International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 15, Issue 02, February 2026)

Integrating Classical and Transformer-Based Model for Cardiovascular Disease Prediction in EHR Data

T. Charanya Nagammal[1], Dr. K. Chitra[2]

[1]*Research Scholar, Department of Computer Science, Madurai Kamaraj University, Madurai, Tamil Nadu, India.*
[2]*Associate Professor, Department of Computer Science, Quaid-E-Millath Government College, Chennai, Tamil Nadu, India*

***Abstract*-- Cardiovascular diseases (CVDs) continue to be the predominant cause of mortality and morbidity. Major trends in CVDs today include variations in prevalence across age and gender, along with risk factors like hypertension, high cholesterol, diabetes, obesity, smoking, and physical inactivity, driving increased research attention. Predictive analytics using machine learning (ML) models has shown potential in identifying individuals at high risk, improving treatment planning, and forecasting outcomes based on clinical data. The advanced utilization of text-based electronic health records (EHRs) supports personalized healthcare and CVD predictive modeling. This study examines the predictive performance of Support Vector Machines (SVM), Random Forest (RF), and Bidirectional Encoder Representations from Transformers (BERT) for cardiovascular disease prediction from clinical text. Our evaluation highlights each model's strengths and weaknesses concerning accuracy and robustness.**

***Keywords*-- Cardiovascular Diseases, Machine Learning, Predictive Analytics, EHR**

I. Introduction

Cardiovascular diseases (CVDs) significantly contribute to global morbidity and mortality among adults. This paper focuses on early disease prediction by analyzing its prevalence, associated risk factors, and physical fitness influences. Major behavioral risk factors include poor diet, physical inactivity, tobacco use, and excessive sugar consumption, while determinants like economic changes, urbanization, aging populations, poverty, stress, and genetics also play key roles. Management strategies involve treating hypertension, diabetes, and high cholesterol to mitigate cardiovascular risks and prevent complications. Accurate prediction and early diagnosis are crucial for effective treatment and improved outcomes.

Data mining plays a significant role in the medical field, demonstrated by research conducted over the past few decades. Factors such as diabetes, high blood pressure, high cholesterol, and abnormal pulse rate are crucial in predicting CVD. Machine learning techniques have become vital in medical applications, especially for disease prediction. In this paper, we investigate the effectiveness of various machine learning algorithms in predicting cardio vascular disease.

We employed Random Forest, Support Vector Machines, and BERT to build predictive models. The focus is on integrating classical machine learning models with BERT and evaluating their performance. The dataset used for this study was obtained from Kaggle and includes structured and unstructured data. All computations, preprocessing, and visualizations were conducted on Google Colab using Python

II. Literature Review

Machine learning has significantly enhanced the ability to uncover hidden patterns, enabling advanced analysis in clustering, classification, regression, and correlation tasks. Its integration has improved medical diagnostics by facilitating early detection of cardiovascular diseases, reducing the need for costly clinical tests, and lowering financial burdens on healthcare systems. Several studies have leveraged machine learning models to predict heart conditions using diverse datasets and algorithms. However, many efforts have yet to optimize predictive accuracy when utilizing electronic health records (EHRs) for CVD prediction, particularly with complex clinical and demographic data [1, 2]. Heart disease, also known as cardiovascular disease, remains a leading cause of death worldwide. Although machine learning methods have demonstrated promising results in forecasting certain medical disorders, they are not widely applied to predicting individual CVD survival in hypertensive patients using routinely collected big digital health data [4].

2.1 Related Works

Many researchers have developed cardiovascular disease prediction frameworks using data mining techniques and various algorithms, presenting future possibilities for improved outcomes. Some notable works include:

- In [5], an accuracy of 86.63% was achieved, utilizing a confusion matrix.
- In [6], machine learning algorithms like Naïve Bayes, Decision Tree, and K-Nearest Neighbors were used for disease severity prediction based on user symptoms.
- In [7], data mining techniques were applied to clinical datasets for classification and pre-processing, improving prediction accuracy.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 15, Issue 02, February 2026)

- In [8], BERT achieved mean AUC areas of 97.9%, 97.8%, and 97.8% for groups with hypertension, diabetes, and dyslipidemia.
- In [9], SVM and Random Forest models achieved accuracies of 55.55% for heart disease prediction.
- In [10], a combined confusion matrix analysis revealed differences in performance between MLP and other models.

The studies indicate that using classical models (SVM and Random Forest) alongside transformer-based models (BERT) yields improved prediction rates. Our proposed model uses a unified dataset for both classical and BERT-based predictions of cardiovascular disease risk.

