Received: 19 April 2026 | Revised: 2 July 2026 | Accepted: 14 September 2026 | Published online: 29 September 2026

RESEARCH ARTICLE

Artificial Intelligence and Applications
2026, Vol. 00(00) 1–18
DOI: 10.47852/bonviewAIA620210055

BON VIEW PUBLISHING

Explainable AI in Healthcare: Text Classification Using Bidirectional LSTM and LIME

Adam Mohd Khairuddin[1,*], Siti Armiza Mohd Aris[2], Azizul Azizan[1] and Noor Jannah Zakaria[1]

[1] *Department of Intelligence Informatics, Universiti Teknologi Malaysia, Malaysia*

[2] *Department of Smart Engineering and Advanced Technology, Universiti Teknologi Malaysia, Malaysia*

Abstract: Although deep learning models can help to improve clinical text classification significantly, their poor interpretability made them unsuitable for adoption in the medical settings. Unlike the past studies, this research adopted the explainable artificial intelligence (XAI) technique to improve the interpretability of the bidirectional long short-term memory (Bi-LSTM) model for clinical text classification. In particular, the study selected the local interpretable model-agnostic explanations (LIME) technique as the XAI technique to interpret the Bi-LSTM model used to classify 24 classes of diseases. The proposed Bi-LSTM model was developed based on the following three main stages: (1) preprocessing; (2) classification; and (3) model explanation. Data for symptoms and diseases were gathered from the Kaggle and used to train and evaluate the model. The outcomes of the study showed that the model achieved precision, recall, F1-score, and accuracy of 98.00%. In addition, the model was further reevaluated by using two additional datasets. By utilizing the 40-class medical transcription dataset, the model attained 76.18% accuracy with F1-score of 0.82. However, by adopting the 30-class symptom-based disease dataset, it accomplished 95.26% accuracy with F1-score of 0.95. The results further unveiled that the LIME technique was not only able to help understand the decision-making process of the model by providing the local explanations, but it also acted as an active diagnostic instrument for refining the preprocessing pipeline during the development of the model.

Keywords: natural language processing, recurrent neural network, explainable AI

1. Introduction

Electronic health record (EHR) or electronic medical record (EMR) systems can be defined as the software system that managed digital version of patient health information for the healthcare providers [1]. These systems contain important information such as patient's medical history, medical health information, diagnoses, medications, immunization dates, radiology images, lab results, and demographic data which are essential for making informed healthcare decisions [2]. In the last 10 years, the EHR systems have drastically changed the way healthcare is delivered with the introduction of electronic medical records instead of the traditional paper-based records. At the same time, they have improved patient record management, reduced administrative burden, enhanced patient care quality and better coordinated care among healthcare providers [3–5].

The EHR systems have been able to enhance the functioning of healthcare institutions, but they have also faced some challenges in adoption and use [6–8]. One of the core problems was that the user interface of the EHR systems was not designed well and as such had limited the clinical workflows as well as decreased the efficiency of the medical staff [9]. It was also found that the unoptimized user interface had caused the cognitive load to increase, resulted in longer interaction times, and contributed to higher rates of human error. In addition, the manual data entry created clinician fatigue that not only reduced the time for patient care but also affected the overall quality of the healthcare services [10]. The other major challenge involved the need to deal with the huge volume of unstructured clinical information such as physician notes, discharge summaries, and diagnostic reports [11]. All these shortcomings have caused the analytical process to become more complex, time-consuming as well as making it more dependent on the manual efforts to obtain the relevant clinical information for predicting patient outcomes.

However, the unstructured clinical text within the EHR systems can be analyzed by using the natural language processing (NLP) [12]. By adopting the machine learning (ML) and the deep learning (DL) techniques, the NLP can automatically extract the relevant clinical features from the symptoms and the test results for diagnosis and patient outcome prediction [13, 14]. More significantly, these techniques which were based on the data-driven insights can also be incorporated in the personalized medicine which allowed treatment options to be developed and tailored according to the specific needs of the individual patients [15].

***Corresponding author:** Adam Mohd Khairuddin, Department of Intelligence Informatics, Universiti Teknologi Malaysia, Malaysia. Email: adam.mk@utm.my

© The Author(s) 2026. Published by BON VIEW PUBLISHING PTE. LTD. This is an open access article under the CC BY License (https://creativecommons.org/licenses/by/4.0/).

In performing the clinical text classification tasks, the adoption of the deep learning models such as the recurrent neural networks (RNNs) and their variants had also been found to be significant. Nonetheless, these models were frequently considered as black box systems that lacked clinical transparency and interpretability [16, 17]. Without transparency and interpretability, there will be no trust, accountability, and explainability which are very essential for their clinical adoption. Fortunately, there are several XAI methods that can be utilized to overcome these limitations. They can also provide insights into the deep learning model's decision-making process. For instance, as one of the XAI methods, the local interpretable model-agnostic explanations (LIME) can highlight the contribution of specific features in the input text that can be used to support the clinicians in validating and trusting the AI models [18–20].

In view of the challenges and issues related to the deep learning models documented in the previous studies, this study proposed the Bi-LSTM model for the clinical text classification of various disease classes. The model was found to be suitable not only for capturing the contextual dependencies of sequential text data but also it had the ability to process information in both the forward and the backward directions [21, 22]. Equally important, the LIME technique was integrated into the model to provide local explanations for the classification outcomes of the Bi-LSTM model. Combining the Bi-LSTM model with the LIME technique aimed to achieve the balance between the predictive performance and the model transparency for the proof-of-concept clinical text classification as well as with the acknowledgement that further validation on larger external datasets would be necessary before the real-world clinical deployment.

The main contributions of this study are summarized as follows: (1) the development of a domain-specific clinical preprocessing pipeline that improved both the classification performance and the LIME explanation quality in a multi-class disease classification task; (2) the systematic use of the LIME not merely as a post-hoc explanation tool but also as an active diagnostic instrument during the model development to detect and correct preprocessing deficiencies in the Bi-LSTM model; and (3) the ablation analysis comparing the baseline and the fine-tuned configurations that quantified the impact of the targeted preprocessing on both the predictive and the explanatory outcomes.

Accordingly, this paper is divided into the following sections. The reviews of the related works on the techniques used to classify clinical texts as well as the XAI techniques adopted for model interpretation are reviewed in Section two below. The following Section three explains the methodology, the data preprocessing, the model architecture, and the details of the implementation of the model proposed in this study. In Section four, the experimental results of the study are discussed. Finally, the conclusion of the paper and suggestions for future research are presented in Section five.

2. Literature Review

According to the review of the literature, increasing number of studies have attempted to adopt the explainable AI (XAI) techniques to address the issue of lack of transparency related to models developed by using both machine learning (ML) and deep learning (DL) techniques in clinical text classification in the healthcare domain [23]. More specifically, the model-agnostic approaches such as the LIME, the shapley additive explanations (SHAP) as well as the gradient-based techniques had been widely adopted to improve interpretability.

The literature also revealed that numerous research had attempted to examine the effectiveness of the LIME technique in explaining the model predictions for clinical text classification. For an example, the study by [24] explored the use of the LIME technique for the support vector machine (SVM) with the RoBERTa-base for the classification of patient safety event (PSE) reports. The results of the study indicated that LIME provided interpretable representations of the textual context. However, the research also pointed out the need to validate the reliability and consistency of the generated explanations.

In addition, there are also several studies that attempted to explore multiple XAI techniques. For instance, the transformer-based models were used in the study by [25] to classify magnetic resonance imaging (MRI) reports for multiple sclerosis. This particular study also employed the LIME, the SHAP, and the integrated gradients (IG) to interpret the models. According to the results of the study, both the SHAP and the LIME techniques were effective in identifying clinically relevant features that corresponded with the domain knowledge. The variations in the explanation outputs across the different XAI techniques however suggested that there were potential challenges in explanation consistency and robustness.

Transformer-based architectures have also been extended to multilingual and low-resource clinical applications. The other work by [26] examined the use of the XLM-RoBERTa for depression classification from the Thai-language transcripts with the LIME applied for interpretability. The study demonstrated that the LIME was able to identify the key linguistic features contributing to depression screening. However, its applicability remained limited when extended to more complex psychiatric conditions such as bipolar disorder type II or schizophrenia.

Other than the model-agnostic approaches, model-specific XAI techniques had also been utilized in other studies. For instance, the medical text processing study by [27] applied the Gradient-weighted Class Activation Mapping (Grad-CAM) to a ResNet-based model. The heatmaps created in this study were capable of giving the visual explanations as well as identifying the most influential features for prediction. Although the Grad-CAM approach was able to provide intuitive visual explanations, its evaluation relied primarily on the qualitative assessment undertaken by the domain experts which also indicated the lack of unified quantitative assessment for explainability.

In the case of the research by [28], the study adopted the convolutional neural network (CNN) model for the ordinal text classification of psychiatric symptom severity and the LIME technique to generate instance-level explanations. Even though the results of the study were able to show the ability of LIME to identify important contributing features, the model exhibited limited generalizability when applied to atypical cases.

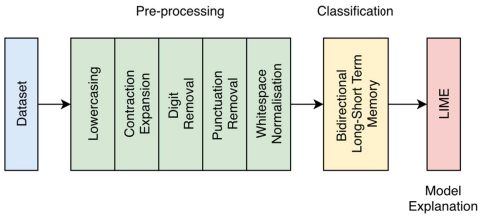
Although the literature showed that there were increasing numbers of research that attempted to address the issue of poor interpretability of AI models, findings of these studies remained limited in terms of their focus and scope. First, limited number of studies have attempted to investigate more complex multi-class clinical text classification problems faced by the healthcare institutions. Second, although the adoption of the LIME and the SHAP techniques has shown promising results, their explanation consistency and clinical reliability across various models and datasets are still not fully addressed. Third, there is still limited investigation on the sequential deep learning models such as the bidirectional long short-term memory (Bi-LSTM), which has been found to be effective for capturing contextual dependencies in clinical narratives.

