Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics e-ISSN: 2656-8624
Homepage: https://ijeeemi.org/; Vol. 8, No. 3, pp. 380-390, August 2026

RESEARCH PAPER **OPEN ACCES**

Enhancing Low-Resource Healthcare Chatbots via Multi-Stage Data Augmentation and IndoBERT Fine-Tuning

Ahmad Wahyu Rosyadi[1], Taufiqur Rohman[1], Moh. Rizki Fajar[1], Muhammad Qomaruz Zaman[2] and Siti Ma'shumah[3]

[1] **Department of Informatic Engineering, Universitas Qomaruddin, Gresik, Indonesia**
[2] **Department of Electrical Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia**
[3] **Department of Electrical Engineering, Universitas Qomaruddin, Gresik, Indonesia**

Corresponding author: Ahmad Wahyu R. (e-mail: awrosyadi@gmail.com), **Author(s) Email**: Taufiqur Rohman (e-mail: taufiqurrohman@uqgresik.ac.id), Moh. Rizki Fajar (e-mail: mohrizkifajar@gmail.com), Muhammad Qomaruz Zaman (e-mail: me@zamzaman.com), Siti Ma'shumah (e-mail: sitimashumah@uqgresik.ac.id)

Abstract **Indonesian healthcare question-answering (QA) systems often operate in low-resource settings and must handle substantial linguistic variability in patient queries, including paraphrasing, informal expressions, and lexical variation. Limited annotated healthcare data further limits QA models' ability to generalize beyond surface-level wording. This study proposes a multi-stage data augmentation method to improve the robustness of an IndoBERT-based healthcare chatbot. The method combines domain-specific fine-tuning with medical-term protection, informal-language normalization, IndoT5-based paraphrasing, and controlled lexical variation through Part-of-Speech (POS)-guided verb synonym replacement. Medical entities are protected during augmentation to preserve domain-specific information while generating diverse question formulations without additional manual annotation. The augmented data are then used to fine-tune IndoBERT for extractive question answering. Experimental results show that full augmentation achieves 68.22% Exact Match (EM) and 82.06% F1, representing relative improvements of 13.2% in EM and 2.6% in F1 over the non-augmented baseline. These improvements are statistically significant at the 0.05 significance level based on paired t-tests. Notably, the ablation study shows that removing slang normalization yields the highest performance, reaching 69.59% EM and 85.21% F1. This finding suggests that preserving informal expressions may better reflect the linguistic characteristics of real-world patient queries than explicitly normalizing them. Overall, the results indicate that multi-stage augmentation can improve Indonesian healthcare QA under limited data conditions by increasing linguistic diversity while preserving critical medical information. The findings also highlight the importance of evaluating individual augmentation components to identify effective strategies for reliable healthcare chatbot development and more effective use of limited annotated healthcare data.**

Keyword: **Chatbot; Low-resource; Data augmentation; IndoBERT; IndoT5**

I. Introduction

The rapid advancement of Artificial Intelligence (AI) has transformed various sectors, including healthcare, by improving service quality, operational efficiency, decision-making, and access to medical information [1], [2], [3]. In this context, timely and accurate information regarding hospital medical services has become a critical public need [4], [5], [6]. Chatbots provide an effective solution by delivering automated responses continuously, improving access to information, and reducing staff workload [4], [5]. In Indonesia, Sahata Sitanggang et al. (2023) reported that chatbot implementation improved customer satisfaction, particularly in response speed and answer relevance, with an average score of 4.01 compared to 3.12 for conventional services [7]. Recent chatbot development has increasingly adopted transformer-based models such as BERT and GPT, which provide stronger contextual understanding for conversational applications [8]. Among these models, IndoBERT has demonstrated effectiveness in various Indonesian-language tasks, including emotion detection, public services, and domain-specific applications [9], [10], [11], [12]. These developments also support adopting AI-based chatbots in healthcare, including hospital and pandemic-related services [13]. However, healthcare question answering remains challenging because patient queries often contain informal expressions, varied sentence structures, and domain-specific terminology. However, limited dataset availability remains a major challenge for Indonesian-language healthcare chatbots due to complex medical terminology and sentence structures. Data augmentation has been shown to address this issue, including Modified Easy Data Augmentation, Knowledge Mixture Data Augmentation, and morphology-based augmentation using Indonesian affixes [14], [15], [16]. Recent studies have also emphasized the importance of controllable diversity and semantic fidelity in low-resource QA, particularly in sensitive domains such as healthcare

Manuscript received 10 February 2026; Revised 19 May 2026; Accepted 20 July 2026; Available online 27 August 2026
DOI: https://doi.org/10.35882/ijeeemi.v8i3.327
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

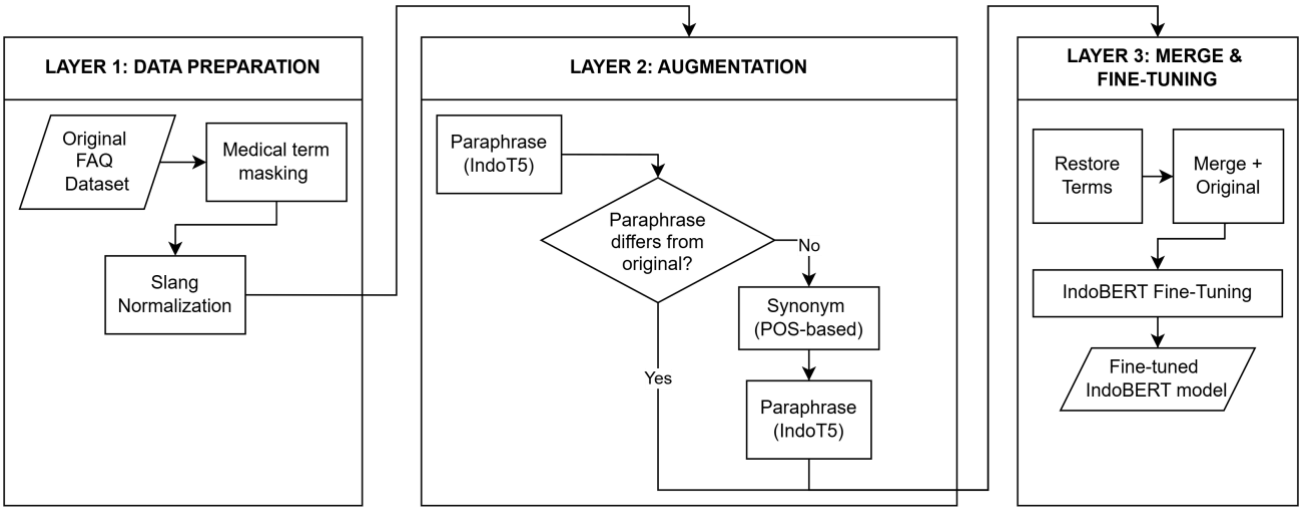


Fig. 1. Multi-Stage Data Augmentation Pipeline with Medical Term Protection for Healthcare QA

[17], [18]. These findings indicate that increasing data volume alone is insufficient, as augmentation must introduce linguistic variation while preserving domain-specific information.