III. Methodology

The proposed methodology integrates structured and unstructured EHR data for CVD prediction. Structured data includes variables like age, cholesterol, and blood pressure, while unstructured data derives from clinical text. Preprocessing involves handling missing values, removing outliers, and normalizing features. Feature engineering for structured data includes scaling and encoding categorical variables. Text data is tokenized and embedded using BERT's pre-trained transformer architecture. Model training was conducted in Python using Google Colab, evaluating SVM, Random Forest, and BERT on the same dataset to ensure consistency. Figure 1 illustrates the proposed model architecture

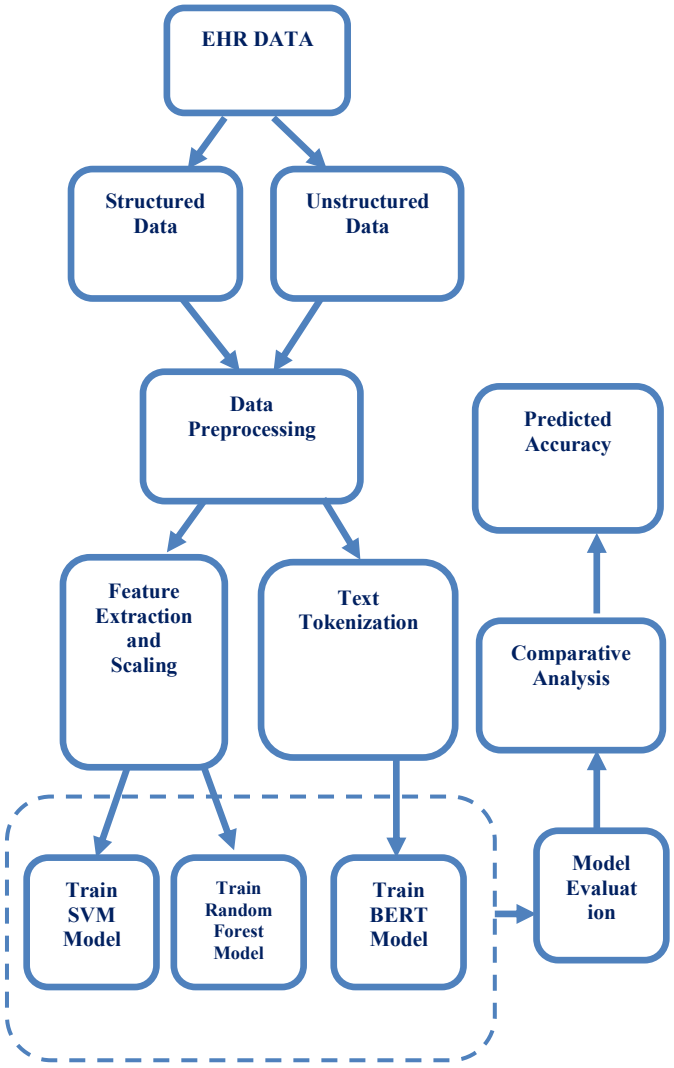

Fig 1. Proposed Model Architecture

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 15, Issue 02, February 2026)

3.1 Data And Preprocessing

The dataset used in this study was obtained from Kaggle, comprising 13 features, including age, gender, height, weight, diastolic and systolic pressures, cholesterol, glucose, alcohol consumption, smoking, and physical activity. It contains nearly 70,000 patient records with a target class 'cardio,' indicating CVD presence. Data cleaning process involved handling missing values, detecting outliers, and normalizing data. For structured data, feature engineering focused on the 'cardio' target, with scaling and categorical encoding.

3.2 Model Description

Feature Engineering is a critical step in the machine learning that involves transforming raw data into meaningful features that improve the performance of predictive models. It plays a crucial role in extracting useful information from datasets, particularly in structured data or text-based data for healthcare applications. The key concepts concentrate here are feature selection, feature transformation, handling categorical variable and missing data handling process. In cardiovascular disease prediction, properly engineered features from EHR data (e.g., age, cholesterol levels, systolic blood pressure) are crucial for achieving high model accuracy. For unstructured clinical text, feature extraction using transformer-based models like BERT helps in understanding the context of health conditions and symptoms. By implementing the classical model (SVM and Random Forest) it is sensitive to feature magnitudes, so scaling continuous variables using StandardScaler or MinMaxScaler ensures all features contribute equally to the model.And also, we could encode the categorical data variables like gender or cholesterol may be one-hot encoded or label-encoded to convert them into numeric form. When comes to transformer-based model (BERT) it requires text to be tokenized into input tokens. In our implementation, Hugging Face's tokenizer was used for this purpose. The embedding process is done by the BERT automatically which generates dense vector representations (embeddings) for each token, capturing the contextual meaning of words in clinical notes. The output embeddings from BERT are used directly as features for classification, replacing manual feature extraction from clinical text. Table 1 displays the comparison of Feature engineering Techniques for the proposed model algorithms.

