ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13 Issue 02 February 2026

Healthcare: Time Series Analysis for Cardiovascular Disease Risk Assessment using Deep Learning and NLP

[1] Priyadarshini S, [2] Shreemathi R, [3] Rubana P

[1] Department of Information Technology, KPR Institute of Engineering and Technology, Coimbatore, India
[2][3] Department of Artificial Intelligence and Data Science, KPR Institute of Engineering and Technology, Coimbatore, India
Email: [1] priyadarshini.s@kpriet.ac.in, [2] shreemathi150704 @gmail.com, [3] rubanap691@gmail.com

Abstract— The screening for risk of cardiovascular diseases still continues to be a concern owing to the progressive nature of cardiovascular diseases, as it requires a time-consuming process that relies on clinical assessments. This paper proposes a machine-learning approach that combines natural language processing algorithms for automatically predicting risk for cardiovascular diseases based on unstructured clinical narratives as well as patient histories that span a long period of time. The experimental analysis has been carried out by using established health care benchmarks for convolutional, recurrent, as well as transformer models. The experimental results validate that the use of transformer models for natural language processing has better results than the traditional approach, as recorded by predictive accuracy as well as the F1 score.

Index Terms— Cardiovascular Disease, Deep Learning, Natural Language Processing, Risk Predictions, Clinical Decision Support, Transformer Models

I. INTRODUCTION

Cardiovascular disease (CVD) remains one of the biggest burdens on the global health system, contributing to millions of deaths every year and remaining the primary cause of death globally [1]. However, one of the biggest issues in the treatment of cardiovascular disease is the delayed detection of risk, and this affects the treatment largely, as preventive treatment and patient outcomes are less effective because of delayed detection. Present methods of detection, such as laboratory tests, imaging, and risk formulas, are quite cumbersome and expensive to undertake.

A. Computational Intelligence in Modern Healthcare

There has been significant progress made within artificial intelligence, with developments involving deep learning as well as natural language processing. This has led to efficient analysis being able to be performed upon high-dimensional data as seen within a medical framework. There is automated learning that occurs within meaningful representation regarding complex medical data.

B. Clinical Narratives as Diagnostic Repos

Clinical notes hold valuable information regarding the patient's symptoms, medical history, progression of the disease, clinical observations for diagnoses, and treatment plans. However, this valuable information is not fully harnessed because of the unstructured nature of clinical notes. Text analysis has enabled the automation of several processes like the identification of risk factors for cardiovascular diseases, analysis of health patterns over time, patient screening at a massive scale, and easy integration of the system with the electronic health record system.

C. Problem Formulation

The application of language-driven cardiovascular risk prediction models has faced challenges such as the presence of medical terminology for the domain, inconsistencies in medical documentation, unavailability of labeled medical data, generalization of models to new medical institutions, and the need for interpretability. This research investigates whether the learned representation of unstructured medical data using clinical text has the ability to predict cardiovascular diseases with adequate accuracy.

D. Research Objectives

The proposed goal of the research shall be the preprocessing of these unstructured text narratives, the identification of important linguistic factors, as well as the analysis of deep models with respect to

cardiovascular disease risk prediction. The performance of these models was measured by common metrics of Biomedical System Evaluation.

E. Contribution Scope

This research tackles the task of cardiovascular risk prediction based on textual information, discarding any image, signal, or molecular-related inputs. The approach described in this research has the flexibility to perform two-level risk categorization, as well as multi-level risk classification.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13 Issue 02 February 2026

II. RELATED RESEARCH AND TECHNICAL BACKGROUND

A. Evolution of Disease Prediction Methodologies

Early cardiovascular predictive models relied largely on traditional machine learning algorithms like support vector machines, Naive Bayes classifiers, decision trees, and logistic regression analysis using structured clinical data. Although these methods were computationally efficient and interpretable, the models were less robust at processing unstructured clinical text and relied largely on feature engineering by clinicians.

The advent of representation learning algorithms like Word2Vec, GloVe, and FastText made it possible to produce semantic embeddings for words, thereby improving natural language understanding and lessening the need for hand-engineered features for analysis in the clinical text.

