Vol. 14 (2026): Proceedings of the 2nd International Conference on Artificial Intelligence, Modern Engineering and Environmental Sustainability (IC-AIMEES 2026)
E-ISSN: 3105-0522 P-ISSN：3105-0514
DOI: https://doi.org/10.70267/ ic-aimees.20260334339

A BERT-Based Interpretable Deep Learning Model for Medical Diagnosis Assisted by Natural Language Processing

Yutong Chen*

School of Finance, Beijing University of Finance and Technology, Beijing, China

**Corresponding author: Yutong Chen.*

Abstract

The combination of deep learning and natural language processing holds great potential in the medical field. This study aims to explore and develop a deep learning-based natural language processing model to assist in medical diagnosis. We used a large, de-identified dataset of electronic health records, which includes patient complaints, medical histories, and final diagnoses. First, we preprocessed these unstructured text data, for example, by tokenization and removing stop words. Then, we built a deep learning model based on a pre-trained language model like BERT. This model can automatically extract key features from clinical texts and learn the complex relationships between these features and specific diseases. Experimental results show that our proposed model achieves significantly higher accuracy and recall than traditional baseline models on several disease prediction tasks. More importantly, we also introduced an attention mechanism, which gives the model a certain level of interpretability: it can highlight the keywords or phrases that most influence the diagnostic decision. We conclude that deep learning and natural language processing can not only improve the accuracy of disease prediction but also provide valuable references for clinicians, thereby enhancing the overall quality and efficiency of medical services.

Keywords

deep learning, natural language processing, medical diagnosis, electronic health records, interpretability

1. Introduction

In recent years, the healthcare industry has generated massive amounts of data every day. A large part of this data exists in the form of unstructured text, such as doctors' outpatient notes, patients' past medical histories, and imaging reports [1]. These texts actually contain a lot of valuable information for diagnosis. In fact, research suggests that up to 80% of clinically relevant information in electronic health records may reside in free text rather than structured coded data [2]. However, effectively extracting and using this information has always been a challenge.

Traditional methods of information extraction mainly rely on manual reading and sorting. This approach is not only inefficient but also prone to bias due to different subjective experiences of doctors. Moreover,

with the explosive growth of medical data, purely manual methods can no longer keep up. As a result, much valuable data just sits in hospital systems and cannot play its role in clinical decision-making.

However, the rapid development of deep learning, especially natural language processing, provides a new way to solve this problem [3]. Deep learning models are very good at automatically learning complex features from large amounts of text, while natural language processing focuses on understanding and processing human language. Combining these two technologies offers the hope of developing intelligent tools to assist doctors in diagnosis.

Although there have been some initial attempts, most current studies have obvious limitations: on the one hand, many models are only targeted at the prediction of a single specific disease, with poor generalization ability; on the other hand, most deep learning models act like a “black box”: they give a conclusion but do not show how they reached it. This makes it challenging for doctors to fully trust and use them. Therefore, there is a clear need to develop a model that is both accurate and transparent about its reasoning.

Thus, the purpose of this study is to develop a deep learning-based natural language processing model to assist medical diagnosis. Our research objectives are twofold: first, to improve the accuracy and efficiency of disease prediction by leveraging the powerful text feature extraction ability of the BERT pre-trained model; second, to endow the model with diagnostic interpretability by introducing an attention mechanism.

2. Methods

2.1 Dataset and Preprocessing

The dataset we used is a publicly available, de-identified English electronic health record dataset called MIMIC-III [4]. This database contains real information from intensive care unit patients, especially the patients' chief complaints (what the patient feels is wrong) and the final diagnoses. This is exactly what we needed.

After obtaining the raw data, the first step was preprocessing, following the clinical text preprocessing norms recommended by relevant research [5], which means converting raw unstructured clinical text into a format that computers can understand more easily. A series of standard text preprocessing procedures were performed tokenization (splitting whole sentences into individual words or phrases), removing stop words (like “the,” “a,” “of,” which have little meaning), and lemmatization (e.g., converting the past tense “ate” to the original form “eat”, the plural “symptoms” to the singular “symptom”) to avoid the model treating different morphological forms of the same word as different features, for example, converting “ate” to “eat” to unify word forms. Research has shown that effective preprocessing of narrative texts—including tokenization, lemmatization, and semantic feature extraction—significantly improves the performance of clinical text analysis tasks [5]. After these steps, the unstructured text became structured data that could be directly input into the deep learning model for training and prediction that the model could directly use.

2.2 Model Architecture

The model architecture we adopted is based on a very powerful pre-trained language model—BERT [6], a state-of-the-art bidirectional transformer-based pre-trained model proposed by Google in 2018. BERT is a pre-trained contextual language model that has learned generalized semantic representations from large-scale corpora. We only need to fine-tune it for our specific task of “medical diagnosis”, avoiding the problem of overfitting caused by training from scratch on small-scale clinical datasets.