In view of these research limitations, this study adopted the Bi-LSTM model as well as the LIME technique for multi-class clinical text classification to enhance both the predictive performance and interpretability.

3. Research Methodology

In this study, the proposed framework was developed based on three stages that involved: (1) preprocessing; (2) classification; and (3) model explanation. The following Figure 1 shows the three stages of the proposed framework.

Figure 1
Three stages of the proposed framework



3.1. Dataset

The symptom and disease dataset available publicly on the Kaggle was used for developing as well as testing the Bi-LSTM model as proposed in this study. The dataset comprised a total of 1200 patient symptom descriptions categorized into 24 distinct disease classes.

The 24 classes of diseases in the dataset included systemic, respiratory, dermatological, and gastrointestinal ailments. More specifically, the 24 classes of diseases included: (1) psoriasis; (2) varicose veins; (3) typhoid; (4) chicken pox; (5) impetigo; (6) dengue; (7) fungal infection; (8) common cold; (9) pneumonia; (10) dimorphic hemorrhoids; (11) arthritis; (12) acne; (13) bronchial asthma; (14) hypertension; (15) migraine; (16) cervical spondylosis; (17) jaundice; (18) malaria; (19) urinary tract infection; (20) allergy; (21) gastroesophageal reflux disease; (22) drug reaction; (23) peptic ulcer disease; and (24) diabetes.

The distribution of the samples across the 24 disease classes is shown in Figure 2 below. The dataset was rather balanced, with each class having exactly 50 textual instances. The uniform distribution was advantageous for training the proposed Bi-LSTM architecture because it eliminated the risk of class imbalance as well as prevented the model from developing any predictive bias toward a majority class.

3.2. Bidirectional LSTM

The 24 classes of diseases were classified by using the Bi-LSTM model. The following Figure 3 highlights the proposed architecture of the classification model. Prior to the model training, the text data from the dataset underwent the preprocessing pipeline. The preprocessing pipeline involved converting the text to lowercase, expanding contractions, and removing the numerical digits as well as standard punctuation. The cleaned text was then tokenized with a maximum vocabulary size of 10,000 words and padded to a uniform maximum sequence length of 200 tokens. The dataset was then stratified and split into training (60%), validation (20%), and testing (20%) sets to ensure a balanced representation of all the 24 disease classes across the training phases. The split resulted in only 10 test samples per disease class which also limited the statistical reliability of the per-class performance estimates. However, to mitigate this limitation,

Figure 2
Distribution of samples across 24 disease classes

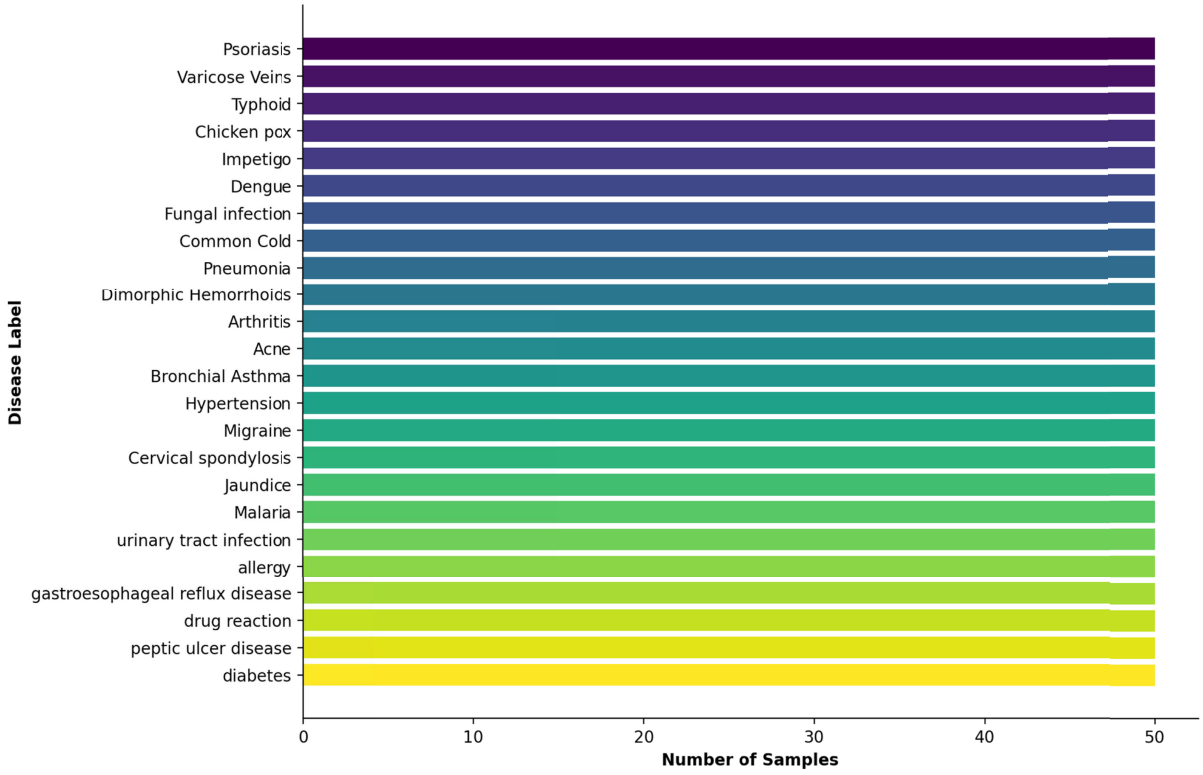
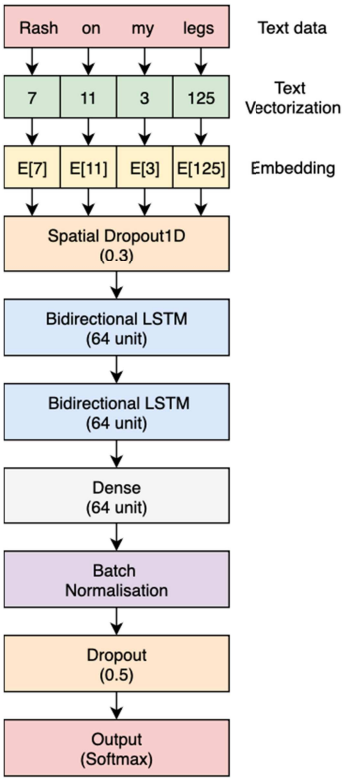


Figure 3
The proposed architecture of the Bi-LSTM model



the 5-fold stratified cross-validation was conducted on the full dataset in addition to the holdout evaluation. The cross-validation results were reported alongside the standard train/validation/test split metrics in order to provide a more robust assessment of the model generalizability.

The sequential architecture of the model was initiated with an embedding layer that converted the input tokens into dense vectors. This layer was configured with an input dimension equal to the vocabulary size of 10,000 words and an output dimension of 128. The SpatialDropout1D layer with a rate of 0.3 dropped the entire 1D feature maps instead of individual elements. The layer also helped to reduce overfitting as well as minimized the model's dependence on specific word embeddings.

Following the embedding stage, the two stacked bidirectional LSTM layers were used to process the sequence. There were 64 neurons in each of them. The first Bi-LSTM layer output was fed into the second Bi-LSTM layer, and the second Bi-LSTM layer returned a single output vector.

The dense layer with 64 neurons was used to obtain higher-level abstract features and employed the rectified linear unit (ReLU) as its activation function. This was followed by the batch normalization layer to accelerate convergence and stabilized the training process. The standard dropout layer with a rate of 0.5 was then employed to further prevent overfitting by randomly disabling 50% of the neurons during the training. The final output layer featured 24 neurons with the Softmax activation function to output the probability distribution for the 24 multi-class diseases.

The Bi-LSTM model was optimized using the Adam optimizer and evaluated utilizing the categorical cross-entropy loss during the compilation phase. The early stopping callback was applied to monitor the validation loss with a patience of 6 epochs automatically halted the training and restored the best model weights upon convergence. In addition, the ReduceLROnPlateau callback was utilized to dynamically decayed the learning rate by a factor of 0.2 if the validation loss stagnated for 2 consecutive epochs.

3.3. Performance metrics

In clinical text classification, using only a single evaluation metric can obscure specific weaknesses of the model. Due to this shortcoming, the model was also assessed using both the holdout test set and the 5-fold stratified cross-validation to address the issue of the small per-class sample size (n = 10 per class in the holdout set).

Given the need to use more than one metric, the overall performance of the model was assessed by using the following four specific performance metrics: (1) accuracy; (2) precision; (3) recall; and (4) F1-score. These metrics were mathematically derived from the four fundamental confusion matrix parameters that included: (1) True Positives (TP); (2) True Negatives (TN); (3) False Positive (FP); and (4) False Negatives (FN) of the predicted outcomes of the model.

The True Positives were the patient symptoms correctly classified as the target disease. The True Negatives were the symptoms correctly identified as not belonging to the target disease. The False Positives were symptoms incorrectly classified as the target disease. The False Negatives were the actual symptoms of the target disease incorrectly classified as another condition.

As shown in Equation 1, the accuracy was the ratio between the number of correct predictions over the total predictions. The accuracy acted as an indicator of the overall model performance without being skewed by a majority class due to the balanced dataset (50 samples per disease).

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{1}$$

The precision quantified the ratio of the correctly predicted positive instances to the total positive predictions produced by the Bi-LSTM model. High precision was vital to minimize the false positives.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

The recall referred to the ratio of the true positives to the total number of the actual positive instances. It evaluated the ability of the model to minimize the false negatives.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

The F1-score was considered as the harmonic mean of precision and recall. For the multi-class disease classification, this measure was essential because of the serious clinical consequences of both false positives and the false negatives.

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

3.4. Model explanation

In the study, the LIME technique was utilized to generate local explanations. This technique emphasized on explaining the prediction made by the model for a specific data point (instance), instead of describing the global behavior of the model. The technique weighed each of the features that reflected their contribution to the model prediction for the specific single instance.

