 DOI: 10.35382/TVUJS.15.3.2025.137

AI-BASED CLINICAL TEXT CLASSIFICATION FOR LUNG DISEASE DIAGNOSIS

Thi-Diem Truong[1,2], Thanh-Nghi Do[3*]

Abstract – *Lung diseases pose a significant challenge to global healthcare, with rising incidence and mortality rates underscoring the need for more accurate and efficient diagnostic methods. Although artificial intelligence has shown enormous promise in enhancing diagnostic accuracy, research on applying natural language processing to Vietnamese clinical texts for lung disease classification remains notably limited. This study addresses the critical gap through two significant contributions. First, a novel clinical dataset comprising 12 categories of lung diseases derived from electronic health records at An Giang Provincial General Hospital was introduced. Second, the study conducted a comprehensive comparative evaluation of text representation techniques, including traditional methods (bags of words and term frequency-inverse document frequency) and modern embeddings (Word2Vec, GloVe, FastText, BERT). These representations were integrated with multiple machine learning models to assess classification performance. Experimental results demonstrate that traditional representations consistently outperform modern embeddings on Vietnamese clinical texts. Significantly, the combination of bags of words with the light gradient boosting machine achieved the highest classification accuracy of 86.26%. These findings provided practical guidance on selecting appropriate natural language processing techniques for Vietnamese medical text analysis and underscore the potential of cost-effective artificial intelligence solutions in resource-limited healthcare settings.*

Keywords: clinical data, electronic medical records, lung disease diagnosis, machine learning, text classification.

I. INTRODUCTION

Lung diseases are currently among the leading causes of death globally [1, 2]. According to the World Health Organization (WHO) [3], pneumonia is the leading cause of death in children, accounting for 14% of all deaths in children under five years old and claiming the lives of 740,180 children in 2019. Tuberculosis is also on the rise, with the number of deaths increasing from 1.4 million in 2019 to 1.6 million in 2021 [4]. Other lung diseases also contribute significantly to the global health burden: asthma affects 262 million people and causes approximately 1,000 deaths daily [5]; chronic obstructive pulmonary disease (COPD) caused 3.5 million deaths in 2021, accounting for 5% of total global deaths [6]; and COVID-19 has resulted in over 700 million infections and more than seven million deaths as of April 2024 [7]. This situation highlights the urgent need for early and accurate diagnosis of lung diseases to ensure timely treatment, reduce mortality rates, and improve patients' quality of life [1]. However, the traditional diagnostic process often faces several challenges, including reliance on physicians' clinical experience and prolonged information processing times. In particular, given the shortage of specialized medical personnel in remote areas and lower-level hospitals, automated diagnostic support has become an urgent necessity.

In recent years, artificial intelligence (AI) and machine learning have achieved significant break-

[1]Postgraduate student, College of Information and Communication Technology, Can Tho University, Vietnam
[2]An Giang University, Vietnam National University Ho Chi Minh City, Vietnam
[3]College of Information and Communication Technology, Can Tho University, Vietnam
*Corresponding author: dtnghi@cit.ctu.edu.vn
Received date: 10 June 2025; Revised date: 13 August 2025; Accepted date: 19 September 2025

throughs, offering tremendous potential to improve the quality of healthcare. From medical image diagnosis and treatment outcome prediction to clinical decision support, AI has demonstrated remarkable effectiveness. Notably, clinical text classification through natural language processing (NLP) techniques has garnered considerable attention from the international research community. Notably, clinical text classification through NLP techniques has garnered considerable attention from the international research community. Within the realm of digital healthcare, electronic medical records (EMRs) serve as an extensive, centralized, and systematic repository of patient information. EMRs contain vast amounts of unstructured textual data, including symptom descriptions, medical histories, test results, imaging reports, and treatment progress notes. By utilizing AI to leverage data sources, physicians can gain invaluable insights to support them in the diagnostic process and the development of optimal treatment plans.

Clinical text classification is a promising research direction aimed at automating the extraction of disease-related information from EMRs. However, this field faces numerous challenges, particularly when applied to the Vietnamese language. Unlike general texts, medical documents are linguistically complex, characterized by a high density of specialized terminology, abbreviations, varied expressions, and domain-specific language. Additionally, clinical datasets often suffer from severe class imbalance, especially in cases involving rare diseases. Using AI to automate EMRs is easier in the English language, as it would focus solely on and extract data from the presented information. Whereas the linguistic characteristics and cultural context of healthcare in Vietnam necessitate customized approaches. Consequently, there is a notable shortage of annotated Vietnamese clinical text datasets, particularly those sourced from local hospitals.

To address this issue, a new Vietnamese clinical text dataset was constructed from EMRs collected at An Giang Provincial General Hospital. The dataset comprises clinical descriptions labeled into 12 common categories of lung diseases: COPD, COVID-19, Asthma, Tuberculosis, Pulmonary Oedema, Respiratory Failure, Pleural Effusion, Pneumothorax, Malignant Neoplasm, Pneumonia, Pulmonary Collapse, and normal conditions. This dataset serves as a valuable resource for training and evaluating models designed to classify lung diseases from Vietnamese clinical texts.