The balance between diversity and semantic fidelity remains a key challenge in text augmentation. Wang et al. (2025) show that while increasing sample diversity is crucial, it must be balanced with label consistency to avoid model overfitting [19]. Similarly, constrained augmentation approaches have been proposed to improve generated sample quality through semantic and structural controls [20]. However, such strategies have been less explored for Indonesian healthcare question answering, particularly for patient queries containing informal language and domain-specific medical terms. To address limited labeled data and linguistic variability in Indonesian patient queries, this study proposes a medically-aware, multi-stage data augmentation pipeline for an IndoBERT-based hospital chatbot. Unlike prior work that applies generic augmentation without preserving medical semantics [14] or focuses solely on implementation without linguistic robustness [13], the proposed approach consists of three stages: (1) medical-term protection using placeholder-based masking to prevent semantic distortion; (2) staged augmentation through optional slang normalization, IndoT5-based paraphrasing, and POS-guided synonym substitution; and (3) evaluation under limited data conditions (730 QA pairs) and high linguistic variability, demonstrating improved extractive QA performance. The augmented dataset is then used to fine-tune IndoBERT, enabling a better understanding of natural and varied patient queries.

II. Materials and Method

A. Method

The proposed method consists of multiple sequential stages, as shown in Fig. 1, to generate a medically preserved augmented training dataset and fine-tune an IndoBERT-based healthcare QA model. The process begins with an Indonesian-language healthcare question dataset, which undergoes text normalization and slang-to-formal conversion. Medical terms are temporarily masked to prevent semantic distortion during augmentation. A transformer-based paraphrasing process is then applied and evaluated through an intent-preservation check. If necessary, re-paraphrasing and limited verb synonym replacement are performed to enhance linguistic diversity. The validated augmented questions are merged with the original dataset and used to fine-tune IndoBERT for robust healthcare question answering.

B. Input Dataset

The dataset was constructed from a hospital FAQ domain consisting of validated question–answer pairs that reflect real patient inquiries, as shown in Table 1. The dataset comprises 730 Indonesian QA pairs covering 15 hospital service intents, sourced from RS Fathma Medika, Gresik. Multiple question variants were generated using GPT-5.6 to represent diverse linguistic styles and formality levels while preserving the original intent. Each question was manually reviewed for relevance, clarity, intent consistency, and answerability, while irrelevant, ambiguous, or duplicate questions were revised or removed. The final dataset was divided into training and testing sets using an 80:20 ratio, resulting in 584 training and 146 testing pairs. Due to institutional policies, the dataset cannot be publicly shared.

C. Protect Medical Terms in Question Text

This stage preserves domain-specific medical terms and service names to prevent semantic distortion and misinterpretation of intent. Specifically, medical entities are identified using a domain-specific stoplist derived from hospital FAQs, standardized Indonesian healthcare terminology, and manual additions. Next, detection is performed using case-insensitive exact matching with word-boundary constraints and a longest-match-first strategy to prioritize multi-word entities. Formally, given a raw query Q and a set of medical entities M, the entity masking process is defined in Eq. (1). In total, 222 entities were identified, with 27.6% of sentences containing at least one entity. During augmentation, each entity was

Manuscript received 10 February 2026; Revised 19 May 2026; Accepted 20 July 2026; Available online 27 August 2026
DOI: https://doi.org/10.35882/ijeeemi.v8i3.327
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

replaced with a placeholder token (e.g., [[MED_1]]) and restored afterward, preserving critical healthcare information while enabling linguistic variation [21].

$$Q_{mask} = Replace(Q, M) \quad (1)$$

$$Y_{final} = \begin{cases} f_{fallback}(Q_{clean}), if\ Y = Q_{clean} \\ Y, \quad if\ Y \neq Q_{clean} \end{cases} \quad (2)$$

D. Convert Slang Words to Formal Words

After preserving domain-specific medical terms, informal or slang expressions are optionally normalized into their formal equivalents using IndoCollex [22]. This step is included to evaluate the contribution of slang normalization within the multi-stage augmentation pipeline rather than as a mandatory process. The full augmentation pipeline applies slang normalization, whereas the ablation configuration (Augmented − Slang) bypasses this step to assess its impact on QA performance. Throughout the process, sentence structure is preserved to avoid syntactic distortion. This design enables a controlled comparison between normalized and non-normalized augmentation strategies for the ablation analysis presented in Section 3.

E. Normalize Spaces and Punctuation

The Normalize Spaces and Punctuation stage performs text cleaning by standardizing excessive whitespace, repeated punctuation symbols, and irregular character usage that frequently appear in user-generated questions. This normalization step creates consistent spelling, ensuring clean input before paraphrasing and synonym replacement are applied.

Table 1. Sample FAQ question–answer pairs from the Indonesian healthcare dataset

No	Type	Representative Question	Answer
1	General Information & Facilities	Does the hospital have a cafeteria?	The hospital provides a cafeteria facility that can be used by patients and visitors.
2	Medical Schedules & Services	Is there information available regarding the doctor's practice schedule?	Doctor schedule information can be accessed through the hospital's official website or by asking directly at the information desk.
3	Payment	What types of payment does the hospital accept?	Payment methods accepted by the hospital include cash, non-cash, and coverage through insurance or BPJS.
4	Registration & Administration	What time does the registration counter open?	The registration counter at the hospital operates in the morning and adjusts to the polyclinic service schedule.
5	Registration & Administration	Can general patients visit more than one polyclinic?	General patients can register and visit more than one service according to their medical needs.
6	Inpatient Care	How much does inpatient care cost at the hospital?	The cost of medical procedures and other services will be adjusted according to the type of care and inpatient class chosen by the patient.