According to the clinical natural language processing, the LIME operated through the systematic text perturbation mechanism. Starting from the original patient symptom description, the LIME generated hundreds of modified variations of the text by randomly removing different combination of words. For instance, consider the patient symptom description in Table 1 which the Bi-LSTM predicted it with a high probability (0.98) and correctly predicted it as Pneumonia.

The localized dataset of perturbations was developed by the LIME to explain the prediction made by the Bi-LSTM model. The perturbed text samples were fed back into the trained model, and the probabilities of the output prediction were compared. The proximity weights were also calculated to measure the similarity between the generated variations and the original sentence. The proximity weights are presented in Table 2.

When the clinical terms were not present (such as "cough" or "pain"), the model's confidence for the diagnosis of Pneumonia fell significantly, and the algorithm mathematically learned that the specific terms were the ones that were most important for the diagnosis.

The LIME was finally used to fit a linear regression model to the perturbed dataset to estimate the local feature importance. This technique yielded a set of quantitative weights that represented the directionality of each word in the prediction of the model as listed in Table 3.

As shown in Table 3, the positive weights indicated the clinical terms that actively supported the predicted disease, while the negative weights represented words that contradicted it. By integrating strict stop-word removal during the preprocessing phase, the linguistic artifacts (*e.g.*, "I," "have") were removed from predictive power (weight = 0) to ensure that the Bi-LSTM model focused on pathology.

4. Results and Discussion

The experiments in this study were conducted on the Jupyter Notebook with the Python version 3.12.7 based on the following PC specifications: (1) AMD Ryzen 9 9900X; (2) 32 GB of DDR5 RAM; (3) RTX 4070 GPU; and (4) Windows 11. The section below describes the performance of the model in classifying the 24 classes of diseases as well as the interpretation of the Bi-LSTM model by using the local interpretable model-agnostic explanations (LIME) technique.

4.1. Data preprocessing

The quality of the data as well as the performance of the Bi-LSTM model were improved by the pre-data processing processes which involved: (1) lowercasing; (2) contraction expansion; (3) digit removal; (4) punctuation removal; and (5) whitespace normalization. The section below briefly explained each of the five preprocessing processes.

First, all the textual data were converted to the lowercase to ensure uniformity. This prevented the Bi-LSTM model from treating capitalized and non-capitalized versions of the same word (e.g., "Fever" and "fever") as distinct entities.

Second, the common English contractions (e.g., "can't" to "cannot") were expanded. This process was crucial to preserve the semantic clarity as well as the grammatical intent of the patient's

Table 1
Patient symptom description and Bi-LSTM prediction

Content	Actual Class	Bi-LSTM Prediction
"I have a severe cough and chest pain"	Pneumonia	Pneumonia (Probability: 0.98)

Table 2
LIME text perturbations, proximity weights, and resulting probabilities for Pneumonia

Perturbed Text Sample	Probability	Weight
"I have a severe cough and chest pain"	0.98	1.00
"have a severe cough chest pain"	0.95	0.75
"I have a severe and"	0.15	0.62
"severe chest pain"	0.45	0.37

Table 3
Estimated local feature weights

Feature	Weight	Clinical Implication
cough	+5.24	Primary diagnostic driver.
chest	+3.12	Secondary diagnostic driver.
pain	+2.80	Secondary diagnostic driver.
severe	+0.50	Minor contextual modifier.
I	0.00	Linguistic artifact (ignored).
have	0.00	Linguistic artifact (ignored).

descriptions before the apostrophes were removed in the later stages.

Third, all the numerical digits were removed to ensure that the network focused on the alphabetical linguistic patterns of the medical symptoms. Fourth, the standard punctuation symbols were also removed to simplify the subsequent tokenization process.

Fifth, the whitespace normalization was applied to collapse the multiple sequential whitespaces into a single space and stripped any leading or trailing spaces.

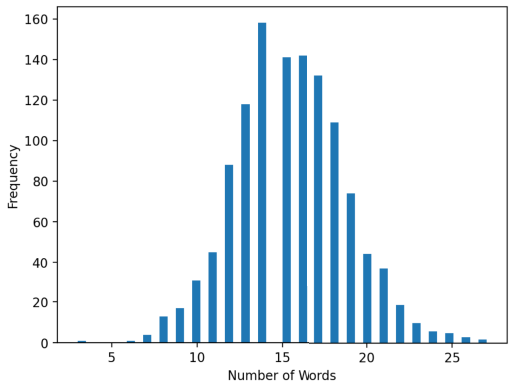
Following the text preprocessing process, the cleaned corpus was fed into the Keras Tokenizer to map the processed texts into a sequence of integer indices. The tokenizer was initially set to a maximum vocabulary size of 10,000 tokens to guarantee that the clinical vocabulary would be sufficiently large enough to capture all the terms in the dataset.

These tokenized sequences were subsequently padded to a uniform maximum length of 200 tokens. The raw text was transformed into structured and uniform numerical arrays that were ready for the sequence modeling.

However, after the statistical analysis of the dataset, it was found that the entire cleaned corpus contained only 1,374 unique clinical terms and the absolute maximum text length did not exceed 27 words. Consequently, maintaining the initial parameters introduced massive matrix sparsity and unnecessary computational overhead due to excessive zero-padding. The tokenization parameters were optimized by setting the vocabulary size to 1,500 words, and the sequences were padded to a uniform maximum length of 30 tokens.

Figure 4 illustrates the text lengths in the pre-processed dataset. The dataset showed a normal distribution with a mean between 14 and 17 words. It also has an absolute maximum of 27 words. This testified the selection of the model padding threshold of 30 tokens.

Figure 4
Distribution of text lengths



4.2. Classification using the bidirectional LSTM model

In the study, the performance of the proposed Bi-LSTM model was evaluated in two distinct phases. The baseline model utilizing the default hyperparameters of the architecture as presented in the above Figure 3 and the fine-tuned model that adopted the domain-specific clinical text preprocessing pipeline with the optimized architecture. Table 4 summarizes the preprocessing and architectural configurations used for the performance evaluation of the proposed model.

In the case of the fine-tuned model, the vocabulary size was adjusted to 1,500 words, the embedding dimension was reduced to 64, and the maximum sequence length was optimized to 30 tokens. The preprocessing pipeline was included with the context-aware stop-word removal to preserve clinical negations (*e.g.*, "no," "not") while filtering out the non-diagnostic noise (*e.g.*, "patient," "clinic") with lemmatization and punctuation handling to protect the alphanumeric conditions like "Covid-19." The bidirectional LSTM layers were expanded to 100 neurons each, and the SpatialDropout1D rate was slightly reduced to 0.2 to retain more feature maps.

The results presented in Tables 5 and 6 indicate the baseline model's performance. The performance of the Bi-LSTM model was assessed using precision, recall, and F1-score, with the per-class metrics displayed in Table 5.

The results in Table 5 showed that the model was able to yield high accuracy for 10 classes of diseases. It predicted these classes with precision, recall, and F1-score of 1.00. The 10 diseases included: (1) varicose veins; (2) peptic ulcer disease; (3) drug reaction; (4) malaria; (5) hypertension; (6) bronchial asthma; (7) common cold; (8) fungal infection; (9) dengue; and (10) typhoid. These findings demonstrated the model's effectiveness in learning the distinct symptom patterns for these diseases.

Nevertheless, the unoptimized Bi-LSTM model classified the gastroesophageal reflux disease and pneumonia with lower performance metrics as compared to the other classes. The gastroesophageal reflux disease achieved a precision, recall, and F1-score of 0.82, 0.90, and 0.86, respectively, while the pneumonia recorded 0.75, 0.90, and 0.82. This reduction in precision can be attributed to the overlapping symptoms representations across the respiratory and the gastrointestinal disorders that introduced classification ambiguity.

Similarly, minor variations and misclassifications were observed in the conditions such as the urinary tract infection which exhibited the lowest recall at 0.60 and the impetigo with a recall of 0.70. These results highlighted a specific limitation of the Bi-LSTM model. It struggled to distinguish diseases that share overlapping symptom descriptions.

As a whole, the model was able to accomplish accuracy of 95.00%. The consistency of the macro and weighted average scores is presented in Table 6. As indicated in Table 6, precision and recall both yielded 0.95, and the F1-score was 0.94 for both macro and weighted averages.

The consistent macro and weighted averages suggested that the Bi-LSTM model did not display significant bias toward any class. The results in Table 6 further highlighted the robustness of the Bi-LSTM architecture in handling the multi-class clinical text classification.

The training and validation loss curves are presented in Figure 5. These curves assessed the learning and generalization capabilities of the Bi-LSTM model. The *x*-axis represented the number of training epochs. The *y*-axis denoted the categorical cross-entropy loss. The training loss (blue line) progressively decreased as the training epochs increased. The training loss stabilized around epoch 20. Additional training epochs did not yield any improvement in performance. This is because the model has reached a local optimum.

