, "Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods," *Inform. Med. Unlocked*, vol. 40, Jan. 2023, doi: 10.1016/j.imu.2023.101286.

[12] S. Ali *et al.*, "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence," *Information Fusion*, vol. 99, Nov. 2023, doi: 10.1016/j.inffus.2023.101805.

[13] S. Sharma, M. Singh, L. McDaid, and S. Bhattacharyya, "XAI-based Data Visualization in Multimodal Medical Data," Jul. 15, 2025. doi: 10.1101/2025.07.11.664302.

[14] K. Zahoor, N. Z. Bawany, and T. Qamar, "Evaluating text classification with explainable artificial intelligence," *IAES International Journal of Artificial Intelligence*, vol. 13, no. 1, pp. 278–286, 2024, doi: 10.11591/ijai.v13.i1.pp278-286.

[15] A. M. Ravichandran, J. Grune, N. Feldhus, A. Burchardt, S. Möller, and R. Roller, "XAI for Better Exploitation of Text in Medical Decision Support," 2024.

[16] K. Kitamura, M. Irvan, and R. Shigetomi Yamaguchi, "XAI for Medicine by ChatGPT Code interpreter," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, 2023, pp. 28–34. doi: 10.1145/3633624.3633629.

[17] A. Zevana and D. Riana, "Text Classification Using Indobert Fine-Tuning Modeling With Convolutional Neural Network And BI-LSTM," *Jurnal Teknik Informatika (Jutif)*, vol. 4, no. 6, pp. 1605–1610, Jan. 2024, doi: 10.52436/1.jutif.2023.4.6.1650.

[18] L. S. Darfiansa, Fitriyani, and S. S. A. Larasati, "Optimizing IndoBERT for Revised Bloom's Taxonomy Question Classification Using Neural Network Classifier," *Journal of Information Systems Engineering and Business Intelligence*, vol. 11, no. 2, pp. 226–237, Jun. 2025, doi: 10.20473/jisebi.11.2.226-237.

[19] O. Ozdemir, "Explainable AI (XAI) in Healthcare: Bridging the Gap between Accuracy and Interpretability," 2024.

[20] A. Hedström *et al.*, "Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond," Apr. 2023, [Online]. Available: http://arxiv.org/abs/2202.06861