F. Paraphrase the Text

The paraphrasing stage generates linguistically diverse question variants while preserving semantic meaning to reduce reliance on rigid phrasing. This step uses an Indonesian adaptation of mT5 with a sequence-to-sequence architecture (IndoT5) to generate reformulations with varied lexical and syntactic structures while preserving intent [23], [24]. IndoT5 is optimized for Indonesian through tokenizer and embedding adaptation, reducing the vocabulary to 30K tokens and model size by approximately 58%, improving efficiency without sacrificing fluency and semantic quality[23]. To ensure quality, paraphrases are accepted only if they introduce lexical variation while preserving protected medical entities via masking. If the output Y is identical to Q_{clean}, it is treated as a failure. The system then retries generation by applying a fallback selection function $f_{fallback}$ that performs verb-level synonym substitution followed by a second paraphrasing attempt as defined in Equation (2). Samples that still show no variation are rejected. In practice, failures mainly occur when outputs replicate the input, while valid paraphrases preserve intent with meaningful variation.

G. Replace Up to Two Verbs with Synonyms

If paraphrasing fails to modify the query, a controlled synonym replacement strategy is applied using the Tesaurus Bahasa Indonesia [25], [26], for verb tokens v_i identified via POS tagging [27]. The candidate set $S(v_i)$ is formulated in Equation (3). To maintain naturalness and avoid semantic drift, substitutions are disambiguated within the same lexical group, limited to at most $k = 2$ verbs per sentence, and strictly exclude protected medical entities M. This strategy is defined in Equation (4). As a result, this approach preserves the original intent while complementing the paraphrasing stage with additional linguistic diversity.

$$S(v_i) = \{s_1, s_2, \ldots, s_k | s_j \in Tesaurus(v_i), POS(v_i) = VERB\} \quad (3)$$

$$Q_{syn} = Substitute(Q_{clean} \backslash M, \{v_i \rightarrow s_k\}_{i=1}^{k}) \quad (4)$$

H. Restore Medical Placeholders and Augmented Dataset Construction

Manuscript received 10 February 2026; Revised 19 May 2026; Accepted 20 July 2026; Available online 27 August 2026
DOI: https://doi.org/10.35882/ijeeemi.v8i3.327
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Table 2. Contribution of each augmentation strategy to dataset expansion and linguistic diversity

Dataset Configuration	Description	Number of QA Pairs	Diversity Contribution
Original Dataset	Original healthcare FAQ dataset without augmentation	730	Baseline linguistic variation
Augmented Full	Dataset with all augmentation strategies (paraphrasing, slang normalization, synonym substitution)	1,424	Combines structural, lexical, and conversational diversity
Augmented without Paraphrase	Augmentation without sentence structure variations (paraphrasing removed)	1,154	Reduced syntactic and structural diversity
Augmented without Slang	Augmentation without informal/slang language variations	1,425	Reduced conversational and informal language coverage
Augmented without Synonym	Augmentation without synonym substitution	1,373	Reduced lexical diversity from controlled word variation

Once a valid augmented question Q_{aug} is generated, the masked medical placeholders are restored to their original domain-specific terms. This restoration step ensures that all augmented questions retain medically accurate information while benefiting from increased linguistic variability. The restored questions are then added to the augmented dataset. To construct the final training corpus D_{train}, the augmented dataset D_{aug} is merged with the original healthcare question dataset D_{ori} as formulated in Equation (5). This combination preserves the original data distribution while enriching the training set with syntactic and lexical variations, resulting in a more diverse and robust dataset for model training.

$$D_{train} = D_{ori} \cup D_{aug} \tag{5}$$

I. IndoBERT Fine-Tuning

Recent advances in natural language processing have been largely driven by architectures built on the Transformer framework, which rely on attention-based mechanisms to model syntactic patterns and sentence-level meaning effectively [28], [29]. Pretrained transformer language models are now widely adopted because they provide rich contextual representations that can be adapted to downstream tasks through fine-tuning [30], [31]. In Indonesian-language NLP, IndoBERT serves as a strong foundation as it is pretrained on large-scale Indonesian corpora and can be further specialized for domain-specific applications such as healthcare question answering [32], [33].

In this study, a customized IndoBERT model (Rifky/Indobert-QA) for extractive question answering was fine-tuned using the translated SQuAD v2.0 dataset. The model predicts answer spans by identifying their start and end positions within the context, making it suitable for healthcare FAQ scenarios. The proposed multi-stage augmentation pipeline expands question variations, and the resulting data are used to fine-tune IndoBERT-QA to improve robustness to paraphrasing, informal expressions, and lexical variation. Fine-tuning used indolem/indobert-base-uncased with the AdamW optimizer, a learning rate of 2×10^{-5}, a batch size of 8, a maximum sequence length of 384, and three epochs with early stopping. Experiments were conducted on Kaggle using a T4 GPU and 30 GB RAM with PyTorch and Hugging Face Transformers. The best model was selected based on validation F1 score.

To ensure statistical reliability, all experiments were repeated five times using different random seeds [42, 123, 2024, 777, 9999]. Mean performance and 95% confidence intervals were calculated using the Student's t-distribution, and performance differences were assessed using paired t-tests at $\alpha = 0.05$, following standard practices for robust evaluation in low-resource NLP [34].

J. Fine-Tuned IndoBERT Healthcare QA Model

The proposed system produces a fine-tuned IndoBERT-based healthcare question answering model trained on a medically preserved augmented dataset. The resulting model exhibits improved robustness to linguistic variability while maintaining accurate medical intent understanding, and serves as the core component of the healthcare chatbot system which is subsequently evaluated using standard question answering metrics.

III. Results

This section evaluates the effectiveness of the proposed multi-stage data-augmentation pipeline for enhancing low-resource Indonesian healthcare chatbots through a series of experimental results. Four model configurations were evaluated: Long Short-Term Memory (LSTM), pre-trained IndoBERT, IndoBERT fine-tuned, and mBERT on the augmented dataset across several augmentation settings, including (1) the original dataset, (2) full augmentation, and (3) ablation settings where paraphrasing, slang normalization, or synonym substitution was removed. To evaluate model performance, this study adopts Exact Match (EM) and F1 metrics, which quantify strict answer correctness and partial overlap in extractive QA tasks [35].