Table 1-
Comparison of Feature Engineering Approaches

Model	Feature Engineering Techniques
SVM	Scaling, feature selection, and encoding categorical variables
Random Forest	Feature importance analysis, handling categorical and continuous data
BERT	Text tokenization and contextual embedding using a pre-trained transformer model

IV. Result And Discussion

As a result of our implementation, we can analyze both classical models together with the transformer-based model. Our result depends on both quantitative analysis and qualitative analysis. The metrics we have taken here to analyze the performance of both SVM and Random Forest are accuracy, precision, recall, and F1-score in Python. In Fig 2 we produce the visualization of performance comparison of three models Random Forest, SVM and BERT by using the performance matrices such as Accuracy, Precision, Recall and F1 Score values.

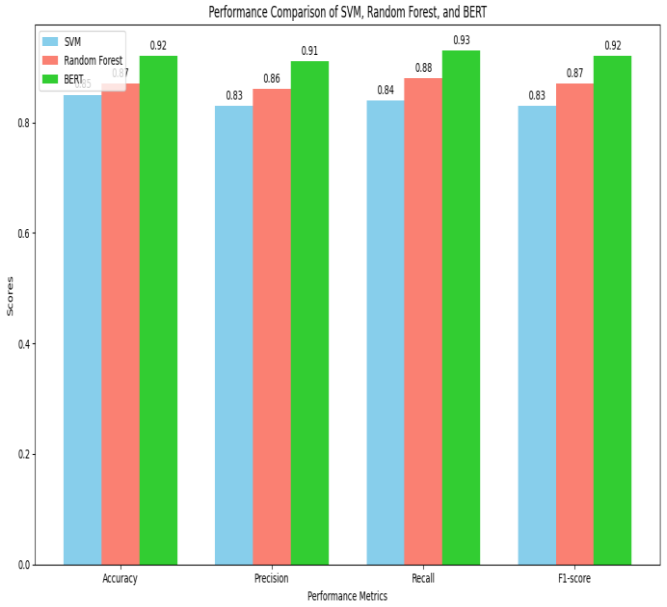

Fig 2. Performance comparison of SVM and RF and BERT

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 15, Issue 02, February 2026)

Each of the metrics can be plotted by using the following calculation methods.

- *Accuracy:* Measures how often the model makes correct predictions.

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}}$$

- *Precision:* Focuses on how many of the predicted positive cases are actually positive. It's critical when false positives need to be minimized.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

- *Recall (Sensitivity):* Measures how many actual positive cases were predicted correctly. Important for detecting true cases**.**

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

- *F1-Score:* Harmonic mean of precision and recall, providing a balanced metric.

$$\text{F1Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

By analyzing all these metrics for both classical models of SVM and Random Forest provides the overall better accuracy.

By using the Precision and recall we can understand whether one model is more conservative (higher precision) or more sensitive (higher recall) in predicting cardiovascular disease. The F1-score indicates which model provides the best balance between precision and recall, useful when both false positives and false negatives have significant costs. When we compare the both classical models with BERT model, it give more accuracy in all the metrics by using pre-trained language model uses contextual embeddings, eliminating the need for traditional text vectorization techniques, which are necessary for Random Forest and SVM in handling the input data. BERT achieves a much higher accuracy (e.g., 92%) compared to SVM (85%) and Random Forest (87%).

Precision for BERT (91%) surpasses both SVM (83%) and Random Forest (86%). Higher precision means BERT minimizes false positives, making it more reliable when predicting positive cases of cardiovascular disease, which helps in reducing the chances of unnecessary treatments or alarms. BERT's recall (93%) outperforms SVM (84%) and Random Forest (88%). This reflects its effectiveness in identifying most positive cases, important for ensuring patients with potential cardiovascular issues are correctly diagnosed. The F1-score, a balanced measure of precision and recall, is highest for BERT (92%). This demonstrates that BERT not only captures most positive cases but does so with fewer errors compared to the classical models.

