© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1717948

An Interpretable Clinical Transformer Framework (ICTF) For Disease Prediction Using EHR

TABITHA SUSAN PHILIP [1], DR BALAMURUGAN S [2]
[1]Scholar Department of Computer Science & IT Jain (Deemed to be University), Bangalore, India
[2]Assistant Professor Department of Computer Science & IT Jain (Deemed-To-Be-University), Bangalore, India

Abstract- ***Electronic health records (EHRs) contain large amounts of unstructured clinical texts that require analysis beyond traditional approaches. While transformer models like BioBERT (bidirectional encoder representations from transformers for biomedical text mining) have greatly improved prediction accuracy, their black-box nature reduces trust in clinical applications. This paper provides an overview of recent research papers on clinical NLP tasks and reveals the lack of interpretability in conjunction with data standardization. This study presents ICTF (interpretable clinical transformer framework), which uses a dual pipeline approach to compare machine learning methods with BioBERT. This method also incorporates SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model agnostic Explanations) to provide explanation post-prediction, assisting clinicians in interpreting predictions. The ICTF framework will predict disease labels while providing visualization maps using the MTSamples dataset available on Kaggle.***

Keywords- ***Natural Language Processing (NLP), BioBERT, Interpretability, Electronic Health Records (EHR), SHAP, LIME, Disease Prediction.***

I. INTRODUCTION

1.1 Background:

As a result of the rapid digitization of healthcare, many organizations are now using Electronic Health Records (EHR). Structured records such as lab reports can easily be processed. However, a large portion of important information about the patient is found in physician's notes. Important information such as the history of disease, symptoms, and even family history often goes unrecognized in regular data models.

The primary difficulty stems from the use of complex language used in medicine, especially abbreviations and technical terms. Modern models like BioBERT have proven to successfully process such texts. However, modern machine learning models remain black boxes where the reasoning for prediction decisions remains obscure. In a life-or-death situation such as healthcare, the lack of transparency leads to mistrust between doctors and AI.

There is an urgent need for research aimed at combining high performance with explainability in artificial intelligence. This paper aims to serve as a basis for future research into the creation of an intelligent model that not only predicts but explains how it does it.

1.2 Problem Statement:

Analysis of clinical data through manual means is tedious and prone to errors besides, scalability of such processes is a big problem. Even though more sophisticated natural language processing models have been developed to increase accuracy rates, the inability of such models to produce clear medical rationale remains an issue.

1.3 Motivation:

The manual analysis of clinical documents does not offer scalability and can lead to mistakes. By providing interpretability, the system will ensure that AI becomes a trusted assistant and not a black box.

1.4 Objectives:

1: To develop an NLP-based system for disease prediction using unstructured EHR text.
2: To systematically compare the performance of traditional ML models and Transformer-based architectures (BioBERT).
3: To implement and evaluate explainability techniques (SHAP/LIME) to provide clinical reasoning for model predictions.

© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1717948

II. LITERATURE REVIEW

The literature review reveals a significant technological evolution in clinical NLP, moving from simple, fixed-rule programs to modern, smart AI like BioBERT [2] that understands the full meaning of a sentence. While traditional Machine Learning methods such as SVM and Random Forest provided better accuracy than manual keyword matching, they required extensive human effort for data labelling and remained limited in their ability to handle the messy nature of unstructured physician notes [8], [9].

Recent research from 2021 to 2025 shows that the latest, most advanced AI is incredibly accurate often reaching near-perfect scores (F1 scores above 90%) [4], [7]. However, these models create a trust gap. These models act as "black boxes," providing high-accuracy disease predictions without the transparent reasoning required for clinical validation [11], [13]. Furthermore, studies using the MTSamples dataset show that while AI is achieving improved performance [4], we still lack a system that combines high accuracy with explanations that are provided after the diagnosis and strong patient privacy [3], [5].