The study employs a comprehensive approach to clinical text classification, aiming to support lung disease diagnosis by combining multiple feature extraction techniques with various machine learning models. Specifically, the research utilizes traditional feature extraction methods such as bag-of-words (BoW) and term frequency–inverse document frequency (TF-IDF), alongside modern word embedding techniques, including Word2Vec, GloVe [8], FastText, and the advanced transformer-based model BERT [9]. For classification, the study compares and evaluates a wide range of machine learning algorithms, from basic to advanced, including logistic regression (LR) [10], support vector machine (SVM) [11], stochastic gradient descent (SGD) [10], Gaussian Naive Bayes (GNB), Multinomial Naive Bayes (MNB) [10], K-nearest neighbors (KNN) [12], multi-layer perceptron (MLP) [10], random forest (RF) [13], extreme gradient boosting (XGB) [14], light gradient boosting machine (LGBM) [15], and CatBoost (CB) [16]. The primary objective is to comprehensively assess the effectiveness of various combinations of feature extraction methods and classification algorithms in identifying lung diseases from clinical text.

The experimental results demonstrate that the choice of both machine learning models and word representation methods influences classification performance. Traditional representation techniques such as BoW and TF-IDF outperform modern embedding methods when applied to Vietnamese clinical texts. In particular, the LGBM model, when combined with BoW, achieved the highest accuracy of 86.26%, highlighting the effectiveness and suitability of this approach for processing and classifying highly

domain-specific Vietnamese clinical texts. This study not only contributes to improving the accuracy of lung disease diagnosis but also paves the way for further research on the application of AI in Vietnamese healthcare.

This paper is organized as follows: Section 2 provides an overview of related work on medical text classification. Section 3 describes the dataset and research methodology. Section 4 presents experimental results and a comparative analysis of the methods. Finally, Section 5 discusses the conclusions and directions for future research.

II. LITERATURE REVIEW

In recent years, the application of machine learning AI models to clinical text data has emerged as a prominent trend, attracting significant attention from domestic and international research communities. These efforts have opened up new opportunities to optimize clinical diagnosis and decision-making by improving accuracy, consistency, and overall efficiency in clinical data sets within modern healthcare systems.

Machine learning algorithms such as RF [13] and SVM [11] have been widely applied in numerous medical studies for disease classification tasks based on textual data. These methods have demonstrated strong effectiveness in NLP, enabling the extraction of meaningful features from clinical texts and improving the accuracy of automatic disease detection and classification [17–20]. Chapman et al. [21] evaluated the performance of three approaches – expert rules, Bayesian networks, and decision trees – in automatically identifying bacterial pneumonia from chest X-ray reports. Using a dataset of 292 reports processed with an NLP tool combined with manual correction, all three methods achieved high performance, with area under the curve scores ranging from 0.940 to 0.960. Notably, the accuracy of these models was comparable to that of medical experts, highlighting their strong potential to support physicians in clinical practice. In a different approach, Buntoro et al. [22] applied text preprocessing techniques to classify 19 different diseases based on symptoms extracted from medical records. Two machine learning models, Naive Bayes and RF, were employed for classification. The results showed that RF achieved the highest accuracy of 99.9%, confirming the effectiveness and promising applicability of this method in medical data processing.

In particular, during the COVID-19 pandemic, AI and machine learning have played a crucial role in supporting rapid and accurate clinical decision-making. Khanday et al. [23] classified clinical reports into four disease categories based on features extracted using TF-IDF and BoW models. Algorithms such as logistic regression and MNB achieved accuracy rates as high as 96.2%. Similarly, Batra et al. [24] employed feature extraction techniques, including TF-IDF, BoW, and text length, to detect suspected COVID-19 cases from medical texts. The Naive Bayes model consistently outperformed other methods in terms of accuracy.

Recent advancements have seen the emergence of transformer-based language models, which have revolutionized clinical NLP. Models such as BioBERT [25], ClinicalBERT [26], and BlueBERT [27] leverage contextual embeddings to capture complex semantic relationships in medical text, significantly outperforming traditional word representation methods. For multilingual applications, XLM-RoBERTa [28] and Medical mT5 [29] have demonstrated strong cross-lingual transfer capabilities, enabling effective classification in languages with limited annotated data. For Vietnamese medical research, PhoBERT [30] and ViHealthBERT [31] have been introduced, offering domain-adapted embeddings for both general and biomedical text. These transformer-based approaches have shown substantial performance gains in various Vietnamese NLP tasks, including named entity recognition, relation extraction, and document classification, making them highly relevant for clinical text processing.

Despite these advances, research on Vietnamese clinical texts remains limited, both in terms of dataset availability and in comparative evaluations across representation methods. A comprehensive benchmarking of traditional

methods (BoW, TF-IDF) versus modern embeddings (Word2Vec, GloVe, FastText, and BERT-based models) on a Vietnamese clinical dataset has yet to be fully explored, contrasted to English-language corpora.