The overall architecture of the model is a three-layer structure with an embedded attention mechanism, and the specific composition and functional logic are as follows:

Input layer: Receives the preprocessed text data, i.e., the patient's chief complaint. The input layer converts the discrete text tokens into vectorized representations that can be processed by the model, laying the foundation for subsequent feature extraction.

BERT encoding layer: This is the core part. It captures the deep contextual semantic information of clinical text and converts the input text into a set of high-dimensional semantic vectors that integrate contextual features. Because BERT has “read” so much, it can understand that “chest pain” and “heart

discomfort" might describe similar symptoms, e.g., recognizing that "chest pain" and "heart discomfort" describe similar cardiovascular symptoms.

Output layer: On top of BERT, we added a fully connected layer. The job of this layer is to transform BERT's understanding into a final diagnostic result. For binary classification disease prediction tasks (e.g., judging whether a patient has pneumonia), the output layer outputs a probability value between 0 and 1; for multi-classification tasks, the output layer outputs the probability distribution of the patient suffering from various possible diseases. For example, if the task is to determine whether it is pneumonia, this layer will output a probability between 0 and 1. The closer to 1, the higher the likelihood of pneumonia.

To enable the model to explain why it made a certain judgment, we also incorporated an attention mechanism into the model, which is integrated into the BERT encoding layer. This mechanism helps us identify which words in the input text the model paid most attention to when making its decision, and the attention weight distribution can be visualized (e.g., generating a heat map). For instance, if the model predicts pneumonia, the attention mechanism might show that it mainly relied on words like "fever," "cough," and "phlegm."

2.3 Experimental Setup

We divided the preprocessed data into three parts, with no overlap between the three datasets to ensure the independence of training, validation and testing: 70% for training the model, 10% for validating and adjusting the model's parameters during training (e.g., adjusting the learning rate, batch size, and number of network layers), avoiding overfitting and underfitting of the model, and the remaining 20% was kept completely separate. This remaining part was used only after the model was trained, to test its final performance objectively. This ensures a fair evaluation.

We compared our proposed model with several traditional methods, such as Naive Bayes and Support Vector Machines. These two models are representative of traditional statistical learning and machine learning methods, and were the mainstream models in clinical text-based disease prediction before the rise of deep learning. The main evaluation metrics were accuracy, precision, and recall. These metrics are also the core indicators for evaluating medical diagnostic models in relevant studies [2, 7], measuring the model's performance from different clinical perspectives. Among them, Recall (Sensitivity) reflects the model's ability to identify positive disease samples and is particularly important for clinical disease detection. We conducted 5-fold cross-validation to verify the model's stability, and calculated the standard deviation (SD) of each metric and p-value (t-test) to verify the statistical significance of performance differences between models.

3. Results

3.1 Comparison of Prediction Performance

The results showed that our BERT-based deep learning model significantly outperformed traditional machine learning models on all evaluation metrics.

Table 1: Performance comparison of different models on disease prediction tasks

Model Name	Accuracy (SD)	Precision (SD)	Recall (SD)	p-value (vs. BERT)
Naive Bayes	82.5% (0.032)	80.1% (0.035)	79.3% (0.038)	<0.001
Support Vector Machine	85.2% (0.028)	83.8% (0.029)	82.7% (0.031)	<0.001
Our BERT Model	91.8% (0.015)	90.5% (0.016)	90.1% (0.017)	-

As shown in Table 1, our model achieved an accuracy of 91.8%, a precision of 90.5% and a recall of 90.1%, with the smallest standard deviation among all models (0.015~0.017), which is far lower than that of Naive Bayes (0.032~0.038) and Support Vector Machine (0.028~0.031), indicating that our model has better performance stability in repeated experiments. Compared with the Support Vector Machine (the best-performing traditional model), our model has an accuracy improvement of 6.6 percentage points, a precision improvement of 6.7 percentage points, and a recall improvement of 7.4 percentage points; compared with the Naive Bayes model, the improvements are 9.3, 10.4 and 10.8 percentage points respectively. The t-test results show that the performance differences between our model and the two traditional models are all

statistically significant ($p<0.001$), which proves that the superior performance of our proposed model is not accidental but has reliable statistical support. This improvement aligns with findings from a large-scale multicenter study, which showed that incorporating natural language processing from clinical free text can significantly enhance case detection, with median sensitivity increasing from 62% (using codes alone) to 78% (using codes plus text) [7]. This indicates that the deep learning model can better understand the complex and implicit information in clinical texts, leading to more accurate and stable predictions.

3.2 Interpretability Analysis

Besides prediction accuracy, we also cared about whether the model "makes sense." By visualizing the attention mechanism (generating a color-coded heat map where darker red represents a higher attention weight), we observed an interesting phenomenon.

For example, when the model analyzed the sentence "The patient presented with sudden chest pain accompanied by radiating back pain," to diagnose "aortic dissection," it assigned high attention weights (shown as dark red in a heatmap) to the words "sudden," "chest pain," and "radiating back pain." This reasoning is actually very close to the diagnostic thinking of a clinician. When a doctor sees these keywords, they would also highly suspect aortic dissection and immediately order relevant tests. This shows that our model is not just relying on statistical correlations but has learned some clinically meaningful diagnostic logic. This makes its judgments more convincing.

4. Discussion

Our study confirms that deep learning-based natural language processing has great potential in medical diagnosis. Not only does it predict more accurately than traditional methods, but more importantly, through the attention mechanism, it makes the model's "thinking process" transparent. This is crucial for clinical application.

However, this study also has some limitations, which we must point out in the discussion. First, although the MIMIC-III dataset we used is authoritative, it mainly comes from Western populations and is in English. A recent study analyzing data from 44 U.S. institutions found that single-institution models generalized poorly to external data, with accuracy dropping by an average of 22.4% [8]. Therefore, further validation is necessary to determine whether our model can directly apply to Chinese medical records or perform equally well in hospitals in other countries. This is a question of external validity.

Second, data quality issues in critical care EHRs are pervasive and multifaceted. Research has shown that missing data rates can exceed 80% for some variables, and EHR-related medication errors comprise 34% of all medication errors in ICUs, with one-third having life-threatening potential [9]. While we have addressed some of these issues through preprocessing, the underlying data quality challenges may still affect model performance.

Third, our current model is still at the level of "assisted diagnosis," meaning it acts more like a consultant offering suggestions to the doctor. It cannot yet completely replace the doctor in making final decisions, because medical decisions are very complex and must consider the patient's personal circumstances, psychological state, and so on.

For future research, we think several directions are worth pursuing. One is to try incorporating multimodal data. Instead of just text, we could also include CT images, genetic test results, and so on. With more comprehensive information, the model would certainly make more accurate judgments. Another direction is to collect more data from diverse sources and languages to train and test the model, making it more generalizable. Semantic technologies, including ontologies and knowledge graphs, have shown significant potential in enhancing EHR data quality, particularly in improving conformance, portability, and usability [10]. Integrating these approaches could further improve model performance. Finally, how to better integrate such tools into doctors' daily workflows and design user-friendly interfaces is also an important area for further study.

5. Conclusion

In summary, this study accomplished several things. We successfully developed and validated a medical diagnostic assistance model based on deep learning and natural language processing, which achieves efficient extraction of clinical key features from unstructured text and endows the BERT-based model with basic clinical interpretability for the first time by introducing an attention mechanism. This model can efficiently extract key information from patients' chief complaint texts and achieve better results than traditional methods on multiple disease prediction tasks, with better performance stability and statistically significant improvement ($p<0.001$), proving that the model's superior performance is not accidental. More importantly, by introducing an attention mechanism, we gave the model a basic level of interpretability. This addresses, to some extent, the problem of clinicians not trusting deep learning models. Therefore, our research provides some valuable ideas and methods for the development of intelligent and interpretable clinical decision support systems, and enriches the research on the application of deep learning and NLP in the medical field. The ultimate goal is to help doctors improve diagnostic efficiency and accuracy, reduce the rate of misdiagnosis and missed diagnosis, benefiting patients. We will further optimize the model by integrating multimodal clinical data and constructing multi-language medical record datasets, striving to develop a model with higher accuracy, stronger generalization ability and better clinical adaptability to better serve clinical medical services.

References

[1] Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. Nature Reviews Genetics, 13(6), 395-405.

[2] Ford, E., Carroll, J. A., Smith, H. E., Scott, D., & Cassell, J. A. (2016). Extracting information from the text of electronic medical records to improve case detection: a systematic review. Journal of the American Medical Informatics Association, 23(5), 1007-1015.

[3] Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ... & Dean, J. (2019). A guide to deep learning in healthcare. Nature Medicine, 25(1), 24-29.

[4] Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L. W. H., Feng, M., Ghassemi, M., ... & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. Scientific Data, 3(1), 1-9.

[5] De Faria, C. L., & Santos, R. P. (2025). Preprocessing narrative texts in electronic medical records to identify hospital adverse events: A scoping review. Artificial Intelligence in Medicine, 162, 102987.

[6] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[7] Ford, E., Carroll, J., Smith, H., Davies, K., Koeling, R., Petersen, I., ... & Cassell, J. (2016). What evidence is there for a delay in diagnostic coding of RA in UK general practice records? An observational study of free text. BMJ Open, 6(6), e010393.

[8] Zhang, Y., Li, Q., & Liu, X. (2025). Generalizing machine learning models from clinical free text. Scientific Reports, 15, 31668.

[9] Fleuren, L. M., Thoral, P., Shillan, D., Ercole, A., Elbers, P. W., & Right Data, Right Now, Right Here Collaborative. (2026). Discovery of data quality issues in electronic health records: profound consequences for critical care medicine applications – a systematized review. Critical Care, 30, 19.

[10] Wang, H., Chen, L., & Zhang, J. (2025). Semantics-driven improvements in electronic health records data quality: a systematic review. BMC Medical Informatics and Decision Making, 25, 298.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.