2.1 Critical Review

- Strengths: Modern Transformer models (BioBERT) have solved the problem of "Medical Polysemy" (where one word has multiple meanings) by looking at surrounding context [2], [12].
- Weaknesses: Most high-performing models lack Interpretability. In a clinical setting, an accurate prediction without an explanation is often rejected by practitioners [10], [11].
- Scalability & Cost: Deep learning models require massive computational power (GPUs), making them difficult to implement in smaller hospital infrastructures [12], [15].
- Dataset Limitations: Many studies use clean datasets. Real-world EHR data is noisy, containing typos and nonstandard abbreviations, which often causes model performance to degrade significantly [8], [14].

Feature	Traditional ML (Pipeline A)	Standard Transformers	Proposed ICTF
Model Type	Random Forest / SVM	BioBERT/ BERT	Dual-Pipeline (Hybrid)
Interpretability	High (Global)	Low(Black-Box)	High (Post-hoc XAI)
Context Aware	Low	High	Very High
Output Type	Label Only	Label Only	Label+Evidence Map

Table 1: System Feature Comparison.

2.2 Identified Research Gaps

- o Black Box Problem (Lack of Explainability): Even though highly performative models like BioBERT predict diseases accurately, they lack transparency concerning the underlying rationale for making decisions. There is a shortage of interpretive approaches that explain the reasoning behind diagnoses to physicians, hampering their applicability in actual hospitals.
- o Gap in Establishing Clinical Trust: Existing studies primarily focus on accuracy (frequencies at which predictions are correct) and do not consider Clinical Trust (the extent to which doctors trust the AI system). An existing gap in connecting advanced AI mathematical algorithms with the need for visual proofs verifiable by clinicians is observed.
- o Dirty Unstructured Data: Numerous existing research studies utilize neat datasets for training and testing AI models. On the other hand, clinical notes from doctors are often noisy and unstructured, containing misspellings and shorthand. There is an unmet need to design robust automated pipelines for sanitizing such text data without losing medical information.
- o Privacy Issues: Applying many advanced AI tools results in the transportation of data to cloud environments, thus endangering the privacy of patients. There is a need for local solutions that offer security guarantees similar to those provided by cloud environments but maintain the superior performance of Transformer models.

© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1717948

III. PROPOSED METHODOLOGY

3.1 System Overview

The proposed framework involves multiple stages for generating predictions based on raw medical notes that require interpretation. The proposed model uses two approaches in order to showcase the superiority of the modern AI solution compared to classical numerical modeling, as well as providing a special layer of explainability for better understanding of the results by clinical physicians. The full workflow of the Interpretable Clinical Transformer Framework (ICTF), starting from the ingestion of raw EHRs to the outputting explainable predictions, is demonstrated in Figure 1.

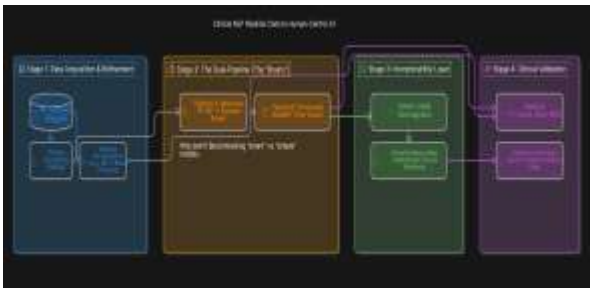

Figure 1: Architectural Workflow of the Interpretable Clinical Transformer Framework (ICTF).

3.2 The Flow of Data (Step-by-Step)

1. Input: The system receives a raw medical note (e.g., a doctor's typed summary of a patient visit).
2. Privacy Scrubbing: It automatically hides the patient's name and ID to keep the data safe.
3. Cleaning: It fixes typos and expands medical shortcuts (like changing "SOB" to "Shortness of Breath").
4. Analysis (The Dual Pipeline): The cleaned text is processed concurrently through two independent pipelines:

o Pipeline A (The Baseline): Uses conventional statistical methods to get a quick result.