This study constructed a Vietnamese clinical text dataset extracted from electronic medical records at An Giang Provincial General Hospital. Using this dataset, a comprehensive evaluation of six text representation methods (BoW, TF-IDF, Word2Vec, GloVe, FastText, and BERT) was conducted by combining them with eleven different machine learning algorithms. This approach not only provides valuable insights into the relative performance of each method on Vietnamese clinical data but also offers practical guidance for developing AI systems supporting lung disease diagnosis.

III. RESEARCH METHODS

A. *Overall clinical text classification process*

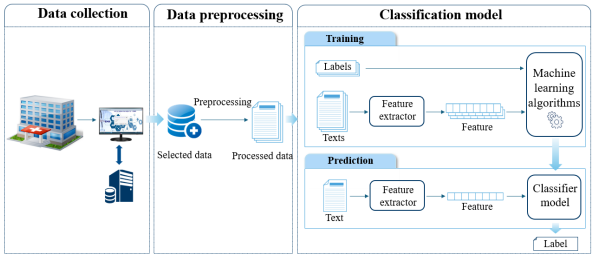

Fig. 1: Overall clinical text classification process

The clinical text classification process to support the diagnosis of lung diseases consists of three main stages. The first stage involves collecting data from EMRs at An Giang Provincial General Hospital. The second stage focuses on data preprocessing, which includes noise removal, language normalization, and labeling according to specific disease categories. The final stage is model construction, where features are extracted from the clinical text and used to train machine learning algorithms that learn the mapping between the input content and disease labels. Once trained, the model is used to classify new clinical texts. Figure 1 provides a visual overview of the entire process, from data collection to prediction.

B. *Data collection*

An Giang Provincial General Hospital has implemented an EMRs system to modernize patient information management, aiming for comprehensiveness and efficiency. This system functions as a centralized repository, integrating data from all departments, including patient admission, clinical examinations, laboratory tests, diagnoses, treatments, and discharges. This facilitates the systematic, consistent, and complete collection of clinical information, supporting both medical treatment and research. Clinical information in textual form is collected through the EMRs administrative interface. Throughout the data collection process, the hospital's IT technicians provided close support to ensure the data gathered was accurate, comprehensive, and compliant with strict patient information privacy regulations.

Focusing on lung-related diseases, the research team selected and extracted valuable information fields to support the diagnosis of lung conditions. Irrelevant attributes or those likely to introduce noise into the analytical model were removed to enhance the quality of the dataset. The clinical text dataset used in this study includes crucial information such as patient medical history, results of general and related organ examinations, pertinent laboratory test results, disease progression during treatment, and discharge outcomes.

C. *Data preprocessing*

Data preprocessing and labeling

Preprocessing clinical text data is a critical step in building a machine learning system, ensuring that the input is high-quality, well-structured, and suitable for analysis. Vietnamese EMR text often contains specialized medical terminology, domain-specific abbreviations, typographical errors, and incomplete or noisy entries, all of which can introduce variability and noise into the modeling process. To address these challenges, this study implemented a multi-stage preprocessing pipeline. First, the simple_preprocess()

function from the Gensim library was applied to normalize the text by lowercasing, removing punctuation and non-alphabetic tokens, and performing basic tokenization. Next, Vietnamese word segmentation was carried out using the ViTokenizer.tokenize() function from the PyVi toolkit. This step is particularly crucial for Vietnamese, a language in which words are often composed of multiple syllables separated by spaces. Without proper segmentation, multi-syllable medical terms may be incorrectly split, leading to semantic distortion and reduced model accuracy. ViTokenizer was chosen for its high accuracy in handling domain-specific terms and its ability to produce linguistically coherent token sequences for downstream feature extraction. Finally, dataset verification and refinement steps were performed, involving manual review to correct residual errors, handling missing values, and removing records with excessive missing data, thereby ensuring dataset integrity and reliability for subsequent modeling.

After completing the preprocessing phase, the text dataset was labeled based on patients' discharge diagnoses, which were directly extracted from the EMRs. The labeled dataset consists of 17,973 texts divided into 12 categories: COPD, COVID-19, Asthma, Tuberculosis, Pulmonary Oedema, Respiratory Failure, Pleural Effusion, Pneumothorax, Malignant Neoplasm, Pneumonia, Pulmonary Collapse, and Normal. Figure 2 illustrates the distribution of clinical texts that were collected, preprocessed, and labeled for each disease category in the study dataset. This dataset serves as a crucial resource for training and evaluating models in classifying lung diseases from clinical texts.