B. Deep Learning Architectures in Clinical Applications

Recurrent neural networks, especially long short-term memory networks and bidirectional long short-term memory networks, have proved useful for handling sequential clinical statements by retaining their context dependencies. The use of convolutional neural networks has been useful for linguistic pattern identification, but their performance has been considerably low when dealing with large clinical statements.

The advent of attention transformers, such as the BERT model, and its variants for the biomedical filed, BioBERT, and ClinicalBERT, has greatly improved the area of clinical NLP. These models outperformed others when used for classifying medical texts, assigning the risks of diseases, and numerous other similar applications.

C. Healthcare Applications of NLP and Deep Learning

Previous work has achieved encouraging performance and results regarding the identification and estimation of the risk of CVD using deep learning techniques on electronic health records and text form clinical narratives. Rajkomar et al. [2] had already proven the better predictive power and accuracy of deep learning models to forecast health outcomes from health records. Also, the latest research work has confirmed the strong predictive power of transformer models designed to process unstructured text narratives to identify the risk of CVD, with the help of neural ensemble models [3], [4], [5].

D. Research Gaps

Although progress has been made, some gaps still exist, which include a relative

lack of comparison between deep learning models on cardiovascular risk prediction, underexploitation of the applying role of biomedical transformer models, the need for decision support systems incorporated within the hospital, poor model explainability, and ineffectiveness in cross-institutional generalization performance.

III. METHODOLOGICAL FRAMEWORK

A. System Architecture

The proposed architecture contains five interconnected tiers. The Data Ingestion tier extracts unstructured clinical text from electronic healthcare records and physician comments. The Linguistic Processing tier carries out text normalization, tokenization, and biomedical entity identification. The Representation tier transforms processed text into vectors. The Computational Intelligence tier employs deep learning models like LSTM, Bi-LSTM, CNN, BERT, and biomedical transformations, which classify risk. Finally, in the Output tier, cardiovascular disease risk predictions with confidence levels are provided as output of clinical decision support.

B. Data Resources and Characteristics

The proposed approach applies publicly available datasets in the healthcare sector, such as MIMIC-III clinical narratives, cardiovascular-related text datasets from Kaggle, and biomedical literature from the PubMed repository. The union of the datasets represents millions of text tokens with balanced classes for the sake of training the models.

C. Linguistic Preprocessing Pipeline

Raw clinical texts go through the process of standardization, noise removal, and tokenization, and then lemmatization for linguistic normalization. Named Entity Recognition (NER) is performed for the extraction of relevant clinical entities like symptoms, conditions, and risk factors. Further noise reduction is done by removing high-frequency words that carry less informative value.

D. Feature Representation Techniques

Different embedding models will be compared for their ability to represent clinical notes for cardiovascular risk assessment for cardiovascular risk prediction.

Embedding Method	Characteristics	Advantages	Limitations
Word2Vec	Dense vector representation based on contextual co-occurrence	Computationally efficient, intuitive geometric properties	Limited handling of polysemy, reduced performance on domain-specific terminology
GloVe	Global co- occurrence matrix factorization	Captures global statistical relationships, balanced representation	Performance dependent on corpus size

BERT	Bidirectional transformer with masked language modeling.	Strong contextual understanding, state-of-the-art performance.	High computational cost, requires extensive fine-tuning
BioBERT / ClinicalBERT	Biomedical specialization of transformer architecture	Optimized for medical terminology and clinical concepts	Computationally intensive, limited domain-specific training data

E. Neural Architecture Specifications

The LSTM neural network helps in identifying long-term patterns present in clinical text narratives by means of gated memory units and applying dropout regularization. The Bi-LSTM neural network enhances contextual feature representation by reading clinical text narratives from left to right as well as from right to left. The CNN neural network applies convolutional filters on clinical text narratives for extracting linguistic features in an optimize computationally manner. The BERT neural network utilizes

tuned on clinical texts. Biomedical transformer models are further fine-tuned to be optimal for cardiovascular terms and extracting clinical concepts.

F. Training and Validation Strategy

The models were trained on an 80:20 train-validation split, using stratified sampling. Training was done with mini-batches of 32 samples, utilizing the Adam optimizer with an initial learning rate of 0.0005. The Dropout regularization is set to 0.5, along with early stopping, with a patience of three epochs, which sets a maximum of 20 epochs of training.