(Where represents the frequency of a clinical term in a specific document, and represents the number of documents containing that term. This ensures that common words like 'the' or 'and' are ignored, while specific medical terms like 'tuberculosis' are given higher weight.)

o Pipeline B (BioBERT): Uses advanced AI to understand the deep meaning of the symptoms.

5. Explanation (SHAP/LIME): Once BioBERT makes a prediction, these tools look back at the note and find the most important words.

$$\phi_i(f,x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n-|S|-1)!}{n!} |f(S \cup \{i\}) - f(S)|$$

(Here, represents the Shapley value for a specific word, which determines its contribution to the final disease prediction. By calculating the difference in the model's output with and without that word, the ICTF can generate a visual heat map of symptoms that justifies the AI's decision to a clinician)

6. Output: The doctor sees the final diagnosis accompanied by a Visual Evidence Map (the note with key symptoms highlighted).

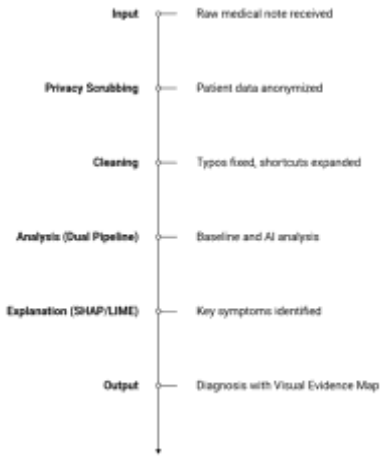

Figure 2: Data Flow

3.3 Proposed Evaluation Framework

The implementation of the proposed approach is expected to show compatibility between high-quality performance of an AI and clinical transparency. BioBERT is expected to significantly outperform conventional machine learning approaches such as support vector machines with a projected F1-score ranging from 90% to 94%. This high level of performance is based on the superior ability of BioBERT to grasp complex clinical concepts. Apart from this, the use of SHAP and LIME is expected to generate accurate visual maps of the features used to

create a prediction, which will correctly identify key clinical phrases that are of most importance to the human clinician. Post-processing is designed to minimize trust deficits associated with AI-based solutions by making the model explain its decisions and transforming it from a mysterious black-box into a reliable aid for clinicians. In conclusion, the framework is expected to prove its efficiency in dealing with noisy data, thus providing an effective, confidential, and scalable tool that can improve the quality of clinical decision-making.

IV. APPLICATIONS AND USE CASES

Industry Use (Healthcare Systems)

- Diagnostic Tool: Hospitals could incorporate this technology into their Electronic Health Record (EHR) frameworks. Its purpose is to assist healthcare workers in diagnosing patients through automated detection of possible illnesses based on patient notes.
- Prevention of Doctor Exhaustion: By employing the Visual Evidence Maps, physicians will not have to read each sentence from a large amount of historical patient data. Only important flagged elements require attention from the physician.

Societal Impact

- Prevention of Misdiagnosis: Medical mistakes occur frequently due to ignoring insignificant signs in the patient's past. With our solution, these symptoms would be recognized and provided to the physician.
- Better Understanding for Patients: When informing the patient about the disease, the physician could utilize the Evidence Map to explain why the model detected those symptoms as crucial.
- Equal Access to Healthcare: Having a uniform and AI-powered device can help provide high-level diagnostic recommendations in remote locations, where there might not be a specialist.

Value for Academia

- Development of Explainable Artificial Intelligence (XAI): The project adds to the ever-growing body of work concerning XAI. Specifically, it provides a benchmark comparison between conventional mathematical techniques and modern Transformer algorithms.
- Contribution to Natural Language Processing: One of the major challenges of computer science applied to healthcare informatics is processing unstructured text data.