V. Conclusion

This study demonstrates a comprehensive comparison of classical machine learning models (SVM and Random Forest) and a transformer-based model (BERT) for cardiovascular disease prediction. Our findings highlight BERT's superior performance in handling unstructured text data due to its contextual embedding capabilities. Unlike traditional models requiring manual feature engineering, BERT's pre-trained embeddings capture deeper semantic relationships within clinical narratives. This study emphasizes the importance of leveraging advanced NLP techniques for EHR-based predictive analytics, providing a robust framework for improving disease detection in real-world healthcare applications.

REFERENCES

[1] Khan MA, Algarn F. A healthcare monitoring system for the diagnosis of heart disease in the iomt cloud environment using msso-anfis. IEEE Access. 2020;8:122259–69.

[2] Javeed A, Zhou S, Yongjian L, Qasim I, Noor A, Nour R. An intelligent learning system based on random search algorithm and optimized random forest model for improved heart disease detection. IEEE Access. 2019;7:180235–43. https:// doi. org/ 10. 1109/ access. 2019. 29521 07.

[3] Health M. Ministry of Health, COVID-19. Accessed: October 2020. 2020. https:// covid 19. moh. gov. sa/.

[4] Ambale-Venkatesh B, Yang X, Wu CO, Liu K, Hundley WG, McClelland R, Gomes AS, Folsom AR, Shea S, Guallar Eet al. Cardiovascular event prediction by machine learning: the multi-ethnic study of atherosclerosis. Circ Res. 2017;121(9):1092–101.

[5] Comparison Of Machine Learning Algorithms For Heart Disease Prediction by Ayat Bahaa ABDULHUSSEIN, Turgay Tugay BİLGİN,Bursa Teknik Üniversitesi, Bilgisayar Mühendisliği Bölümü, Bursa, Türkiye, Bursa Teknik Üniversitesi, Journal of Technology and Applied Sciences

[6] Sonam Nikhar, A.M. Karandikar, "Prediction of Heart Disease Using Machine Learning Algorithms", International Journal of Advanced Engineering, Management and Science (IJAEMS), Vol-2, Issue-6, June- 2016, ISSN: 2454-1311, pp 617-621.

[7] Amin Ul Haq , Jian Ping Li , Muhammad Hammad Memon , Shah Nazir and Ruinan Sun, "A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms", Hindawi Mobile Information Systems, 2018, pp. 01-21.

[8] BERT-based model to predict cardiovascular disease by analyzing healthcare utilization behavior of patients newly diagnosed with metabolic diseases, Dongyup Shin, CONNECT-AI Research Center, October 9, 2023.

[9] Comparative Prediction Performance with Support Vector Machine and Random Forest Classification Techniques, International Journal of Computer Applications (0975 – 8887) Volume 69– No.11, May 2022

[10] Risk prediction of cardiovascular disease using machine learning classifiers, Madhumita Pal, Smita Parij, Ganapati Panda, Kuldeep Dhama, Ranjan K. Mohapatra, May 23, 2022

[11] Transformers for cardiac patient mortality risk prediction from heterogeneous electronic health records. VTT Technical Research Centre of Finland Ltd., 33101 Tampere, Finland, Sci Rep 2023 Mar.

[12] Fine Tuning Bert Based Approach for Cardiovascular Disease Diagnosis, Naga Suneetha, A. R. V. ., & Mahalngam, T. . (2023). Fine Tuning Bert Based Approach for Cardiovascular Disease Diagnosis. *International Journal of Intelligent Systems and Applications in Engineering*, *11*(6s), 59–66.

[13] "Adapting Transformer-Based Language Models for Heart Disease Prediction"Houssein, E.H., Mohamed, R.E., Hu, G. *et al.* Adapting transformer-based language models for heart disease detection and risk factors extraction. *J Big Data* **11**, 47 (2024).

[14] "Med-BERT: Pretrained Contextualized Embeddings on Large-Scale Structured Electronic Health Records for Disease Prediction"Rasmy, L., Xiang, Y., Xie, Z. *et al.* Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digit. Med.* **4**, 86 (2021).

[15] "Heart Disease Risk Factors Detection from Electronic Health Records Using Transformer Models" Houssein, E.H., Mohamed, R.E. & Ali, A.A. Heart disease risk factors detection from electronic health records using advanced NLP and deep learning techniques. *Sci Rep* **13**, 7173 (2023).

[16] "Heart Disease Prediction Using Machine Learning Algorithms: Performance Analysis" - Conference: 2024 3rd International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE).