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.

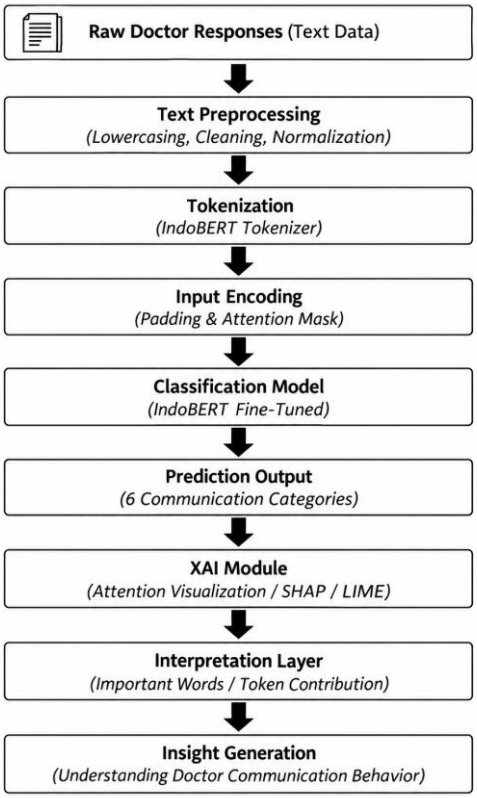


Fig. 1. XAI Pipeline for Interpreting Doctor Response Classification

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.

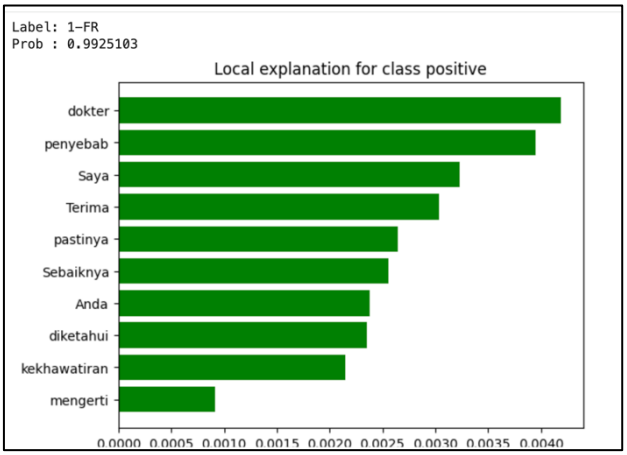


Fig. 4. Lime XAI Result

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