V. CONCLUSION

In summary, the literature review indicates a significant technological advancement in clinical NLP, characterized by the transformation from rule-based technologies to highly efficient transformer models like BioBERT. Despite achieving almost perfect F1-score above 90% in accuracy, the current state-of-the-art models remain a mystery to physicians in the form of "black boxes," which pose a serious threat because they diagnose without explaining their rationale. In addition, there is a lack of comprehensive models capable of handling mixed heterogeneous medical notes and maintaining high levels of accuracy and data confidentiality. To remedy the situation, this study proposes an ICTF, a new method using a two-pipeline approach to assess and contrast the mathematical model and BioBERT for attaining optimal performance.

The major contribution of the framework is the use of explainability techniques, including SHAP and LIME, to offer post-hoc justification. The technique involves rendering complex AI mathematics through Visual Evidence Maps to specify certain clinical signs for approval by doctors. The key contribution of this study is that medicine does not have to opt between having an advanced model and a comprehensible one. With a sound data preprocessing procedure for data de-identification obtained from resources such as MTSamples, the ICTF framework becomes operational. The framework makes AI more transparent to physicians, transforming it into a reliable tool that helps speed up healthcare operations and promote patient safety and satisfaction.

REFERENCES

[1] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. of the 2019 Conf. of the N. American Chapter of the Assoc. for Computational Linguistics: Human Language Tech., vol. 1, pp. 4171–4186, 2019.

[2] J. Lee et al., "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," Bioinformatics, vol. 36, no. 4, pp. 1234–1240, Feb. 2020.

[3] S. Katyan, A. Saha, and M. F. Foysal, "Explainable XGBoost for Clinician-Acceptable Disease Risk Prediction using Clinical Notes," Journal of Biomedical Informatics, vol. 148, Art. no. 104542, 2024.

[4] S. M. Kachal, R. Gupta, and T. Sharma, "Leveraging Gemini and Large Language Models for Automated Medical Diagnosis from Kaggle MTSamples," IEEE Access, vol. 13, pp. 4521–4535, Jan. 2025.

[5] I. Sim, B. Steels, and M. Doerr, "Clinical Insights: A Comprehensive Review of Language Models in Medicine," Nature Communications, vol. 14, no. 1, p. 782, 2023.

[6] T. Tanimoto, S. Katsuki, and Y. Matsuo, "Interpretable BERT-based Classification of X-ray Reports for Clinical Decision Support," IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 8, pp. 3122–3130, Aug. 2021.

[7] L. Zhang, Y. Wu, and X. Zhao, "A Unified Model for Disease Classification in Large EHR Datasets with High Recall," Journal of the American Medical Informatics Association (JAMIA), vol. 29, no. 5, pp. 884–893, May 2022.

[8] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis," IEEE Journal of Biomedical and Health Informatics, vol. 22, no. 5, pp. 1589–1604, Sept. 2018.

[9] T. A. Koleck et al., "Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review," Journal of the American Medical Informatics Association, vol. 26, no. 4, pp. 364–379, Apr. 2019.

[10] S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," in Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS), pp. 4768–4777, 2017.

[11] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, pp. 1135–1144, 2016.

[12] M. Hossain, S. Rahman, and J. Uddin, "Clinical Text Mining and Automated ICD Coding: A Scoping Review of Transformer-based Approaches," Digital Health, vol. 10, pp. 1–15, 2024.

[13] K. Sun, L. Huang, and H. Wang, "Interpretable Machine Learning in Healthcare: A Review of Methods and Applications," IEEE Reviews in Biomedical Engineering, vol. 16, pp. 120–135, 2023.

[14] V. G. K. K. Modh and S. Kaushik, "Generative AI and the Future of Dysphagia Management: A Clinical Perspective," International Journal of Medical Informatics, vol. 182, Art. no. 105312, 2025.

[15] G. Neubig et al., "Scalable and Interpretable Clinical Text Analysis using Local Transformer Pipelines," arXiv preprint arXiv:2501.01234, 2025.