This section evaluates the baseline performance of LSTM, pretrained IndoBERT, IndoBERT Fine-Tuned and mBERT on the original non-augmented dataset as a

Manuscript received 10 February 2026; Revised 19 May 2026; Accepted 20 July 2026; Available online 27 August 2026
DOI: https://doi.org/10.35882/ijeeemi.v8i3.327
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

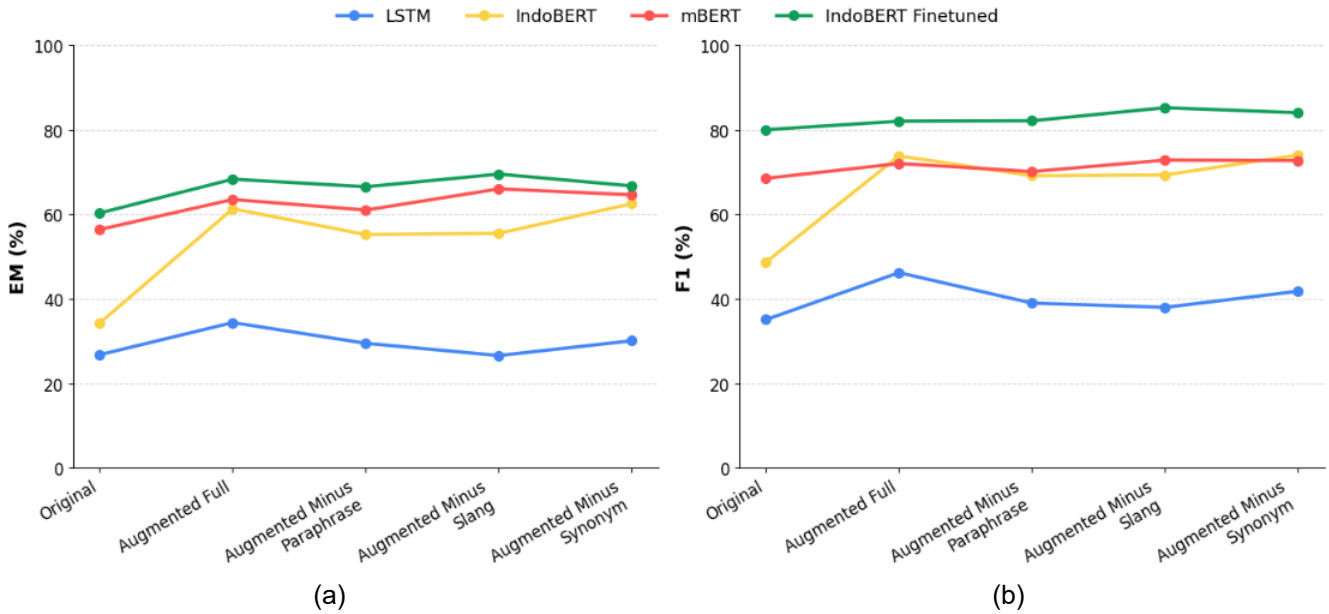
Table 3. Representative Qualitative Examples Comparing IndoBERT Predictions Before and After Augmentation (ORI = baseline/non-augmented model; AUG = augmented model)

Aspect	Example Question (Augmented Test Set)	Prediction (ORI)	Prediction (AUG)	Ground Truth	Observation
Robust handling of paraphrased queries	"My child suddenly has a fever at night, can I register directly to the hospital?"	The hospital registration counter operates 24 hours.	The hospital registration counter operates 24 hours.	The hospital registration counter operates 24 hours.	Augmentation improves robustness to paraphrasing.
Improved interpretation of informal language	"Does the hospital provide a place to sleep over?"	The hospital has a cafeteria facility next to the back parking lot.	The hospital provides special accommodation facilities (guest house) or cooperates with nearby hotels.	The hospital provides special accommodation facilities (guest house) or cooperates with nearby hotels.	Slang/ abbreviations handled better after augmentation.
Lexical flexibility	"If I pay with a credit card, is there an additional fee?"	There is no additional charge for credit card payments.	There is no additional charge for credit card payments.	There is no additional charge for credit card payments.	Reduced dependence on exact wording.
More precise span boundaries	"When is the allowed time to see the patient's condition in the treatment room?"	Yes	08.00 AM to 11.00 AM WIB and 04.00 PM to 05.00 PM WIB. For intensive care units (ICU), visiting hours are shorter and not open to the public.	Visiting hours for visitors are divided into two sessions: morning 08.00 AM to 11.00 AM WIB and afternoon 04.00 PM to 05.00 PM WIB. For intensive care units (ICU), visiting hours are shorter and not open to the public.	Span boundaries become more precise.
Remaining difficulty with implicit intent	"Are patients who have never been treated before allowed to come directly to register?"	Come immediately to the Emergency Installation (IGD) of the hospital which is available 24 hours.	Come immediately to the Emergency Installation (IGD) of the hospital which is available 24 hours.	There are 2 options: 1. New patients can register themselves or be registered by others by coming directly to the hospital, taking a queue number, filling out a biodata form, and bringing identification (Patient ID Card). 2….	Implicit intent remains challenging.

reference for assessing the impact of data augmentation. As shown in Table 2, the original dataset contains 730 Indonesian healthcare FAQ question–answer pairs, representing a controlled linguistic setting. The same table shows that full augmentation expands the dataset to 1,424 pairs, while the Augmented without Slang configuration produces one additional sample (1,425). This difference occurs because slang normalization occasionally generates paraphrases too similar to the original, which the validation criteria reject, whereas removing normalization allows one additional variation to pass. These results provide the basis for analyzing the performance improvements.

The LSTM model achieved the lowest performance, with 26.85 EM and 35.06 F1, as shown in Fig. 2, indicating limited ability to capture the semantic relationships required for Indonesian healthcare question answering. The large gap between EM and F1 suggests that the model often identifies parts of the correct answer but struggles to predict the exact answer span. In contrast, pretrained IndoBERT achieved 34.38 EM and 48.77 F1, demonstrating the benefit of transformer-based contextual representations for handling semantic variation and medical terminology. mBERT further improved performance to 56.44 EM and 68.43 F1, indicating stronger generalization across linguistic variations. However, the results also show that pretraining alone is insufficient to address domain-specific variability under limited labeled data.

Manuscript received 10 February 2026; Revised 19 May 2026; Accepted 20 July 2026; Available online 27 August 2026
DOI: https://doi.org/10.35882/ijeeemi.v8i3.327
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).



(a) (b)

Fig. 2. (a) EM and (b) F1 performance of LSTM, IndoBERT, fine-tuned IndoBERT, and mBERT models on the original dataset and various data augmentation settings

The fine-tuned IndoBERT model achieved the strongest baseline performance, with 60.27 EM and 79.96 F1, highlighting the effectiveness of domain adaptation for healthcare QA. However, the model still exhibits limitations in handling informal language, implicit medical intent, and ambiguous or multi-intent queries, motivating the use of data augmentation to improve robustness to linguistic variation.

Accordingly, integrating IndoBERT Fine-Tuned with the proposed data augmentation pipeline leads to a clear and consistent performance improvement, increasing from 60.27 EM and 79.96 F1 to 68.22 EM and 82.06 F1. The augmented IndoBERT Fine-Tuned model also consistently outperformed mBERT across all augmentation settings. This improvement indicates that the system is not only more accurate but also more robust in handling the linguistic variability commonly found in real-world Indonesian healthcare queries. Trained on a broader range of paraphrased questions, informal expressions, and lexical variations, the model learns more flexible representations of question intent, reducing its reliance on fixed wording and exact keyword matches.

The performance improvements from full data augmentation mainly stem from better handling of paraphrased questions, informal language, and lexical variation, as shown in Table 3, which provides a direct side-by-side comparison between baseline (ORI) and augmented (AUG) predictions. After augmentation, the fine-tuned IndoBERT model more consistently links differently worded questions with the same meaning to the correct answer spans, showing reduced reliance on fixed sentence structures and exact keyword matching. Informal expressions and abbreviations that previously led to errors are also interpreted more reliably, indicating better alignment between conversational user input and structured healthcare information.

As shown in Table 4, the Augmented Full configuration achieves 68.22% EM and 82.06% F1, representing improvements of 13.2% EM and 2.6% F1 over the fine-tuned IndoBERT baseline, while also outperforming

Table 4. Ablation analysis for IndoBERT fine-tuned showing mean ± 95% CI (N=5 runs), with Δ representing absolute difference vs. Augmented Full

Training Configuration	EM ± CI	F1 ± CI	Δ EM vs Full	Δ F1 vs Full	Interpretation (Linguistic Impact)
Augmented Full	68.22 ± 7.50	82.06 ± 5.47	-	-	Balanced robustness across paraphrastic, informal, and lexical variation.
Augmented − Paraphrase	66.57 ± 3.15	82.27 ± 2.93	-1.65	0.21	Less tolerant to sentence structure changes; EM drops.
Augmented − Slang	**69.59 ± 2.92**	**85.21 ± 2.78**	**1.37**	**3.15**	Best. Preserving slang improves informal query understanding.
Augmented − Synonym	66.71 ± 5.16	84.06 ± 3.73	-1.51	2.00	Less flexible in word choice → EM drops, but F1 remains robust as paraphrase and slang provide sufficient diversity.

Manuscript received 10 February 2026; Revised 19 May 2026; Accepted 20 July 2026; Available online 27 August 2026
DOI: https://doi.org/10.35882/ijeeemi.v8i3.327
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Table 6. Representative failure cases identified from qualitative error analysis of model predictions

Category	Example Question	Model Issue	Expected Intent
Implicit Intent	"How to register if I have never been to this hospital?"	Misclassified as emergency-related query	Registration procedure for new patients (not emergency)
Ambiguous Wording	"Is the IGD open 24 hours a day?"	Confused emergency service with registration desk	IGD operating hours (24/7 emergency service)
Medically Sensitive	"If I use corporate insurance, is it possible?"	Incorrectly mapped to BPJS cesarean policy	Payment method: corporate insurance accepted
Multi-part Fragmentation	"What documents are required for inpatient admission?"	Extracted partial answer only	Complete list: SPRI, ID card, insurance card, personal items (clothes, toiletries, assistive devices, medications); avoid jewelry/valuables
Service Conflation	"Does the hospital provide a place to stay?"	Confused lodging service with other facilities	Guest house / hotel cooperation for family accommodation

mBERT by 11.78% EM and 13.63% F1. Among the ablation settings, −Slang achieves the best performance (69.59% EM, 85.21% F1), suggesting that explicit slang normalization is unnecessary when other augmentation stages are applied. In contrast, removing paraphrasing or synonym replacement slightly reduces EM, indicating that both components contribute to model performance, although challenges in understanding implicit medical intent remain.

IV. Discussion

Table 4 shows that multi-stage augmentation improves healthcare question-answering performance, with the Full Augmentation configuration achieving 68.22 EM and 82.06 F1. Notably, the −Slang configuration performs even better, reaching 69.59 EM and 85.21 F1, suggesting that explicit slang normalization is not necessary when paraphrasing and synonym replacement are applied. Preserving informal expressions may help the model learn colloquial patterns that better reflect real-world patient queries. These findings suggest that slang normalization can be treated as an optional component of the augmentation pipeline. To position our work within the broader literature, Table 5 compares this study with previous work. Direct quantitative comparison is limited by differences in tasks, datasets, and evaluation metrics [33]. Nevertheless, the proposed approach improves EM by 13.2% over the fine-tuned IndoBERT baseline, demonstrating the effectiveness of the augmentation strategy. Unlike previous studies that apply generic augmentation [14], [15], [16], this study combines medical-term protection with staged augmentation and evaluates its effectiveness in a low-resource Indonesian healthcare QA setting. The ablation results further show that the augmentation components are not fully additive. Removing slang normalization improves F1 by 3.15 points, while removing paraphrasing or synonym replacement results in smaller changes (−0.21 to +2.00 F1). These findings suggest that preserving informal expressions and introducing structural variation are more beneficial than lexical substitution alone. The higher

Table 5. Methodological comparison with related Indonesian NLP and low-resource augmentation studies

Study	Task / Domain	Augmentation Method	Model	Evaluation Approach	Our Distinction
Zamakhsyari & Fatwanto [33]	Mental Health QA	None	IndoBERT	BERTScore (91.8%)	Extractive QA with EM/F1 metrics
Yanti et al. [13]	Chatbot Implementation Review	N/A (Systematic Review)	N/A	Qualitative synthesis	Technical development with empirical evaluation
Kartika et al. [14]	Paraphrase Identification	Modified EDA	IndoBERT	Accuracy & F1 (86.7%)	Medical term masking in augmentation
Wang et al. [15]	Low-resource QA	Knowledge Mixture	Various (survey)	General framework	Controlled POS-guided synonym substitution
Ashar & Siahaan [16]	Sentiment Analysis	Text Augmentation	Various	Accuracy	Healthcare QA with intent preservation

Manuscript received 10 February 2026; Revised 19 May 2026; Accepted 20 July 2026; Available online 27 August 2026
DOI: https://doi.org/10.35882/ijeeemi.v8i3.327
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

performance of −Slang also indicates that explicit normalization may introduce unnecessary changes to the linguistic characteristics of patient queries.

Despite these improvements, several failure cases remain, particularly in implicit intent understanding, ambiguous wording, medically sensitive questions, multi-part answer extraction, and service disambiguation as summarized in Table 6. These findings indicate that augmentation mainly improves surface-level linguistic robustness but does not fully address deeper semantic reasoning. Therefore, the system should be used as an informational assistant rather than a medical decision support tool, with appropriate escalation mechanisms for critical or low-confidence cases. The study is also limited to a single hospital domain (RS Fathma Medika) with 730 QA pairs and has not been externally validated. Future work should explore explicit intent modeling, external knowledge integration, and context-aware conversational reasoning, as well as evaluation on external datasets and real-world healthcare queries to assess generalizability, deployment readiness, and safety.

V. Conclusion

This study aimed to improve the robustness of Indonesian healthcare question answering in low-resource settings through multi-stage data augmentation and IndoBERT fine-tuning. The proposed approach improved performance over the non-augmented baseline, with full augmentation achieving 68.22% EM and 82.06% F1, representing relative gains of 13.2% in EM and 2.6% in F1. Notably, removing slang normalization yielded the highest performance (69.59% EM, 85.21% F1), suggesting that preserving informal expressions better matches real-world user queries. However, the model still struggles with implicit medical intent, and the findings are limited to a single-domain dataset from RS Fathma Medika without external validation. Therefore, the system should be used as an informational assistant with appropriate escalation mechanisms. Future work should explore explicit intent modeling and evaluate the approach on external datasets and real-world conversational queries to assess its generalizability and safety.

Acknowledgment

The authors would like to thank RS Fathma Medika, Gresik, for providing access to the hospital FAQ data used in this study.

Funding

This research was funded by the Research and Community Service Institute (LPPM) of Universitas Qomaruddin.

Author Contributions

All authors contributed to the study design, data preparation, methodology, experiments, analysis, and manuscript revision, and approved the final manuscript.

Data Availability

The dataset contains hospital-specific information and cannot be publicly shared due to institutional restrictions. We used it with permission from RS Fathma Medika, Gresik, and it is available from the corresponding author upon reasonable request and institutional approval.

Code Availability

The source code for the post-OCR correction system developed in this study is not publicly available but can be provided by the corresponding author upon reasonable request.

Ethical Approval

This study used hospital FAQ data without direct patient interaction or new human-subject data collection. The data were used with institutional permission from RS Fathma Medika, Gresik.

Consent To Participate

Informed consent was not applicable because the study did not involve direct participation or interaction with human subjects.

Consent for Publication

Consent for publication was not applicable because the study did not report identifiable individual patient information.

Competing Interests

The authors declare no competing interests.

Generative AI Use Declaration

GPT-5.6 was used to generate question variants for data augmentation. The authors manually reviewed the generated questions for relevance, clarity, intent consistency, and answerability. The authors remain responsible for the accuracy and final content of the manuscript.

References

[1] J. Ridha, K. Saddami, M. Riswan, and R. Roslidar, “An Explainable Artificial Intelligence Framework for Breast Cancer Detection,” *Indones. J. Electron. Electromed. Eng. Med. Informatics*, vol. 7, no. 2, pp. 298–311, 2025, doi: 10.35882/ijeeemi.v7i2.78.

[2] V. S. Laurenxius, C. N. Ginting, and R. Ginting, “Efficiency Of Transition From Manual Medical Records To Electronic Medical Records On The Speed Of Patient Service At RSU Royal Prima Medan,” *Indones. J. Electron. Electromed. Eng. Med. Informatics*, vol. 7, no. 2, pp. 245–252, 2025, doi: 10.35882/ijeeemi.v7i2.79.

[3] F. M. Aprihapsari, A. A. Octavia, K. Ain, and B. Ariwanto, “Multi-Channel Trans-Admittance Imaging for Anomaly Detection,” *Indones. J. Electron. Electromed. Eng. Med. Informatics*, vol. 7, no. 2, pp.

Manuscript received 10 February 2026; Revised 19 May 2026; Accepted 20 July 2026; Available online 27 August 2026
DOI: https://doi.org/10.35882/ijeeemi.v8i3.327
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

331–343, 2025, doi: 10.35882/ijeeemi.v7i2.88.

[4] O. Dogan and O. Faruk Gurcan, "Enhancing Hospital Services: Utilizing Chatbot Technology for Patient Inquiries," in *Lecture Notes in Networks and Systems*, C. Kahraman, S. Cevik Onar, S. Cebi, B. Oztaysi, A. C. Tolga, and I. Ucal Sari, Eds., Cham: Springer Nature Switzerland, 2024, pp. 233–239. doi: 10.1007/978-3-031-67192-0_29.

[5] A. Khamaj, "AI-enhanced chatbot for improving healthcare usability and accessibility for older adults," *Alexandria Eng. J.*, vol. 116, no. October 2024, pp. 202–213, 2025, doi: 10.1016/j.aej.2024.12.090.

[6] E. Grassini, M. Buzzi, B. Leporini, and A. Vozna, "A systematic review of chatbots in inclusive healthcare: insights from the last 5 years," *Univers. Access Inf. Soc.*, vol. 24, no. 1, pp. 195–203, 2025, doi: 10.1007/s10209-024-01118-x.

[7] A. Sahata Sitanggang, R. F. Syafariani, F. W. Sari, W. Wartika, and N. Hasti, "Relation of Chatbot Usage Towards Customer Satisfaction Level in Indonesia," *Int. J. Adv. Data Inf. Syst.*, vol. 4, no. 1, pp. 86–96, 2023, doi: 10.25008/ijadis.v4i1.1261.

[8] G. Yigit and R. Bayraktar, "Chatbot development strategies : a review of current studies and applications," *Knowl. Inf. Syst.*, vol. 67, no. 9, pp. 7319–7354, 2025, doi: 10.1007/s10115-025-02462-x.

[9] K. Hulliyah, F. Rayyan, and N. S. A. A. Bakar, "Development of A Chatbot for the Online Application Telegram Chat with An Approach to the Emotion Classification Text Using the Indobert-Lite Method," *2022 4th Int. Conf. Cybern. Intell. Syst. ICORIS 2022*, pp. 1–4, 2022, doi: 10.1109/ICORIS56080.2022.10031483.