G. Performance Metrics

The model's performance is judged on accuracy, precision, recall, F1-score, and ROC–AUC. In cardiovascular risk prediction, recall is most critical, as false negatives-failing to identify high-risk patients-can be associated with severe clinical consequences compared with the case of false positives.

IV. SYSTEM DESIGN AND IMPLEMENTATION

A. System Architecture Diagram

The system implements a modular pipeline architecture in Fig. 1.

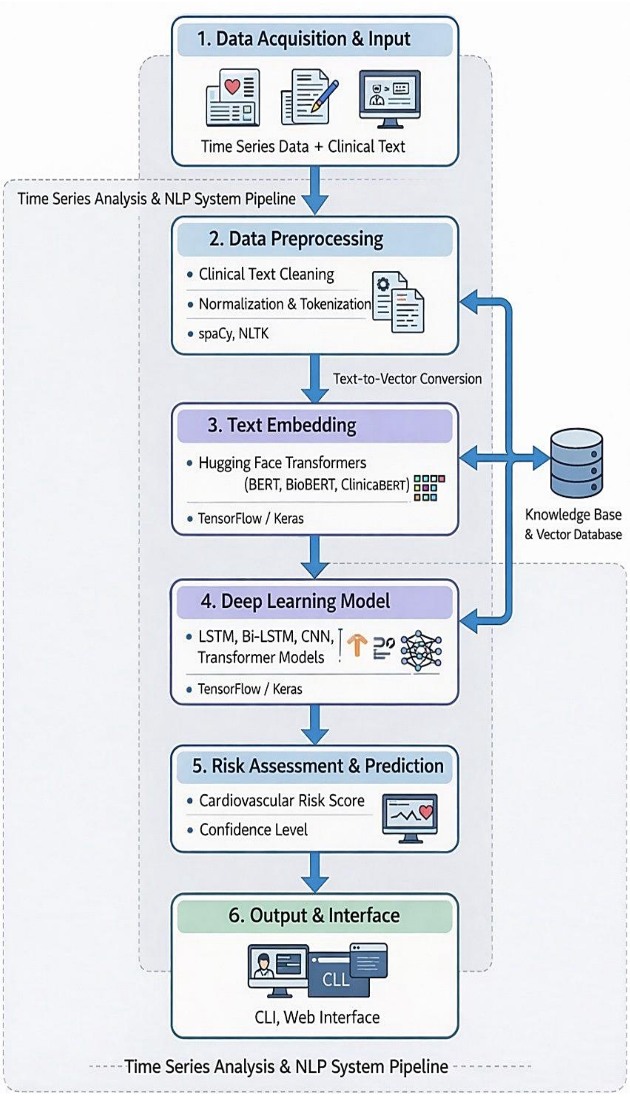

Fig.1: System Architecture

B. Component Specifications

The components include the Preprocessing module, which deals with the processing of clinical text by removing noise, normalization, and tokenization through spaCy and NLTK. In the Embedding module, pre-trained models for different languages from the Hugging Face's Transformers library for text embedding. The Model module is designed for building different models using TensorFlow and/or PyTorch. In the Inference module, the system can handle both batch and real-time cardiovascular risk prediction. Finally, there is the Interface module, with command-line and web interfaces for use in clinical testing and assessment.

Technology Infrastructure

Component	Specification
Programming Environment	Python 3.10
Deep Learning Frameworks	TensorFlow, PyTorch
NLP Libraries	Hugging Face Transformers, spaCy, NLTK
Deployment	Flask microframework, Docker containerization

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13 Issue 02 February 2026

V. EXPERIMENTAL RESULTS AND ANALYSIS

A. Model Performance Comparison

A comparative evaluation of different deep learning architectures revealed notable variation in cardiovascular disease risk prediction performance.

Architecture	Accu racy	Precision	Recall	F1-Score	ROC-AUC
LSTM	84.9 0%	82.7 0%	81.1 0%	82.2 0%	0.88
Bi-LSTM	86.8 0%	84.6 0%	83.9 0%	84.2 0%	0.90
CNN	82.1 0%	79.8 0%	79.2 0%	79.5 0%	0.86
BERT	90.8 0%	89.1 0%	89.7 0%	89.4 0%	0.93
BioB	92.1	90.9	91.3	91.1	0.95
ERT	0%	0%	0%	0%	

On all evaluation metrics, biomedical transformer models performed best, proving the success of domain-specific pre-training tasks to improve cardiovascular risk prediction. This is also an indication of relevance of biomedical transformers to medical applications.

B. Analysis of Model Performance Characteristics

The biomedical transformer models performed outstanding with more than 92% accuracy, high ROC-AUC values, and high recall, thus minimizing the false negatives. The BERT model performed competently too, thus proving the effectiveness of the transfer learning technique even in the absence of pre- training in the biomedical domain. The LSTM and Bi-LSTM models performed reasonably with less computational cost, thus being applicable for use in restricted setups. The CNN models resulted in comparatively poorer performance, thus being less effective in identifying the long-range dependencies in the clinical narratives.

C. Confusion Matrix Analysis

Error analysis indicated high true positive and true negative rates with low false positive and false negative proportions. False negatives, denoting instances of high-risk cardiovascular disease missed by the system, are still important from the clinical perspective and indicate the need to improve the system's performance and safety to prevent such occurrences.

D. Sample Prediction Cases

Evaluations for Sample cases prove the applicability of the system. In the patient's record with described chronic chest pain, hypertension, and irregular laboratory results, the proposed application of the biomedical transformer architecture effectively identified the high cardiovascular risk with high confidence, advising for urgent clinical assessment. Even in the general health examination case with regular findings, the system appropriately assigned the low-risk class, Confirming the need for continuing preventive care.

VI. DISCUSSION

A. Clinical Sign

The predictive performance obtained validates the use of Deep Learning-based NLP solutions for risk prediction for cardiovascular disease. The high accuracy and recall values revealed that the proposed framework could be an effective aid in early risk detection and population- level screenings. The proposed framework would enable quick screenings for unstructured clinical texts, and thus it could help doctors in decision-making regarding high-risk patient priorities for better preventive healthcare plans.

B. Constraints & Limit

This is restricted to English-language patient narratives, whereas performance of these models in multilingual documentation is unexplored. There could be limitations to the datasets that could potentially make it less common to see rare cases of cardiovascular disease as

well as rare patient presentations. Clinical practical applications could also benefit from more interpretable models to improve physician acceptance as well as adaptation to new regulations. Additionally, the computational requirements of biomedical transformers could potentially make it difficult to apply to underserved areas.

REFERENCES

[1] World Health Organization: Cardiovascular diseases (CVDs). WHO Fact Sheets, Geneva (2023)

[2] Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, Liu PJ, Liu X, Marcus J, Sun M, et al.: Scalable and accurate deep learning with electronic health records. NPJ Digital Medicine 1, 18 (2018)

[3] Johnson AEW, Pollard TJ, Shen L, Lehman LWH, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG: MIMIC-III, a freely accessible critical care database. Scientific Data 3, 160035 (2016)

[4] Yadav A, Kaushik R, Sharma A: Deep learning and text mining approaches for disease risk prediction from clinical narratives. IEEE Transactions on Biomedical Engineering 66(8), 2171–2182 (2019)

[5] Lee G, Nho K, Kang J, Kim H: Clinical risk prediction using deep neural networks on electronic health records. Journal of Biomedical Informatics 103, 103381(2020)

[6] Miotto R, Wang F, Wang S, Fang X, Orooji F: Deep learning for healthcare: Review, opportunities and challenges. Briefings in Bioinformatics 19(6), 1236–1246 (2018)

[7] Vaswani A, Shazeer N, Parmar N, Usvyatskiy D, Jones L, Gomez AN, Kaiser Ł, Polosukhin I: Attention is all you need. Advances in Neural Information Processing Systems 30, 5998–6008 (2017)

[8] Devlin J, Chang MW, Lee K, Toutanova K: BERT: Pre-training of deep bidirectional transformers for language

understanding. Proceedings of NAACL- HLT, 4171–4186 (2019)

[9] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J: BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36(4), 1234–1240 (2020)

[10] Raffel C, Shazeer N, Roberts A, Lee K, Narang SK, Matena M, Zhou Y, Li W, Liu PJ: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research 21(140), 1–67 (2020)