The validation loss (orange line) also followed a downward trend. It was observed that the loss on the validation set was slightly higher than the loss on the training set during the first

Table 4
Comparison of baseline and fine-tuned model configurations

Component	Baseline	Fine-Tuned
Text preprocessing	Basic cleaning	Domain-specific
Vocabulary size	10 000 tokens	1 500 tokens
Sequence length	200 tokens	30 tokens
Embedding dimension	128	64
SpatialDropout1D	0.3	0.2
Bi-LSTM layers	2 layers *x* 64 neurons	2 layers *x* 100 neurons
Dense layer	64 neurons (ReLU activation)	64 neurons (ReLU activation)
Dropout	0.5	0.5

Table 5
The evaluation metrics per class

Class	Precision	Recall	F1-score	Support
Psoriasis	0.91	1.00	0.95	10
Varicose veins	1.00	1.00	1.00	10
Peptic ulcer disease	1.00	1.00	1.00	10
Drug reaction	1.00	1.00	1.00	10
Gastroesophageal reflux disease	0.82	0.90	0.86	10
Allergy	0.91	1.00	0.95	10
Urinary tract infection	0.86	0.60	0.71	10
Malaria	1.00	1.00	1.00	10
Jaundice	0.91	1.00	0.95	10
Cervical spondylosis	0.91	1.00	0.95	10
Migraine	1.00	0.90	0.95	10
Hypertension	1.00	1.00	1.00	10
Bronchial asthma	1.00	1.00	1.00	10
Acne	1.00	0.90	0.95	10
Arthritis	0.83	1.00	0.91	10
Dimorphic hemorrhoids	1.00	0.90	0.95	10
Pneumonia	0.75	0.90	0.82	10
Common cold	1.00	1.00	1.00	10
Fungal infection	1.00	1.00	1.00	10
Dengue	1.00	1.00	1.00	10
Impetigo	1.00	0.70	0.82	10
Chicken pox	0.91	1.00	0.95	10
Typhoid	1.00	1.00	1.00	10
Diabetes	1.00	0.90	0.95	10

Table 6
The macro and weighted average of precision, recall, and F1-score

	Precision	Recall	F1-score	Support
Macro average	0.95	0.95	0.94	240
Weighted average	0.95	0.95	0.94	240

epochs (epochs 0-10). The validation curve was flat during epochs 15 to 25.

Following this, Figure 6 depicts the training and validation accuracy curves during the training of the Bi-LSTM model. The number of epochs was plotted on the *x*-axis and accuracy on the *y*-axis.

As indicated in Figure 6, the training accuracy (blue line) rose rapidly during the first 6 epochs, surpassed the 90% threshold. Subsequently, it reached high accuracy by epoch 10. It remained consistently high, which suggested that the Bi-LSTM model was efficient at capturing the long-term dependencies.

Figure 5
The training and validation loss of the baseline Bi-LSTM model

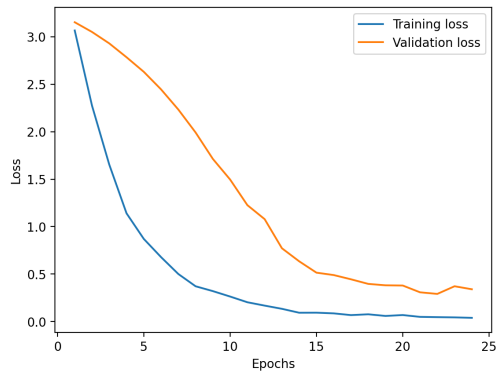
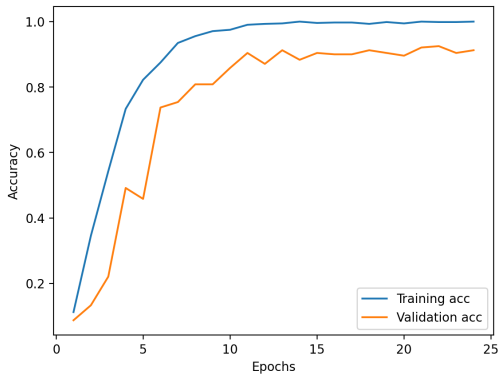


Figure 6
The training and validation accuracy of the baseline Bi-LSTM model



The validation accuracy (orange line) also increased sharply in the initial phase (epochs 0 to 7). After epoch 10, the validation accuracy entered the steady-state phase, indicating that the model did not exhibit overfitting and was able to generalize with unseen test instances.

The results presented in Tables 7 and 8 show the performance of the fine-tuned Bi-LSTM model utilizing the 5-fold stratified cross-validation. As shown in Table 7, the optimized Bi-LSTM model accurately predicted 8 classes of diseases with perfect precision, recall, and F1-scores of 1.00. The 8 diseases included: (1) Malaria; (2) Jaundice; (3) Cervical spondylosis; (4) Acne; (5) Arthritis; (6) Dimorphic hemorrhoids; (7) Pneumonia; (8) Impetigo.

The precision for the gastroesophageal reflux disease and the pneumonia improved significantly to 0.93 and 1.00, respectively. This indicates that the linguistic noise was removed with the help of the context-aware preprocessing. Similarly, the recall for the urinary tract infection which previously struggled at 0.60 achieved a score of 0.96 under this fine-tuned model by reducing false negatives for this condition.

Despite these improvements, the fine-tuned Bi-LSTM model still exhibited minor variations in some specific classes. For instance, the dengue yielded the lowest precision of 0.89 under cross-validation, indicating minor false positives. Meanwhile, the allergy recorded the lowest recall score at 0.86, and the chicken pox showed the reduction in the recall score from 1.00 to 0.92. These metrics highlighted a specific challenge. Medical conditions with similar visual or textual descriptions are difficult to differentiate by the model.

Overall, the results in Table 8 indicated that the fine-tuned Bi-LSTM model using the 5-fold stratified cross-validation accomplished an accuracy of 98.00%. Both the macro and weighted average scores were highly consistent, achieving 0.98 for precision, recall, and the F1-score. These results showed that the vocabulary size, sequence padding, and increased recurrent capacity of the fine-tuned model can maximize computation efficiency as well as improve the overall classification results.

The training and validation loss as well as the accuracy curves for the fine-tuned Bi-LSTM model utilizing the 5-fold stratified cross-validation are presented in Figures 7 and 8, respectively. As shown in Figure 7, the fine-tuned Bi-LSTM improved in reaching optimal generalization threshold at a faster pace because it was able to perform effective text preprocessing across all 5 folds within the first 10 epochs.

As shown in Figure 8, the early stopping mechanism was able to detect the validation loss which had reached a plateau and could not be improved anymore. Therefore, the training process stopped early to restore the best model weights and prevented overfitting.

4.3. Model explanation

This section presents the results of the Bi-LSTM model's local interpretations obtained by using the LIME technique. The LIME technique was chosen because it is model-agnostic, meaning that it can be applied to the Bi-LSTM model without understanding its internal architecture. The instance 30 of Psoriasis class, the instance 100 of typhoid class, and the instance

Table 7
The evaluation metrics per class across 5-folds

Class	Precision	Recall	F1-score	Support
Psoriasis	0.92	0.96	0.94	50
Varicose veins	0.98	1.00	0.99	50
Peptic ulcer disease	0.96	0.96	0.96	50
Drug reaction	0.96	0.96	0.96	50
Gastroesophageal reflux disease	0.93	1.00	0.96	50
Allergy	1.00	0.86	0.92	50
Urinary tract infection	1.00	0.96	0.98	50
Malaria	1.00	1.00	1.00	50

(*Continued*)

Table 7
(*Continued*)

Class	Precision	Recall	F1-score	Support
Jaundice	1.00	1.00	1.00	50
Cervical spondylosis	1.00	1.00	1.00	50
Migraine	0.98	0.98	0.98	50
Hypertension	0.96	1.00	0.98	50
Bronchial asthma	0.96	1.00	0.98	50
Acne	1.00	1.00	1.00	50
Arthritis	1.00	1.00	1.00	50
Dimorphic hemorrhoids	1.00	1.00	1.00	50
Pneumonia	1.00	1.00	1.00	50
Common cold	0.98	0.98	0.98	50
Fungal infection	1.00	0.98	0.99	50
Dengue	0.89	0.94	0.91	50
Impetigo	1.00	1.00	1.00	50
Chicken pox	0.98	0.92	0.95	50
Typhoid	0.96	0.96	0.96	50
Diabetes	0.98	0.96	0.97	50

Table 8
The macro and weighted average of precision, recall, and F1-score across 5-folds

	Precision	Recall	F1-score	Support
Macro average	0.98	0.98	0.98	1200
Weighted average	0.98	0.98	0.98	1200

Figure 7
The training and validation loss of the baseline Bi-LSTM model across 5-folds

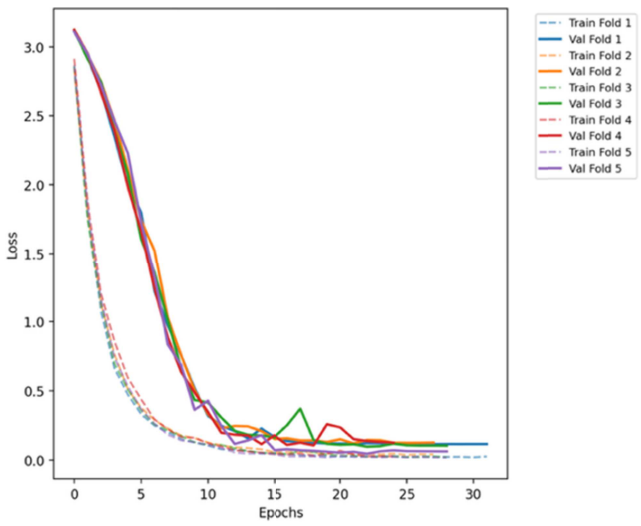
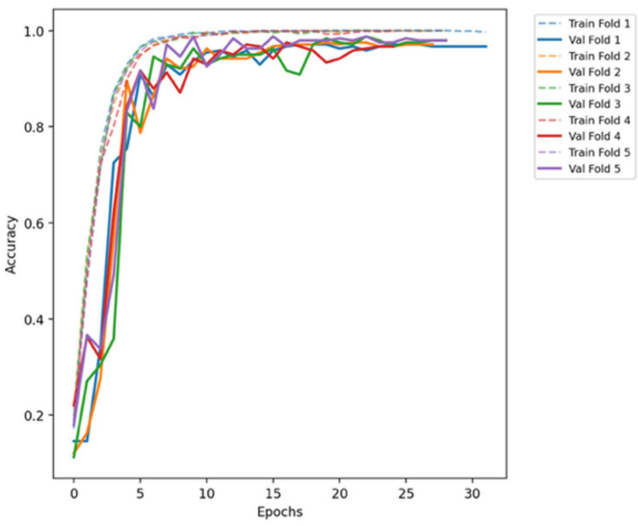


Figure 8
The training and validation accuracy of the fine-tuned Bi-LSTM model across 5-folds



400 of pneumonia class were chosen randomly for local explanations. The LIME technique was able to present the features that contributed the most to the target outcome of Psoriasis, Typhoid, and Pneumonia, respectively.

The feature importance scores for psoriasis class between the baseline model and the fine-tuned model are presented in Figures 9 and 10, respectively.

As depicted in Figure 9, the baseline model predicted the psoriasis class with a high probability of 1.00. The model identified several key anatomical markers such as "joints" (0.11), "elbows" (0.11), "knees" (0.09), and "skin" (0.08). The LIME explanations also revealed weaknesses in the unoptimized processing pipeline. For instance, the baseline model relied on ubiquitous grammatical stop-words, specifically "on" (0.06), "in" (0.06), and "is" (0.05).

Figure 9
Local explanation of baseline Bi-LSTM prediction for Psoriasis

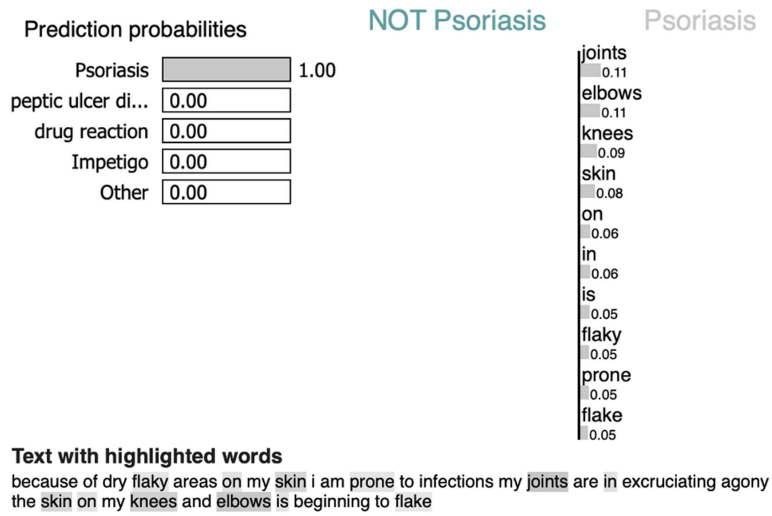
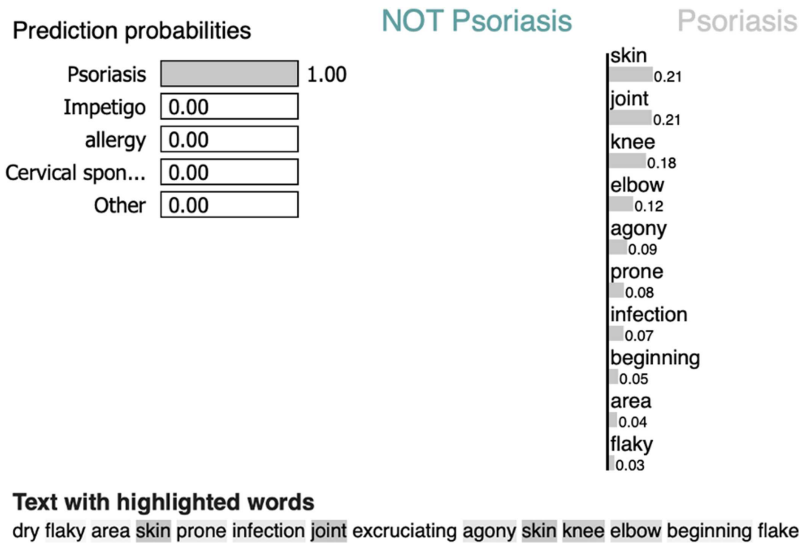


Figure 10
Local explanation of fine-tuned Bi-LSTM prediction for Psoriasis



This indicates that the baseline model was learning linguistic structures rather than pathological characteristics. It is also not reliable if the model depended on irrelevant words such as "is" or "in" for diagnosis in the clinical settings.

On the other hand, the fine-tuned model had a positive effect on the local explanation of the Psoriasis class as shown in Figure 10. It can be observed that the grammatical stop-words were filtered out, indicating that the fine-tuned model's prediction was driven by the clinical descriptors. The effect of lemmatization was clearly visualized. Plural terms such as ("joints," "elbows," and "knees") have been standardized to their root forms ("joint," "elbow," "knee"). The "skin" and "joint" emerged as the highest-weighted features (both at 0.21), followed closely by the localized anatomical markers "knee" (0.18) and "elbow" (0.12).

The output of the LIME technique as presented in Figure 10 also captured the additional contextual and pathological descriptors that reinforced the psoriasis prediction. For instance, the clinical narrative elements "agony" (0.09), "prone" (0.08), and "infection" (0.07) acted as positive contributors alongside the lower-weighted features like "beginning" (0.05), "area" (0.04), and "flaky" (0.03). The LIME analysis in Figure 10 also validated the methodological improvements of the fine-tuned model to prioritizing the precise dermatological and rheumatological entities that clinically defined psoriasis as well as ensured diagnostic transparency and medical trustworthiness.

The LIME analysis presented in Figures 11 and 12 below highlights that the baseline model used to rely on the stop-word "been" and the fine-tuned model corrected this by entirely filtering out the linguistic noise and shifting weights toward a clearer clinical narrative featuring terms like "belly," "constipation," "pain," and "chill."

As shown in Figure 11, the baseline Bi-LSTM model predicted the Typhoid with a probability of 1.00. While the model correctly identified gastrointestinal markers such as "belly" (0.16) and "constipation (0.08)", the distribution of the feature weights revealed significant vulnerability to the linguistic noise. The

Figure 11
Local explanation of baseline Bi-LSTM prediction for Typhoid

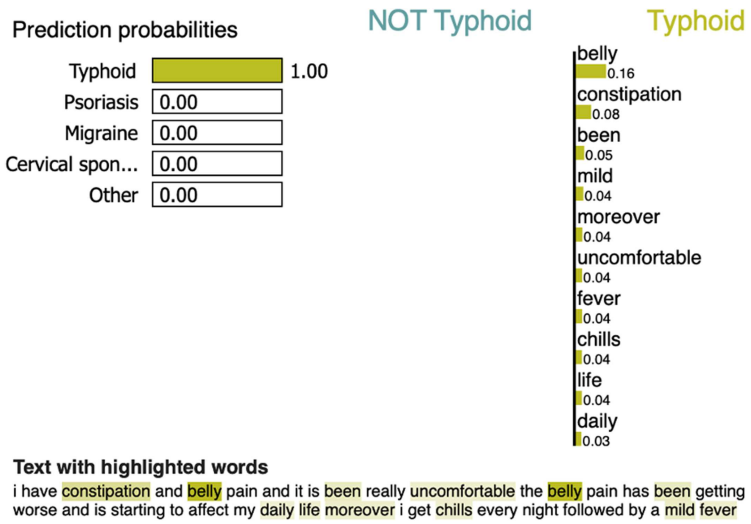
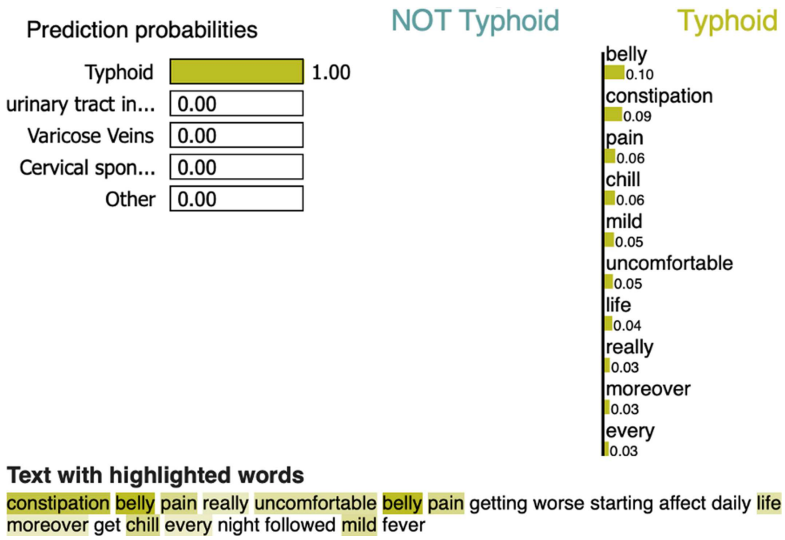


Figure 12
Local explanation of fine-tuned Bi-LSTM prediction for Typhoid



grammatical stop-word "been" (0.05) emerged as the third most important feature contributing to the prediction of the Typhoid, outweighing actual systemic clinical symptoms like "fever" (0.04) and "chills" (0.04). Additionally, the non-diagnostic filler words or adverbs such as "mild" (0.04), "moreover" (0.04), "uncomfortable" (0.04) and "life" (0.04), and "daily" (0.03) heavily cluttered the model's decision-making process. According to these results, the baseline model was prioritizing irrelevant syntactic structures over the actual pathological information. The following Figure 12 illustrates the ability of the fine-tuned model to focus on the relevant clinical descriptors.

The weights for the primary diagnostic features had been increased. The grammatical noise like "been," "fever," "chills," and "daily" dropped out of the top features. The feature "belly" remained the most dominant contributor at 0.10, followed closely by "constipation" at 0.09. The removal of the underlying noise allowed the critical symptom "pain" (0.06) and the lemmatized root form "chill" (0.06) to emerge directly underneath the gastrointestinal descriptors. While a few contextual modifiers and filler elements remained such as "mild" (0.05), "uncomfortable" (0.05), "life" (0.04), "really" (0.03), "moreover" (0.03), and "every" (0.03).

Figure 12 indicated that by preventing the model to rely on grammatical artifacts like "been," the fine-tuned Bi-LSTM model was able to direct its diagnostic confidence on evidence-based symptomatology.

The baseline Bi-LSTM model predicted the pneumonia with a high probability of 1.00 as shown in Figure 13. The examination of the feature weights revealed severe diagnostic vulnerabilities due to unoptimized text processing. While the baseline model captured the relevant clinical markers like "heart" (0.25), "phlegm" (0.23), and "color" (0.15), it incorrectly split its mathematical weight between the syntactic variations of the same symptom, specifically "cough" (0.14) and "coughing" (0.13). This was due to the baseline pipeline lacking lemmatization which treated these as distinct features rather than a single unified symptom.

Figure 13
Local explanation of baseline Bi-LSTM prediction for Pneumonia

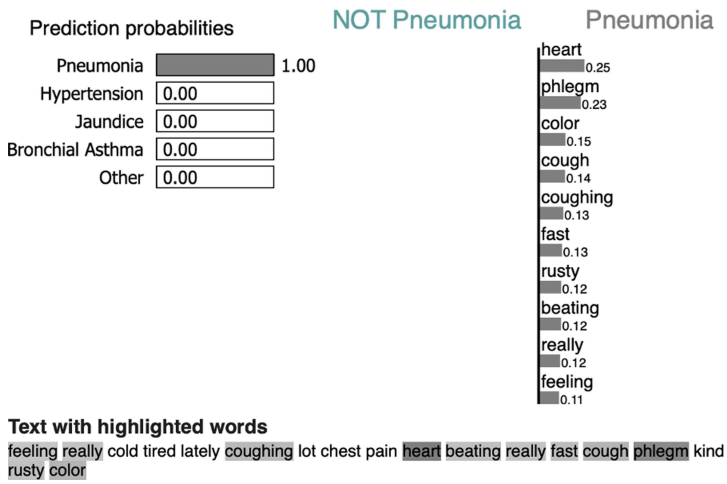
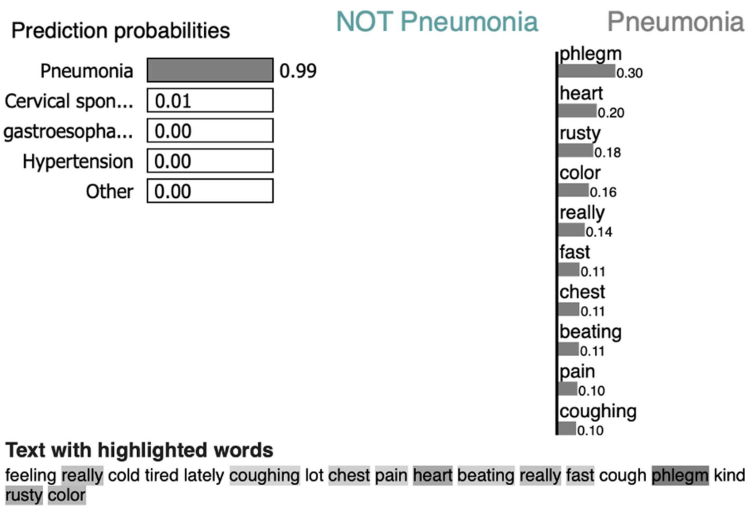


Figure 14
Local explanation of fine-tuned Bi-LSTM prediction for Pneumonia



Furthermore, the conversational filler words and adverbs such as "really" (0.12) and "feeling" (0.11) diluted the model's focus, competing closely with highly specific pathological features like "rusty" (0.12), "beating" (0.12), and "fast" (0.13).

Figure 14 shows that the weight of the primary symptom, "phlegm" increased from 0.23 to 0.30. The removal of the linguistic noise uncovered critical contextual symptoms that were entirely suppressed in the baseline model. Specifically, the emergence of "heart" (0.20), "fast" (0.11), and "beating" (0.11) illustrated the model's capacity to aggregate weights around multi-word physiological concepts like a "fast-beating heart." Similarly, the emergence of localized descriptors such as "chest" (0.11) and "pain" (0.10) captured classic diagnostic presentations of the infection. Additionally, the identification of "rusty" (0.18) and "color" (0.16) indicated that the model's ability to recognize sputum descriptors associated to the bacterial pneumococcal infections.

Results presented in Figures 9, 11, and 13 demonstrated that the explanations produced by the LIME technique were accurate, but not necessarily clinically reliable. It sometimes gave high feature importance to linguistic artifacts and grammatical structures as shown by the baseline bidirectional long short-term memory (Bi-LSTM) model.

On the other hand, if a model was too dependent on the linguistic artifacts, it would suffer from poor performance in real-world clinical settings with varying patient phrasing. The LIME technique was also used as a diagnostic tool in the development process to identify the text preprocessing errors and the fine-tuned Bi-LSTM model to correct them.

It is also equally important to know that the outcomes demonstrated the importance of domain expert oversight. The XAI outputs required evaluation by medical professionals. The evaluation by the medical professionals is crucial to ensure that the model's decision mainly relies on the pathognomonic evidence rather than the linguistic artifacts.

4.4. Evaluation on additional datasets

The proposed Bi-LSTM model was also reevaluated by using two other datasets obtained from the Kaggle to verify its

robustness and generalizability. First, the Dataset-2 which was the medical transcriptions dataset sourced from the mtsamples.com comprised 5,000 anonymized and synthetic clinical notes. This dataset included the full-text medical transcriptions, their corresponding medical specialties, short descriptions, and relevant keywords. Second, the Dataset-3 which was the symptoms-based medical test recommendations dataset that consisted of 1,498 records structured across the following three primary variables: (1) questions; (2) corresponding target diseases; and (3) recommended medical tests.

The results presented in Tables 9 and 10 showed the performance of the fine-tuned Bi-LSTM model based on the Dataset-2. A 5-fold stratified cross-validation was performed for the 40 medical specialty classes to test the model. The test showed that model was able to achieve accuracy of 76.18%. Although the Dataset-2 exhibited class imbalance with certain medical specialties heavily

Table 9
The evaluation metrics per class across 5-folds for Dataset-2

Class	Precision	Recall	F1-score	Support
Allergy	1.00	0.45	0.62	1103
Autopsy	0.10	1.00	0.18	1103
Bariatrics	1.00	1.00	1.00	1103
Cardiovascular / Pulmonary	1.00	0.77	0.87	1103
Chiropractic	1.00	0.28	0.43	1103
Consult – History and physical	1.00	0.44	0.61	1103
Cosmetic / Plastic surgery	1.00	0.93	0.96	1103
Dentistry	1.00	0.88	0.94	1103
Dermatology	1.00	0.84	0.91	1103
Diets and nutrition	1.00	1.00	1.00	1103
Discharge summary	1.00	0.68	0.81	1103
ENT – Otolaryngology	1.00	0.87	0.93	1103
Emergency room reports	1.00	0.37	0.54	1103
Endocrinology	1.00	0.81	0.90	1103
Gastroenterology	1.00	0.87	0.93	1103
General medicine	1.00	0.55	0.71	1103
Hematology – Oncology	1.00	0.67	0.80	1103
Hospice – Palliative care	1.00	0.82	0.90	1103
IME-QME-Work Comp	1.00	0.23	0.38	1103
Lab medicine – Pathology	1.00	1.00	1.00	1103
Letters	1.00	0.88	0.93	1103
Nephrology	1.00	0.76	0.87	1103
Neurology	1.00	0.73	0.84	1103
Neurosurgery	1.00	0.85	0.92	1103
Obstetrics / Gynecology	1.00	0.83	0.91	1103
Office notes	0.69	0.88	0.78	1103
Ophthalmology	1.00	0.95	0.98	1103
Orthopedic	1.00	0.82	0.90	1103
Pain management	1.00	0.95	0.98	1103
Pediatrics – Neonatal	1.00	0.55	0.71	1103
Physical medicine - Rehab	1.00	0.51	0.67	1103
Podiatry	1.00	0.90	0.95	1103
Psychiatry / Psychology	1.00	0.35	0.51	1103
Radiology	1.00	0.92	0.96	1103
Rheumatology	1.00	0.67	0.81	1103
SOAP / Chart / Progress notes	1.00	0.86	0.92	1103
Sleep medicine	1.00	0.90	0.95	1103
Speech - Language	1.00	0.88	0.93	1103
Surgery	1.00	0.93	0.96	1103
Urology	1.00	0.89	0.94	1103

Table 10
The macro and weighted average of precision, recall, and F1-score across 5-folds for Dataset-2

	Precision	Recall	F1-score	Support
Macro average	0.97	0.76	0.82	44 120
Weighted average	0.97	0.76	0.82	44 120

overrepresented, the synthetic minority over-sampling technique (SMOTE) was used in the preprocessing phase to prevent the model from becoming biased towards the majority classes.

As indicated in Table 9, the Bi-LSTM model achieved a high F1-score ($F1 \geq 0.95$) for the following eight specific classes: (1) Bariatrics; (2) Diets and Nutrition; (3) Lab Medicine – Pathology; (4) Ophthalmology; (5) Pain Management; (6) Podiatry; (7) Radiology; and (8) Surgery. These 8 classes contained distinct vocabulary, procedure names, and anatomical terms, and they allowed the model to distinguish them from the other medical specialties.

The other specialties that indicated low recall included the independent medical examinations, the qualified medical evaluations, and the workers' compensation (IME-QME-Work Comp) (0.23) and the chiropractic (0.28). These categories overlapped with the broader classes like the orthopedic and the pain management. The independent medical examinations (IME) covered the musculoskeletal complaints. Hence, the text often resembled standard orthopedic consults.

Additionally, the psychiatry / psychology and the emergency room reports classes showed low recall of 0.35 and 0.37, respectively. These document types frequently shared structural phrasing and symptoms with the general medicine and the consult – history and physical classes which may also lead to false negatives prediction.

Table 10 showed that the model yielded macro and weighted average of 0.97 (precision), 0.76 (recall), and 0.82 (F1-score). The precision of 0.97 denoted low false-positive rates. For the moderate recall of 0.76, this figure seemed to suggest that the model was impacted by ambiguous or overlapping specialty categories.

The results of the performance of the fine-tuned Bi-LSTM model using the Dataset-3 are presented in Tables 11 and 12. The model was also evaluated by adopting the 5-fold stratified cross-validation across the 30 classes of diseases.

According to the results in Table 11, there were several classes that were correctly classified with perfect classification metrics scores (Precision = 1.00, Recall = 1.00, F1-score = 1.00). The classes included: (1) asthma; (2) breast cancer; (3) colorectal cancer; (4) prostate cancer; and (5) skin cancer. They were categorized correctly because the patient-reported symptoms for these classes contained highly distinct keyword markers. There is however a slight decrease in the performance for the cardiovascular conditions. For instance, the myocardial infarction yielded the lowest precision of 0.90 and recall of 0.88. In the same way, arrhythmias and coronary artery disease showed minor reduction in the F1-score of 0.91 and 0.94, respectively. These occurred due to the conditions that frequently overlapped and shared non-specific presentations such as in cases such as chest pain, shortness of breath, fatigue, and palpitations. The lowest performing categories in the dataset were chronic obstructive and pulmonary disease which obtained F1-score of 0.80 and 0.81, respectively.

As shown in Table 12, evaluating the model using the Dataset-3 yielded macro and weighted averages of 0.95 for precision, recall, and F1-score. These results illustrated that the model was able to match the patient-reported symptoms to the correct disease with an accuracy of 95.26%.

Table 11
The evaluation metrics per class across 5-folds for Dataset-3

Class	Precision	Recall	F1-score	Support
Addison's Disease	0.94	0.90	0.92	50
Arrhythmias	0.94	0.88	0.91	50
Asthma	1.00	1.00	1.00	50
Atherosclerosis	1.00	0.96	0.98	50
Brain cancer	0.96	1.00	0.98	50
Breast cancer	1.00	1.00	1.00	50
COVID-19	0.98	0.96	0.97	50
Cholera	0.98	0.98	0.98	50
Chronic Obstructive	0.80	0.80	0.80	50
Colorectal Cancer	1.00	1.00	1.00	49
Coronary Artery Disease	0.96	0.92	0.94	50
Dengue fever	0.98	0.96	0.97	50
Diabetes	0.96	0.94	0.95	50
Ebola virus disease	0.92	0.94	0.93	50
HIV / AIDS	0.94	1.00	0.97	50
Myocardial Infarction	0.90	0.88	0.89	49
Hepatitis	0.94	0.98	0.96	50

(*Continued*)

Table 11
(*Continued*)

Class	Precision	Recall	F1-score	Support
Hypertension	0.98	0.98	0.98	50
Hyperthyroidism	0.94	0.98	0.96	50
Hypothyroidism	0.96	0.98	0.97	50
Influenza	1.00	0.90	0.95	50
Interstitial lung disease	0.89	0.96	0.92	50
Leukemia	0.98	0.98	0.98	50
Lung cancer	1.00	0.98	0.99	50
Malaria	0.86	0.96	0.91	50
Prostate cancer	1.00	1.00	1.00	50
Pulmonary disease	0.80	0.82	0.81	50
Skin cancer	1.00	1.00	1.00	50
Stroke	1.00	0.96	0.98	50
Tuberculosis	1.00	0.98	0.99	50

Table 12
The macro and weighted average of precision, recall, and F1-score across 5-folds for Dataset-3

	Precision	Recall	F1-score	Support
Macro average	0.95	0.95	0.95	1498
Weighted average	0.95	0.95	0.95	1498

4.5. Comparison with previous studies

The performance of the Bi-LSTM model proposed in this study was also compared to the other models documented in previous studies. As an example, the study by Hassan et al. [29] also adopted the Bi-LSTM model for disease prediction using the identical Symptom2disease dataset. The Bi-LSTM model in the study which optimized the Hyperopt Bayesian hyperparameter search algorithm achieved accuracy, F1-score, recall, and precision of 97.08%, 97.05%, 97.08%, and 97.37%, respectively. However, in comparison, the fine-tuned Bi-LSTM model developed in the present study which relied on the domain-specific preprocessing pipeline was able to accomplish a marginally higher accuracy, precision, recall, and F1-score of 98.00% using the 5-fold stratified cross-validation. These results indicated that the preprocessing pipeline adopted in this study was able to perform better than the independently hyperparameter-optimized Bi-LSTM model that adopted the same dataset.

The study by [29] also evaluated the transformer-based MCN-BERT model by using the same dataset and attained accuracies of 98.33% and 99.58% with the AdamW and AdamP optimizers, respectively. Although the transformer-based models were able to achieve marginally higher accuracy, their architecture appeared to have more trainable parameters and higher computational overhead. However, unlike the MCN-BERT model, the architecture of the Bi-LSTM model as proposed in the present study required substantially fewer trainable parameters and considerably less computational overhead. The proposed Bi-LSTM and LIME framework presented in this study was not only comparably accurate but also considered to be a more lightweighted alternative to the transformer-based approaches which may be also advantageous when adopted in resource-constrained clinical settings.

In terms of the model interpretability, the findings of this research were also consistent with those studies conducted by [24] and [25]. Findings of these studies also found that the LIME technique was able to offer clinically meaningful local explanations for categorizing patient safety event reports and multiple sclerosis MRI reports, respectively. These previous studies however mainly utilized the technique as a post-hoc explanatory tool. Unlike the prior studies, the present study adopted and expanded the LIME technique as an active diagnostic instrument during the model development phase. The linguistic artifacts (e.g., stop-words "been," "is," and "in") identified through the LIME explanations of the baseline model were subsequently corrected through the domain-specific preprocessing pipeline of the fine-tuned model as presented in Section 4.3. This iterative approach differed from the studies by [26–28], which revealed that despite the ability of the LIME technique to identify important contributing features, the interpretability findings were not systematically used to refine the underlying text classification models.

As far as the comparison of the classification performance of the various models is concerned, the Bi-LSTM model developed in this study appeared to be compatible to the performance of the Bi-LSTM and the transformer-based models as documented in the existing literature. In addition, the systematic use of the LIME technique as a diagnostic tool during the model development represented a methodological advancement over the post-hoc interpretability approaches adopted in previous studies [24–28].

5. Conclusion

In an attempt to classify different classes of diseases, this study proposed and tested the Bi-LSTM model which combined with the LIME technique. The framework used to develop the classification model consisted of the following stages: preprocessing, classification, and model explanation.

The proposed model achieved 98.00% across precision, recall, F1-score, and accuracy. The generalizability and robustness of the model were further tested using two additional datasets. The model attained an accuracy of 76.18%, precision of 0.97, and recall of 0.76 when it utilized the 40-class medical transcription dataset. These slightly lower results indicated that overlapping and ambiguous clinical categories negatively have impacted the classification. However, the adoption of the 30-class symptom-based disease dataset provided the model with a stronger accuracy of 95.26% with precision, recall, and F1-score of 0.95. These results suggested that the proposed model generalized well to the datasets with clearer class boundaries. However, its performance decreased when utilized datasets with greater class overlap and imbalance.

Taken together, the findings of this study unveiled that the performance of the Bi-LSTM model proposed in this study was not only marginally higher than the previously reported Bi-LSTM models that optimized automated hyperparameter search but also comparable to the transformer-based MCN-BERT models.

Additionally, the findings also indicated that the LIME technique adopted in this study was able to provide local explanation by determining the feature importance scores of words that contributed to the class prediction of the specific instance. The study also discovered that the LIME technique had the ability to enhance the interpretability of the Bi-LSTM model, which initially possessed low transparency for clinical insights. Most importantly, applying the LIME technique enabled the study to identify and correct problematic linguistic artifacts within the preprocessing pipeline. These findings also presented as a methodological advancement that has not been addressed in the literature and had mainly viewed the LIME technique primarily as a post-hoc explanatory tool.

Nonetheless, there were also several limitations in the study. As an example, the dataset used in the study consisted of 1,200 instances (50 per class) and yielded a rather small test set of 10 samples per class. Although the 5-fold stratified cross-validation reached accuracy of 98.00%, the overall dataset size restricted the generalizability of the model. In the study, the issue of generalizability was addressed through the reevaluation of the model by using two other larger datasets. Of the two datasets, the medical transcription dataset indicated that class imbalance and overlapping clinical categories presented as a challenge to improving the generalizability of the model.

The contributions of this work included its integration within the domain-specific clinical preprocessing pipeline for the multi-class disease classification and also in the adoption of the LIME technique as an active diagnostic tool during the model development rather than solely using it as a post-hoc explainer.

Given the limitations, this work should also be considered as a proof-of-concept study rather than a clinically validated system. As such, the claim of clinical readiness should be withheld until the study can be further validated by using more independent and larger-scale clinical sources of datasets.

The emphasis on transparency and interpretability becomes increasingly more important as AI models evolve and become more complex in the healthcare industry. In addition to the need to focus on transparency and interpretability, future research should extend the proposed framework by incorporating multi-modal datasets. For instance, clinical text can be combined with medical imaging to enhance classification accuracy. Furthermore, the future proposed framework should also be validated on larger and more diverse clinical datasets. Those datasets with greater class imbalance can help to establish generalizability. Last but not least, when adopting the Bi-LSTM approach, researchers should also attempt to compare this approach with the other transformer-based models such as the ClinicalBERT and the BioBERT.

Acknowledgement

The authors would like to express their gratitude to the Faculty of Artificial Intelligence, Universiti Teknologi Malaysia (UTM), Kuala Lumpur for their support and assistance throughout this research.

Funding Support

The authors would like to thank Universiti Teknologi Malaysia for sponsoring this study under the Geran Penyelidik Baru (GPB) research grant no.: Q.K130000.2757.04K71.

Ethical Statement

The authors declare that this study did not require formal ethical approval because Universiti Teknologi Malaysia does not require Research Ethics Committee (UTM REC) approval for research that does not involve human or animal subjects. This exemption is based on the UTM Research Ethics Guidelines issued by the UTM REC. The dataset used in this study was obtained from a public repository (Kaggle) and contains no personally identifiable information.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in Kaggle at https://www.kaggle.com/datasets/niyarrbarman/symptom2disease, https://www.kaggle.com/datasets/tboyle10/medicaltranscriptions, and https://www.kaggle.com/datasets/shoaibuddin/symptoms-based-medical-test-recommendations, reference number [30–32].

Author Contribution Statement

Adam Mohd Khairuddin: Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing, Visualization. **Siti Armiza Mohd Aris:** Writing – review & editing, Validation. **Azizul Azizan:** Writing – review & editing, Validation. **Noor Jannah Zakaria:** Writing – review & editing, Validation.

References

[1] Saravanan, V., Aneja, I., Yang, H., Pillai, A. S., & Singh, A. (2021). Impact of big data in healthcare system—A quick look into electronic health record systems. In K. K. Singh, M. Elhoseny, A. Singh, & A. A. Elngar (Eds.), *Machine learning and the internet of medical things in healthcare* (pp. 251–262). Academic Press. https://doi.org/10.1016/B978-0-12-821229-5.00009-4

[2] Moreira, L. H. (2025). Adoption of electronic health records in pathology: Challenges and opportunities. *Jornal Brasileiro de Patologia e Medicina Laboratorial*, *61*(1), 74–79.

[3] Madden, A., & Bekker, A. (2025). *Artificial intelligence for EHR: Use cases, costs, challenges*. ScienceSoft Healthcare.

[4] Rahman, S., Supto, M. M. U. B., Rahman, M. A., Rouf, R. A., & Hasan, M. M. (2025). Assessing the impact of electronic health records on health services. In *2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking*, 1–6. https://doi.org/10.1109/QPAIN66474.2025.11171835

[5] Bednorz, A., Mak, J. K., Jylhävä, J., & Religa, D. (2023). Use of electronic medical records (EMR) in gerontology: Benefits, considerations and a promising future. *Clinical Interventions in Aging*, 2171–2183. https://doi.org/10.2147/CIA.S400887

[6] Tayefi, M., Ngo, P., Chomutare, T., Dalianis, H., Salvi, E., Budrionis, A., & Godtliebsen, F. (2021). Challenges and opportunities beyond structured data in analysis of electronic health records. *Wiley Interdisciplinary Reviews: Computational Statistics*, *13*(6), e1549. https://doi.org/10.1002/wics.1549

[7] Rose, C., & Chen, J. H. (2024). Learning from the EHR to implement AI in healthcare. *NPJ Digital Medicine*, *7*(1), 330. https://doi.org/10.1038/s41746-024-01340-0

[8] Mamae, A. A., & Mamo, E. (2025). Assessment of implementation challenges of electronic medical record in Yekatit 12 hospital medical college. *PLOS One*, *20*(8), e0329896. https://doi.org/10.1371/journal.pone.0329896

[9] Makhni, S., Cerrato, P., Rico, J., Niazi, S., O'Horo, J., Peters, S., ... & Halamka, J. (2025). Meeting the challenges of electronic health record (EHR) optimization.npj Digital Medicine, *9*(1), 8. https://doi.org/10.1038/s41746-025-02178-w

[10] Gardner, R. L., Cooper, E., Haskell, J., Harris, D. A., Poplau, S., Kroth, P. J., & Linzer, M. (2019). Physician stress and burnout: the impact of health information technology. *Journal of the American Medical Informatics Association*, *26*(2), 106–114. https://doi.org/10.1093/jamia/ocy145

[11] Waghmare, I., Dad, R., Hole, D., Khare, K., Jadhav, R., & Waghmare, K. (2025). Unlocking clinical data: Structuring and summarizing unstructured medical records. In *International Conference On Innovative Computing And Communication*, 499–511. https://doi.org/10.1007/978-981-96-7523-4_37

[12] Jayashree, M., & Anitha, N. (2024). Framework for analyzing and predicting the unstructured clinical text data using natural language processing. In *International Conference on Data Science, Computation and Security*, 489–500. https://doi.org/10.1007/978-981-96-4880-1_39

[13] Chowdary, D. K., Jahnavi, D., & Kathambari, P. (2025). Developing an NLP-powered solution to automate the extraction and summarization of medical information from unstructured clinical notes. In *2025 8th International Conference on Circuit, Power & Computing Technologies (ICCPCT)*, 804–809. https://doi.org/10.1109/ICCPCT65132.2025.11176620

[14] Sivarajkumar, S., Edupuganti, S., Lazris, D., Bhattacharya, M., Davis, M., Dressman, D., . . . , & Wang, Y. (2026). Extraction of treatments and responses from non–small cell lung cancer clinical notes using natural language processing. *JCO Clinical Cancer Informatics*, *10*(1), e2500138. https://doi.org/10.1200/CCI-25-00138

[15] Lin, R. (2024). AI-driven personalized healthcare: Leveraging multimodal data for precision medicine. In *2024 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology*, 693–697. https://doi.org/10.1109/WI-IAT62293.2024.00112

[16] Gupta, G. (2023). Post hoc explanations for RNNs using state transition representations for time series data. In *Late-Breaking Work, Demos, Doctoral Consortium*, 249–255.

[17] Qin, C., Lin, K., & Huang, C. (2026). Beyond the black box: Precise temporal localization of arrhythmias via deep learning on class-balanced clinical datasets. *Biomedical Signal Processing and Control*, *120*, 110127. https://doi.org/10.1016/j.bspc.2026.110127

[18] Angiulli, F., De Luca, F., Fassetti, F., & Nisticò, S. (2025). LLiMe: Enhancing text classifier explanations with large language models. *Machine Learning*, *114*(12), 271. https://doi.org/10.1007/s10994-025-06920-z

[19] Angiulli, F., De Luca, F., Fassetti, F., & Nisticò, S. (2024). Large language models-based local explanations of text classifiers. In *International conference on discovery science*, 19–35. https://doi.org/10.1007/978-3-031-78977-9_2

[20] Ling, T., Jake, L., Adams, J., Osinski, K., Liu, X., & Friedland, D. (2023). Interpretable machine learning text classification for clinical computed tomography reports–A case study of temporal bone fracture. *Computer Methods and Programs in Biomedicine Update*, *3*, 100104. https://doi.org/10.1016/j.cmpbup.2023.100104

[21] Shukla, N., Singh, A. K., Dwivedi, V. K., Shukla, P., Srivastava, J., & Srivastava, V. (2025). An LSTM-oriented approach for next word prediction using deep learning. *AI-Based Advanced Optimization Techniques for Edge Computing*, 181–202. https://doi.org/10.1002/9781394287062.ch7

[22] Krichen, M., & Mihoub, A. (2025). Long short-term memory networks: A Comprehensive Survey. *AI*, *6*(9), 215. https://doi.org/10.3390/ai6090215

[23] Madi, I. A. E., Redjdal, A., Bouaud, J., & Seroussi, B. (2024). Exploring explainable AI techniques for text classification in healthcare: A scoping review. In *Digital Health and Informatics Innovations for Sustainable Health Care Systems: Proceedings of MIE 2024*, 846–850. https://doi.org/10.3233/SHTI240544

[24] Chen, H., Cohen, E., Wilson, D., & Alfred, M. (2024). A machine learning approach with human-AI collaboration for automated classification of patient safety event reports: Algorithm development and validation study. *JMIR Human Factors*, *11*(1), e53378. https://doi.org/10.2196/53378

[25] Rietberg, M. T., Nguyen, V. B., Geerdink, J., Vijlbrief, O., & Seifert, C. (2023). Accurate and reliable classification of unstructured reports on their diagnostic goal using BERT models. *Diagnostics*, *13*(7), 1251. https://doi.org/10.3390/diagnostics13071251

[26] Munthuli, A., Pooprasert, P., Klangpornkun, N., Phienphanich, P., Onsuwan, C., Jaisin, K., . . . , & Tantibundhit, C. (2023). Classification and analysis of text transcription from Thai depression assessment tasks among patients with depression. *PLOS One*, *18*(3), e0283095. https://doi.org/10.1371/journal.pone.0283095

[27] Zhang, H., & Ogasawara, K. (2023). Grad-CAM-based explainable artificial intelligence related to medical text processing. *Bioengineering*, *10*(9), 1070. https://doi.org/10.3390/bioengineering10091070

[28] Rios, A., & Kavuluru, R. (2017). Ordinal convolutional neural networks for predicting RDoC positive valence psychiatric symptom severity scores. *Journal of Biomedical Informatics*, *75*, S85–S93. https://doi.org/10.1016/j.jbi.2017.05.008

[29] Hassan, E., El-Hafeez, Abd., T, & Shams, M. Y. (2024). Optimizing classification of diseases through language model analysis of symptoms. *Scientific Reports*, *14*(1), 1507. https://doi.org/10.1038/s41598-024-51615-5

[30] Barman, N. R., Karim, F., &; Sharma, K. (2022). Symptom 2 Disease dataset [Data set]. Kaggle. https://www.kaggle.com/datasets/niyarrbarman/symptom2disease

[31] Boyle, T. (2018). Medical Transcriptions dataset [Data set]. Kaggle. https://www.kaggle.com/datasets/tboyle10/medicaltranscriptions

[32] Shoaibuddin. (2023). Symptoms based medicaltest recommendations dataset [Data set]. Kaggle. https://www.kaggle.com/datasets/shoaibuddin/symptoms-based-medical-test-recommendations

How to Cite: Khairuddin, A. M., Aris, S. A. M., Azizan, A., & Zakaria, N. J. (2026). Explainable AI in Healthcare: Text Classification Using Bidirectional LSTM and LIME. *Artificial Intelligence and Applications*. https://doi.org/10.47852/bonviewAIA620210055