[10] A. Fadhlurohman, A. Sulasikin, Y. Nugraha, N. L. R. Husna, M. E. Aminanto, and J. I. Kanggrawan, "Development of Indonesian Language Intelligent Chatbot for Public Services in JAKI Application," *Proc. 2023 IEEE Int. Smart Cities Conf. ISC2 2023*, pp. 1–7, 2023, doi: 10.1109/ISC257844.2023.10293514.

[11] A. Aziz, M. A. Khadija, W. Nurharjadmo, D. C. Febrianto, M. R. u. Firmansyah, and R. A. Pratama, "Enhancing Freelancer Project Matching with a BERT-Powered Deep Learning Indonesian Chatbot," *Int. Conf. Comput. Control. Informatics its Appl. IC3INA*, no. 2024, pp. 213–218, 2024, doi: 10.1109/IC3INA64086.2024.10732347.

[12] J. I. T. Krisna, A. Luthfiarta, L. D. Cahya, S. Winarno, and A. Nugraha, "Comparing Optimizer Strategies For Enhancing Emotion Classification In IndoBERT Models," *Adv. Sustain. Sci. Eng. Technol.*, vol. 6, no. 2, pp. 1–8, 2024, doi: 10.26877/asset.v6i2.18228.

[13] E. Yanti, A. Mardhiah, P. T. Novita, A. Ardilla, and F. H. Khana, "The Implementation of AI Chatbots in Indonesian Public Hospitals: Opportunities and Challenges," *Med. Res. Nursing, Heal. Midwife Particip.*, vol. 5, no. 2, pp. 329–337, 2024, doi: 10.59733/medalion.v5i2.149.

[14] B. V. Kartika, M. J. Alfredo, and G. P. Kusuma, "Fine-Tuned IndoBERT based model and data augmentation for indonesian language paraphrase identification," *Rev. d'Intelligence Artif.*, vol. 37, no. 3, pp. 733–743, 2023, doi: 10.18280/ria.370322.

[15] Y. Wang *et al.*, "Knowda: All-in-One Knowledge Mixture Model for Data Augmentation in Low-Resource Nlp Tasks," *11th Int. Conf. Learn. Represent. ICLR 2023*, pp. 1–25, 2023, [Online]. Available: https://openreview.net/forum?id=2nocgE1m0A

[16] M. N. Ashar and D. O. Siahaan, "Text Augmentation to Overcome Data Limitations in Sentiment Analysis for Bahasa Indonesia," *Proc. 2024 IEEE Int. Conf. Data Softw. Eng. Data-Driven Innov. Transform. Ind. Soc. ICoDSE 2024*, pp. 217–222, 2024, doi: 10.1109/ICODSE63307.2024.10829895.

[17] B. K. Rachmat, T. Gerald, Z. Zhang, and C. Grouin, "QA Analysis in Medical and Legal Domains: A Survey of Data Augmentation in Low-Resource Settings," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, vol. 4, pp. 1132–1144, 2025, doi: 10.18653/v1/2025.acl-srw.89.

[18] Y. Wei, Q. Li, and J. Pillai, "Structured LLM Augmentation for Clinical Information Extraction," *Stud. Health Technol. Inform.*, vol. 329, pp. 971–976, 2025, doi: 10.3233/SHTI250984.

[19] Z. Wang, J. Zhang, X. Zhang, K. Liu, P. Wang, and Y. Zhou, "Diversity-oriented Data Augmentation with Large Language Models," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, vol. 1, no. 3, pp. 22265–22283, 2025, doi: 10.18653/v1/2025.acl-long.1084.

[20] F.-Y. Su, G.-H. Ngo, B. Phan, and J.-H. Chiang, "CAS: enhancing implicit constrained data augmentation with semantic enrichment for biomedical relation extraction and beyond," *Database*, vol. 2025, p. baaf025, Jan. 2025, doi: 10.1093/database/baaf025.

[21] A. Karimi, L. Rossi, and A. Prati, "AEDA: An Easier Data Augmentation Technique for Text Classification," *Find. Assoc. Comput. Linguist. EMNLP 2021*, pp. 2748–2754, 2021, doi: 10.18653/v1/2021.findings-emnlp.234.

[22] H. A. Wibowo *et al.*, "IndoCollex: A Testbed for Morphological Transformation of Indonesian Colloquial Words," *Find. Assoc. Comput. Linguist. ACL-IJCNLP 2021*, no. 2017, pp. 3170–3183, 2021, doi: 10.18653/v1/2021.findings-acl.280.

[23] M. Fuadi, A. D. Wibawa, and S. Sumpeno, "Adaptation of Multilingual T5 Transformer for Indonesian Language," in *2023 IEEE 9th Information Technology International Seminar (ITIS)*, 2023, pp. 1–6. doi: 10.1109/ITIS59651.2023.10420049.

[24] Winda Puspitasari, Dani Ramdani, and Aditya Muhamad Maulana, "IndoT5 (Text-to-Text Transfer Transformer) Algorithm for Paraphrasing Indonesian Language Islamic Sermon Manuscripts," *Khazanah J. Relig. Technol.*, vol. 2, no. 2, pp. 63–73, 2025, doi: 10.15575/kjrt.v2i2.1093.

[25] P. Bahasa, *Tesaurus Bahasa Indonesia*, Edisi Pert.

Manuscript received 10 February 2026; Revised 19 May 2026; Accepted 20 July 2026; Available online 27 August 2026
DOI: https://doi.org/10.35882/ijeeemi.v8i3.327
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Jakarta: Departemen Pendidikan Nasional, 2008.

[26] M. D. Ferdiansyah and S. Suyanto, "Data Augmentation and Fine-Tuning to Improve IndoBART Performance," *2024 ASU Int. Conf. Emerg. Technol. Sustain. Intell. Syst. ICETSIS 2024*, pp. 1668–1672, 2024, doi: 10.1109/ICETSIS61505.2024.10459464.

[27] A. Z. Arifin, M. Z. Abdullah, A. W. Rosyadi, D. I. Ulumi, A. Wahib, and R. W. Sholikah, "Sentence extraction based on sentence distribution and part of speech tagging for multi-document summarization," *Telkomnika (Telecommunication Comput. Electron. Control.*, vol. 16, no. 2, pp. 843–851, 2018, doi: 10.12928/TELKOMNIKA.v16i2.8431.

[28] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," *28th Int. Conf. Comput. Linguist. Proc. Conf.*, pp. 757–770, 2020, doi: 10.18653/v1/2020.coling-main.66.

[29] H. Bao, L. Dong, W. Wang, N. Yang, S. Piao, and F. Wei, "Fine-tuning pretrained transformer encoders for sequence-to-sequence learning," *Int. J. Mach. Learn. Cybern.*, vol. 15, no. 5, pp. 1711–1728, 2024, doi: 10.1007/s13042-023-01992-6.

[30] V. C. Mawardi, J. M. K. Jonathan, J. Syahwalina, S. Monika, S. Widoatmodjo, and E. Wijaya, "Dataset Alignment for Fine-Tuning Large Language Models for PJOK Educational Chatbot," in *2024 Ninth International Conference on Informatics and Computing (ICIC)*, 2024, pp. 1–6. doi: 10.1109/ICIC64337.2024.10956758.

[31] M. A. K. Fata, S. Sumpeno, A. D. Wibawa, and D. A. Feryando, "Evaluating the Sentiment Analysis from Auto-Generated Summary Text Using IndoBERT Fine-Tuning Model in Indonesian News Text," in *2023 IEEE 15th International Conference on Computational Intelligence and Communication Networks (CICN)*, 2023, pp. 822–829. doi: 10.1109/CICN59264.2023.10402345.

[32] W. O. Vihikan and I. N. P. Trisna, "Indonesian Health Question Multi-Class Classification Based on Deep Learning," *J. Inf. Syst. Informatics*, vol. 6, no. 3, pp. 1931–1944, 2024, doi: 10.51519/journalisi.v6i3.838.

[33] F. Zamakhsyari and A. Fatwanto, "Language model optimization for mental health question answering application," *Int. J. Electr. Comput. Eng.*, vol. 15, no. 5, pp. 4829–4836, 2025, doi: 10.11591/ijece.v15i5.pp4829-4836.

[34] P. C. Chang, "95% Confidence Interval BT - Data Analysis of Medical Studies: Reading and Reporting," P. C. Chang, Ed., Cham: Springer Nature Switzerland, 2025, pp. 385–391. doi: 10.1007/978-3-031-49984-5_12.

[35] S. Satheesh, K. Beckh, K. Klug, H. Allende-Cid, S. Houben, and T. Hassan, "Robustness Evaluation of the German Extractive Question Answering Task," in *Proceedings of the 31st International Conference on Computational Linguistics*, O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. Di Eugenio, and S. Schockaert, Eds., Abu Dhabi, UAE: Association for Computational Linguistics, Jan. 2025, pp. 1785–1801. [Online]. Available: https://aclanthology.org/2025.coling-main.121/

Author Biography

Ahmad Wahyu Rosyadi serves as an academic staff member and researcher in Informatics at Qomaruddin University. He earned his bachelor's degree from Universitas Islam Negeri Maulana Malik Ibrahim Malang (2014) and his master's from Institut Teknologi Sepuluh Nopember (ITS) Surabaya (2018). His research spans AI-driven systems, machine learning, image analysis, visual computing, and Indonesian text retrieval. He actively teaches and supervises undergraduate research while contributing to applied AI studies. His current interests include intelligent systems development, OCR-based applications, and healthcare information retrieval, focusing on practical implementations that address real-world problems and support local technological advancement in Indonesia.

Taufiqur Rohman is an academic and researcher affiliated with Qomaruddin University. He completed his undergraduate education at Universitas Trunojoyo Madura, followed by postgraduate studies in information systems at Institut Teknologi Sepuluh Nopember Surabaya. His academic work primarily addresses web-based information systems to support educational and administrative digitalization. His scholarly publications contribute to information systems in education, database-driven applications, and practical software engineering solutions tailored to organizational needs. He regularly teaches and supervises undergraduate students. His ongoing research explores software engineering practices, educational information systems, web application development, and IT applications for real-world challenges in educational environments.

Moh. Rizki Fajar is an alumnus of Informatics Engineering at Qomaruddin University with strong interests in intelligent application development and computer vision. During his undergraduate studies, he completed a final project developing a computer vision-based system for recognizing Javanese script, involving image preprocessing, feature extraction, and model evaluation to improve character recognition accuracy. Through this project, he gained hands-on experience in machine learning and visual data processing. His academic interests include artificial intelligence, pattern recognition, and applied computer vision for preserving local cultural heritage through digital technologies. He continues developing his technical skills and aspires to contribute innovative AI-based solutions in practical domains.

Manuscript received 10 February 2026; Revised 19 May 2026; Accepted 20 July 2026; Available online 27 August 2026
DOI: https://doi.org/10.35882/ijeeemi.v8i3.327
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).

Muhammad Qomaruz Zaman earned his Bachelor and Master of Engineering in Electrical Engineering from Institut Teknologi Sepuluh Nopember (ITS) in 2018, and completed his Ph.D. at the National Taipei University of Technology, Taiwan, in 2025. He is currently a faculty member in the Electrical Engineering academic unit at Institut Teknologi Sepuluh Nopember. His scholarly work centers on robotics and intelligent systems, particularly fuzzy inference systems, evolutionary algorithms, and machine learning for autonomous applications. He actively conducts academic research, teaches, and supervises student projects, emphasizing the integration of artificial intelligence with engineering solutions for advanced control systems.

Siti Ma'shumah serves as a lecturer and researcher in the Department of Electrical Engineering at Universitas Qomaruddin, Gresik, Indonesia. She obtained her Bachelor's degree in Electrical Engineering from Institut Teknologi Nurul Jadid in 2015, followed by a Master's degree in the same discipline from Institut Teknologi Sepuluh Nopember (ITS), Surabaya, in 2018. Her research interests span analog electronics, Internet of Things (IoT) systems, artificial neural networks, and their integration into electrical and biomedical applications. She has actively participated in numerous research initiatives focusing on intelligent systems and electronic instrumentation. As an IEEE member, she remains engaged with the broader academic community. She can be reached via email at sitimashumah@uqgresik.ac.id.

Manuscript received 10 February 2026; Revised 19 May 2026; Accepted 20 July 2026; Available online 27 August 2026
DOI: https://doi.org/10.35882/ijeeemi.v8i3.327
Copyright © 2026 by the authors. Published by Jurusan Teknik Elektromedik, Politeknik Kesehatan Kemenkes Surabaya Indonesia. This work is an open-access article and licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0).