Results of clinical text dataset construction

After completing the data collection and processing phase at An Giang Provincial General Hospital, this study successfully constructed a dataset comprising 17,973 clinical text files. Each file contains information extracted from EMRs, which provides comprehensive knowledge about the patient's medical condition. These include fields such as symptom descriptions, general

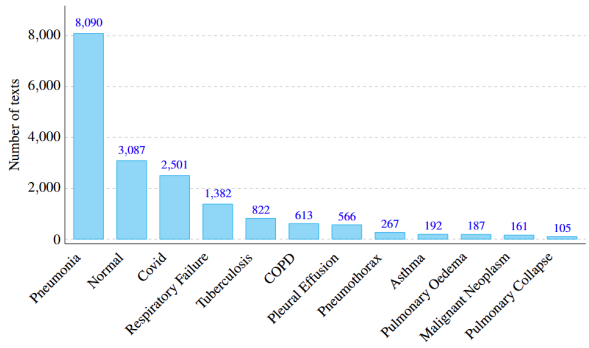

Fig. 2: Number of clinical texts by disease category

physical examinations, respiratory and circulatory assessments, medical history, laboratory test summaries, and discharge summaries. Table 1 presents the detailed contents of a typical clinical text file. This descriptive information serves as essential input for machine learning models and diagnostic support within the study.

Table 2 provides an overview of the basic descriptive statistics for the clinical text dataset used in this study. With a total of 17,973 documents, the dataset is sufficiently large to meet the requirements of modern machine learning tasks. On average, each document contains approximately 1,014.29 characters, which is equivalent to about 184.05 words. This suggests that each medical record is documented in considerable detail, encompassing comprehensive information about the patient's examination, diagnosis, treatment, and disease progression. Non-alphabetic characters account for around 29.68% of the text, reflecting the unique characteristics of medical data. These include specialized symbols, laboratory values, and distinctive note formats, which differ from regular text and highlight the specific style of expression and annotation used in medical records.

Table 1: Example of information in a clinical text

Information field name	Patient condition information
Symptom descriptions	The patient has experienced sudden chest heaviness for the past three days, radiating as a sharp pain to the left back, accompanied by shortness of breath when lying down.
General physical examinations	The patient is alert and well-oriented. Skin and mucous membranes are pink. Extremities are warm, with palpable pulses. Sharp pain is reported in the left back. Shortness of breath happens when lying down. SpO_2 was at 94% on room air.
Respiratory assessments	The chest is symmetrical and moves evenly with respiration. Tactile fremitus is equal bilaterally, and chest expansion is symmetrical. Percussion reveals resonant sounds. Breath sounds are soft and vesicular on both sides, with fine, moist crackles and wheezing.
Circulatory assessments	The apex beat is located at the 4^{th} intercostal space along the left midclavicular line. No thrills are detected. S1 and S2 are clear and regular.
Medical history	Hypertension, heart failure
Laboratory test summaries	WBC: 9.21, RBC: 3.68, HGB: 11.7, HCT: 37.5, PLT: 223; Posteroanterior (PA) chest X-ray
Discharge summaries	HA: 110/80 mmHg. LDVV: Productive cough and fever. Doctor's note: The illness began three days before hospital admission. The patient presented with a productive cough producing a large amount of opaque white sputum, without blood. Symptoms did not improve with medication. Urination and defecation were normal. Appetite was poor. Medical history: Hypertension. Examination: The patient was conscious and responsive. Mucous membranes were pink. There was no shortness of breath and no chest pain. Heart rhythm was regular. Lungs revealed crackles, moist rales, and wheezing. The abdomen was soft, with tenderness in the epigastric region.

Table 2: Statistics of clinical text dataset characteristics

Total number of clinical texts	Average number of characters per text	Average number of words per text	Ratio of non-alphabetic characters (%)
17,973	1,014.29	184.05	29.68

Figure 3 visually illustrates the distribution of clinical text lengths based on word counts. The distribution exhibits a clear right-skewed pattern, with most texts concentrated between 100 and 250 words. Notably, the peak occurs within the 150–180 word range, where the number of texts reaches nearly 2,000 texts. This indicates that most medical records are documented at a moderate length with a relatively consistent level of detail. This length is generally sufficient to capture the necessary clinical information, facilitating effective data mining and enabling machine learning models to comprehend important contexts and features. The shape of the distribution closely resembles a right-skewed normal distribution, reflecting consistency and stability in the documentation of medical records. Overall, these findings suggest that the textual data quality is quite good and well-suited for NLP tasks in the medical field.

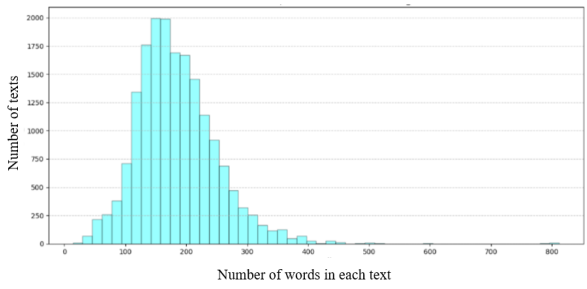

Fig. 3: Distribution of clinical text lengths

D. Building a clinical text classification model for lung disease diagnosis

The development of the clinical text classification model to support lung disease diagnosis

involves two main stages: model training and label prediction for new texts.

In the training stage, clinical texts first undergo preprocessing procedures, including cleaning, normalization, and tokenization. The preprocessed texts are then converted into numerical feature representations using various feature extraction techniques, such as BoW, TF-IDF, Word2Vec, GloVe, FastText, and BERT. Each of these employs a distinct approach to transform natural language into feature vectors that capture either frequency or semantic information. These feature vectors, combined with their corresponding disease labels, are then used to train various machine learning models, including LR, SVM, SGD, GNB, MNB, KNN, MLP, RF, XGB, LGBM, and CB.

In the prediction stage, a new, unlabeled clinical text is subjected to the same preprocessing and feature extraction procedures used during training. The resulting feature vector is then fed into the trained model to predict the corresponding disease label. The classification task involves 12 lung disease categories: COPD, COVID-19, Asthma, Tuberculosis, Pulmonary Oedema, Respiratory Failure, Pleural Effusion, Pneumothorax, Malignant Neoplasm, Pneumonia, Pulmonary Collapse, and Normal. To evaluate the model's effectiveness, key statistical metrics such as Precision, Recall, F1-score, and Accuracy are employed. These metrics assess the reliability and practical applicability of each model in accurately classifying lung diseases from clinical texts, thereby providing valuable support to physicians in diagnosis and treatment.

IV. RESULTS AND DISCUSSION

A. Experiment environment

To evaluate the performance of the models in classifying lung diseases from clinical texts, all experiments were conducted in a Python environment utilizing popular libraries such as TensorFlow [32] and Scikit-learn [33].

The experiments were performed on a Linux Fedora Core 33 operating system, running on a high-performance computer equipped with an Intel(R) Core i7-4790 processor (3.6 GHz, 4 cores), 32 GB of RAM, and a GeForce RTX 2080 Ti GPU with 4,352 CUDA cores and 11 GB of GDDR6 memory.

B. Dataset description

The study newly constructed a dataset of 17,973 clinical texts, collected and processed from EMRs at An Giang Provincial General Hospital. This dataset was labeled into 12 disease categories. To ensure effective training and evaluation of the machine learning models, the dataset was randomly split into two subsets with an 80:20 ratio. The training set contains 14,371 texts, providing sufficient data for the model to learn and recognize important features of each disease group. The test set comprises 3,602 texts, which are used independently to objectively and accurately assess the predictive performance of the trained model. This division helps ensure the model's generalizability and practical applicability. Table 3 presents detailed information about the distribution of disease labels within the clinical text dataset.

Table 3: Dataset description

No.	Labels	Training set	Test set	Total
1	Normal	2,469	618	3,087
2	COPD	490	123	613
3	Covid	2,000	501	2,501
4	Asthma	153	39	192
5	Tuberculosis	657	165	822
6	Pulmonary Oedema	149	38	187
7	Respiratory Failure	1,105	277	1,382
8	Pleural Effusion	452	114	566
9	Pneumothorax	213	54	267
10	Malignant Neoplasm	128	33	161
11	Pneumonia	6,472	1,618	8,090
12	Pulmonary Collapse	83	22	105
	Total	**14,371**	**3,602**	**17,973**

C. Experimental results

In the experimental phase, a comprehensive evaluation of various feature extraction tech-

niques combined with machine learning models for clinical text classification was conducted to support the accurate diagnosis of lung diseases. Six popular feature extraction techniques were utilized in this study, including BoW, TF-IDF, Word2Vec, GloVe, FastText, and BERT. Following the transformation of clinical texts into vector representations, 11 different machine learning models were trained and evaluated: LR, SVM, SGD, GNB, MNB, KNN, MLP, RF, XGB, LGBM, and CB.

To ensure optimal performance, the machine learning models were carefully fine-tuned with specific hyperparameters. The SVM employed an RBF kernel with a regularization parameter C = 105 and γ = 0.000045. The SGD model was trained using the modified_huber loss function for up to 100 iterations, with a constant learning rate and an initial learning rate (eta0) of 0.1. The KNN model was configured with k = 1 and used Euclidean distance. The MLP utilized the Adam optimizer, featuring two hidden layers with 256 neurons each, an initial learning rate of 0.001, and a maximum of 200 iterations. The RF model was set with max_features = 50 and n_estimators = 200. For both the XGB and LGBM models, the gamma parameter of 0.1 provides optimal regularization to prevent overfitting while maintaining model flexibility. This parameter was selected through hyperparameter tuning as it yielded the best validation performance.

The experimental results, summarized in Table 4, show the accuracy of the models across various text representation methods. These findings provide a comprehensive overview of the effectiveness of each model and representation technique, guiding the selection of the most suitable approach for clinical text classification in the context of pulmonary disease diagnosis.

The experimental results demonstrate the superiority of certain models when combined with traditional feature extraction methods. The LGBM model showed the highest accuracy of 86.26% with BoW, outperforming all other models in the study. The XGB model also performed strongly, reaching accuracies of 86.15% with BoW and 85.15% with TF-IDF. The SVM delivered high accuracy as well, achieving 84.48% with BoW and 84.81% with TF-IDF, confirming its status as a traditional yet highly effective model. The MLP showed notable potential in clinical text classification, particularly with BoW (83.09%) and TF-IDF (80.82%). Meanwhile, models such as RF, CB, and KNN achieved moderate accuracy levels ranging from approximately 65% to 82%, depending on the feature extraction method used. It indicates that GNB exhibited the lowest performance across all experiments, especially when combined with BERT, with only 23.15% accuracy.

Table 4: The accuracy results of clinical text classification

Machine learning models	Accuracy (%)					
	BoW	TF-IDF	Word2Vec	GloVe	FastText	BERT
LR	81.04	79.07	72.32	70.41	71.38	68.41
SVM	84.48	84.81	79.84	77.57	**73.07**	78.51
SGD	73.76	81.09	66.63	66.69	70.49	63.21
GNB	41.70	41.87	33.65	30.62	58.05	23.15
MNB	69.32	69.93	59.52	47.95	70.13	48.83
KNN	75.93	77.54	74.65	72.74	64.49	68.93
MLP	83.09	80.82	78.90	78.51	64.80	**79.46**
RF	81.48	81.01	79.40	76.82	70.29	74.88
CB	80.93	80.23	80.57	**79.15**	69.85	78.26
XGB	86.15	85.15	80.84	79.04	71.07	78.29
LGBM	**86.26**	**86.15**	**81.09**	79.12	67.93	77.96

Regarding feature extraction techniques, BoW and TF-IDF continue to demonstrate the significant value of traditional text representation methods. When combined with strong models like LGBM, XGB, SVM, and MLP, these two techniques consistently deliver the highest performance. In contrast, modern word embedding methods such as Word2Vec, GloVe, and FastText provide results ranging from average to good but generally do not outperform BoW and TF-IDF in the context of clinical text classification. These results emphasize that, for this specific task, traditional text representation methods remain highly effective and reliable. Furthermore, models like LGBM, XGB, and SVM emerge as the optimal choices for lung disease classification based on text data.

Table 5: Best model performance for each feature extraction method

Models	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
LGBM-BoW	**86.26**	85.57	69.18	74.85
LGBM-TFIDF	86.15	86.16	69.04	75.04
LGBM-W2V	81.09	85.67	58.33	65.16
CB-Glove	79.15	84.06	56.16	63.39
SVM-FastText	73.07	36.66	35.44	35.75
MLP-BERT	79.46	71.90	60.44	64.39

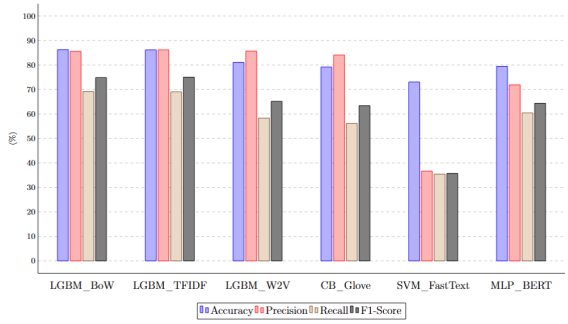

Fig. 4: Best model performance for each feature extraction method

Table 5 presents the experimental results, illustrating the performance of the best machine learning models selected for each feature extraction method in the clinical text classification task (the visualized results are shown in Figure 4). The LGBM-BoW model achieved the highest accuracy of 86.26% and an F1-score of 74.85%, highlighting the superior effectiveness of traditional text representations when combined with the LGBM model. The LGBM-TFIDF model also achieved comparable performance (86.15% accuracy, 75.04% F1-score), further confirming the strength of traditional text representation methods combined with LGBM. The CB-GloVe model produced average results (79.15% accuracy, 63.39% F1-score). In contrast, the SVM-FastText model recorded the lowest performance, with precision, recall, and F1-score all around 35%, indicating the incompatibility between FastText and the SVM classifier for this task. Finally, the MLP-BERT model demonstrated the potential of deep learning with semantic representations from BERT, achieving 79.46% accuracy and a 64.39% F1-score, despite its weaker performance compared to traditional methods like BoW or TF-IDF combined with LGBM. Overall, these findings suggest that the combination of traditional feature representations and the LGBM model continues to be an effective and robust choice for clinical text classification.

Figure 5 illustrates the confusion matrix of the LGBM-BoW model, which achieved the highest accuracy in the study, detailing its performance across each specific disease class. The model attained high accuracy for classes with abundant data, such as pneumonia (1,507 out of 1,618 texts correctly classified) and normal cases (542 out of 618). COVID-19 was also effectively classified, with 478 out of 501 texts correctly predicted, reflecting its distinct disease features. However, confusion persisted among diseases with similar symptoms, particularly between respiratory failure, tuberculosis, pneumonia, and pleural effusion. For example, 60 respiratory failure texts were misclassified as pneumonia, and tuberculosis texts were frequently confused with both pneumonia and respiratory failure. Diseases with fewer texts, such as asthma, pulmonary collapse, and malignant neoplasm, exhibited higher misclassification rates and lower accuracy, highlighting the impact of data imbalance. Asthma was correctly predicted in only 20 out of 39 texts, with most misclassified as pneumonia. Similarly, pulmonary collapse had 14 correctly predicted texts, many of which were also misclassified as pneumonia. These results clearly demonstrate the significant influence of class imbalance on classification performance.

Generally, the LGBM-BoW model shows strong generalization ability for common lung conditions. Nevertheless, to optimize performance, the model may require further fine-tuning for rare disease classes. Addressing these cases will improve overall accuracy and enhance the model's practical applicability in clinical settings.

Fig. 5: Confusion matrix of LGBM-BoW model

V. CONCLUSION AND RECOMMENDATIONS

This study highlights the significant potential of AI in assisting the diagnosis of lung diseases through clinical text classification. To support this, a novel real-world clinical dataset was developed, derived from electronic medical records at the An Giang Provincial General Hospital. Through a comprehensive evaluation of text representation methods and machine learning algorithms, the results indicate that traditional techniques such as BoW and TF-IDF remain more effective for Vietnamese medical language processing compared to modern embedding models. The combination of BoW with the LGBM model achieved an impressive accuracy of 86.26%, confirming the effectiveness of this approach for classifying lung diseases from clinical text. The contributions of this study provide a valuable dataset and offer deep insights into NLP techniques for lung disease diagnosis, paving the way for broader AI applications in healthcare.

Based on these results, future research will focus on expanding and diversifying the dataset by collecting data from multiple hospitals to improve the model's generalizability and reliability. Concurrently, integrating multimodal data such as medical images, laboratory test results, and genetic data will create a comprehensive and accurate diagnostic system. To enhance trust and acceptance within the medical community, future studies should emphasize developing Explainable AI methods to make AI decision-making processes transparent and understandable.

ACKNOWLEDGMENTS

This research has received support from the Vietnamese Ministry of Education and Training's scientific research project, code B2025-TCT-01. We would like to thank the College of Information Technology, Can Tho University for their support of this project.

REFERENCES

[1] Kieu STH, Bade A, Hijazi MH, Kolivand H. A survey of deep learning for lung disease detection on medical images: state-of-the-art, taxonomy, issues and future directions. *Journal of Imaging*. 2020;6(12): 131. https://doi.org/10.3390/jimaging6120131.

[2] Yadav P, Menon N, Ravi V, Vishvanathan S. Lung-GANs: Unsupervised representation learning for lung disease classification using chest CT and X-ray images. *IEEE Transactions on Magnetics*. 2023;70(8): 2774–2786. https://doi.org/10.1109/TEM.2021.3103334.

[3] World Health Organization. *Pneumonia in children*. https://www.who.int/news-room/fact-sheets/detail/pneumonia [Accessed 25 May 2025].

[4] World Health Organization. *Global tuberculosis report 2022*. https://www.who.int/teams/global-tuberculosis-programme/tb-reports/global-tuberculosis-report-2022 [Accessed 13 June 2024].

[5] Network GA. *The global asthma report 2022*. http://globalasthmareport.org/ [Accessed 13 June 2024].

[6] World Health Organization. *Chronic obstructive pulmonary disease (COPD)*. https://www.who.int/news-room/fact-sheets/detail/chronic-obstructive-pulmonary-disease-(copd) [Accessed 25 May 2025].

[7] UNICEF. *COVID-19 response | UNICEF Serbia*. https://www.unicef.org/serbia/en/-covid-19-response [Accessed 25 May 2025].

[8] Pennington J, Socher R, Manning CD. GloVe: Global vectors for word representation. In: Moschitti A, Pang B, Daelemans W (eds.). *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. October 2014; Doha, Qatar. United States of America: Association

for Computational Linguistics; 2014. p.1532–1543. https://doi.org/10.3115/v1/D14-1162.

[9] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T (eds.). *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Volume 1 (long and short papers). June 2019; Minneapolis, Minnesota. United States of America: Association for Computational Linguistics; 2019. p.4171–4186. https://doi.org/10.18653/v1/N19-1423.

[10] John Lu ZQ. The elements of statistical learning: data mining, inference, and prediction. *Journal of the Royal Statistical Society Series A: Statistics in Society*. 2010;173(3): 693–694. https://doi.org/10.1111/j.1467-985X.2010.00646_6.x.

[11] Vapnik NV. *The nature of statistical learning theory*. New York: Springer-Verlag; 1995.

[12] Cover T, Hart P. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*. 1967;13(1): 21–27. https://doi.org/10.1109/TIT.1967.1053964.

[13] Breiman L. Random forests. *Machine Learning*. 2001;45(1): 5–32. https://doi.org/10.1023/A:1010933404324.

[14] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, in KDD '16*. New York, NY, USA: Association for Computing Machinery; 2016. p.785–794. https://doi.org/10.1145/2939672.2939785.

[15] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: a highly efficient gradient boosting decision tree. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Red Hook, NY, USA: Curran Associates Inc.; 2017. p.3149–3157. https://dl.acm.org/doi/10.5555/3294996.3295074 [Accessed 25 May 2025].

[16] Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18)*. Red Hook, NY, USA: Curran Associates Inc.; p.6639–6649. https://dl.acm.org/doi/pdf/10.5555/3327757.3327770 [Accessed 25 May 2025].

[17] Alam MdZ, Rahman MS, and Rahman MS. A random forest based predictor for medical data classification using feature ranking. *Informatics in Medicine Unlocked*. 2019;15: 100180. https://doi.org/10.1016/j.imu.2019.100180.

[18] Karimah S, Setiawan EB, Kurniawan I. Implementation of random forest in classification model of diabetes prediction based on drug review content. In: *2021 International Conference on Data Science and Its Applications (ICoDSA)*. Bandung, Indonesia: IEEE; 2021. p.228–232. https://doi.org/10.1109/ICoDSA53588.2021.9617218.

[19] Sheng L, Qing S, Wenjie H, Aize C. Diseases classification using support vector machine (SVM). In: *Proceedings of the 9th International Conference on Neural Information Processing, 2002 (ICONIP '02)*. 18–22 November 2002; Singapore. IEEE; 2003. p.760–763. https://doi.org/10.1109/ICONIP.2002.1198160.

[20] Nabilah 'Izzaturrahmah A, Nhita F, Kurniawan I. Implementation of support vector machine on text-based GERD detection by using drug review content. In: *2021 International Conference Advancement in Data Science, E-learning and Information Systems (ICADEIS)*. 13–14 October 2021; Bali, Indonesia. IEEE; 2021. p.1–6. https://doi.org/10.1109/ICADEIS52521.2021.9702073.

[21] Chapman WW, Fizman M, Chapman BE, Haug PJ. A comparison of classification algorithms to automatically identify chest X-ray reports that support pneumonia. *Journal of Biomedical Informatics*. 2001;34(1): 4–14. https://doi.org/10.1006/jbin.2001.1000.

[22] Buntoro GA, Wibawa AD, Purnomo MH. Text mining in healthcare for disease classification using machine learning algorithm. In: *International Electronics Symposium (IES)*. 29–30 September 2021; Surabaya, Indonesia. IEEE; 2021. p.97–101. https://doi.org/10.1109/IES53407.2021.9593998.

[23] Khanday AMUD, Rabani ST, Khan QR, Rouf N, Mohi Ud Din M. Machine learning based approaches for detecting COVID-19 using clinical text data. *International Journal of Information Technology*. 2020;12(3): 731–739. https://doi.org/10.1007/s41870-020-00495-9.

[24] Batra R, Mahajan M, Shrivastava VK, Goel AK. Detection of COVID-19 using textual clinical data: a machine learning approach. In: Mishra S, Mallick PK, Tripathy HK, Chae GS, Mishra BSP (eds.). *Impact of AI and data science in response to coronavirus pandemic*. Singapore: Springer; 2021. p.97–109. https://doi.org/10.1007/978-981-16-2786-6_5.

[25] Lee J, Yoon W, Kim S, Kim D, Kim S, So, CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36(4): 1234–1240. https://doi.org/10.1093/bioinformatics/btz682.

[26] Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T, et al. Publicly available clinical BERT embeddings. *arXiv* [Preprint] 2019. Version 3. https://doi.org/10.48550/arXiv.1904.03323.

[27] Peng Y, Yan S, and Lu Z. Transfer learning in biomedical natural language processing: an eval-

uation of BERT and ELMo on ten benchmarking datasets. *arXiv* [Preprint] 2019. Version 2. https://doi.org/10.48550/arXiv.1906.05474.

[28] Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, et al. Unsupervised cross-lingual representation learning at scale. *arXiv* [Preprint] 2020. Version 2. https://doi.org/10.48550/arXiv.1911.02116.

[29] García-Ferrero I, Agerri R, Salazar AA, Cabrio E, de la Iglesia I, Lavelli A, et al. Medical mT5: an open-source multilingual text-to-text LLM for the medical domain. *arXiv* [Preprint] 2024. Version 2. https://doi.org/10.48550/arXiv.2404.07613.

[30] Nguyen DQ, Nguyen AT. PhoBERT: Pre-trained language models for Vietnamese. *arXiv* [Preprint] 2020. Version 3. https://doi.org/10.48550/arXiv.2003.00744.

[31] Minh N, Tran VH, Hoang V, Ta HD, Bui TH, Truong SQH. ViHealthBERT: Pre-trained language models for Vietnamese in health text mining. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association; 2022. p.328–337. https://aclanthology.org/2022.lrec-1.35/ [Accessed 11 August 2025].

[32] Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: large-scale machine learning on heterogeneous distributed systems. *arXiv* [Preprint] 2016. https://doi.org/10.48550/arXiv.1603.04467.

[33] Pedregosa F. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*. 2011;12: 2825–2830. https://doi.org/10.48550/arXiv.1201.0490.