International Journal of Future Engineering Innovations

Enhancing CVD Diagnosis with Machine Learning Algorithms

BNSL Rajeshwari Renu [1*], Kopalle Laxmi Meenakshi [2], B Vibhooshitha [3], Dr. D Shravani [4]
[1] BE, AI&DS, VI Sem, Stanley College of Engineering & Technology For Women, Telangana
[2] BE, AI&DS, VI Sem, Stanley College of Engineering & Technology For Women, Telangana
[3] AI&DS, VI Sem, Stanley College of Engineering & Technology For Women, Telangana
[4] Associate Professor, ADCE dept, Stanley College of Engineering & Technology For Women, OU, Hyderabad, Telangana

* Corresponding Author: **BNSL Rajeshwari Renu**

Article Info

ISSN (online): 3049-1215
Volume: 02
Issue: 04
July – August 2025
Received: 10-05-2025
Accepted: 12-06-2025
Published: 05-07-2025
Page No: 46-50

Abstract
Cardiovascular diseases are one amongst the top killers worldwide. The way we live, the food that we eat, metabolic rate, activity level leads to such health issues. Some of these include health factors like obesity, high blood pressure, high cholesterol levels, smoking, diabetes, and age. The food that we eat also plays an important role in heart risk. These days people are preferring spicy, deep fried food, this food in turn causes obesity and high bp leading to heart problems. Stress also plays a vital role in cardiovascular risk. Quick detection of this risk in individuals allows them to take a quick medical action. The standard methods cannot be completely reliable and are slow. To provide an effective solution to this, we used machine learning algorithms. We trained some models like Decision Tree model and Random Forest model on our dataset that includes most of the key factors that lead to this risk. We used feature selection techniques to boost our model's performance and reliability. We used bagging with extra randomness to predict the results. Prevention is always better than cure so early detection is always better than being at risk.

DOI: https://doi.org/10.54660/IJFEI.2025.2.4.46-50

Keywords: Cardiovascular Disease, Machine Learning, Decision Tree, Random Forest, Bagging, Risk Prediction, AI in Healthcare

1. Introduction

Mortality rates are increasing worldwide due to cardiovascular diseases. Some of the conditions of CVD include heart attacks, strokes, coronary disease, artinery diseases. Health factors like obesity, high blood pressure, high cholesterol levels, smoking, diabetes, and age contribute to such risks. The food that we eat also plays an important role in heart risk. Detecting such deadly diseases on beforehand can save lives.

In our project, we implemented machine learning algorithms like Decision Tree and Random Forest algorithms but they were prone to over fitting. To build an efficient model that is reliable we used bagging with extra randomness that gave us the best results. The models were trained on the latest dataset that consists of key factors like age of the person, blood pressure level, cholesterol level, glucose levels, body mass index (BMI), and lifestyle habits which include smoking, alcohol consumption.

The primary contributions of this project are:

1. We used models like Random Forest, Decision Tree to predict the risk of cardiovascular diseases.
2. The evaluation of the model was done using metrics like accuracy, precision, recall, and F1-score.
3. The models were prone to overfitting, to reduce overfitting we used the bagging technique with extra randomness.
4. This project was implemented using Google Colab, which helped to test real-time data and also visualise the results. This structure of this paper is as follows: Section 2 explains the related work followed by proposed framework, result analysis and future enhancements.

2. Related Work

Machine learning is one of the important areas in the medical field, particularly for predicting diseases like cardiovascular disease there were many researches held on the machine learning algorithms to help in early detection and prediction of diseases.

Akanbi [1] implemented the Naïve Bayes classifier on a Nigerian hospital based patient records, and explained about its ability to handle in medical data. In a separate study, Biswas *et al.* [2] created a CVD prediction framework that incorporated multiple feature selection techniques and classifiers, demonstrating the significant impact of feature optimization on model performance.

Khan *et al.* [3] introduced a novel machine learning methodology that utilized ensemble models like Random Forest, which exhibited enhanced predictive accuracy.

A comparative analysis by Osei-Nkwantabisa and Ntumy [4] evaluated various algorithms, including Support Vector Machines (SVM) and k-Nearest Neighbors (kNN), reinforcing the reliability of machine learning-based diagnostics compared to traditional scoring systems.

Dritsas and Trigka [5] implemented and tested advanced classifiers on real-world cardiac datasets.

Sarra *et al.* [6] chose a Chi-Square-based feature selection technique to boost the accuracy and reducing the risk of overfitting.

Similarly, Baghdadi *et al.* [7] developed a hybrid ensemble model that combined deep learning with traditional techniques, aimed at identifying early signs of CVD in asymptomatic patients.

Muibideen and Prasad [8] in their research used Bayesian networks for predicting the risk of heart diseases. Ahmad *et al.* [9] focused on the importance of optimal feature selection in hybrid machine learning and deep learning models, highlighting the detrimental effects of redundant or irrelevant features.

García-Ordás *et al.* [10] applied deep learning techniques with feature augmentation to manage high-dimensional cardiac data, achieving impressive classification accuracy.

Addressing the needs of underserved healthcare areas, Paulino-Moreno and Iparraguirre-Villanueva [13] advocated for swift machine learning-driven diagnostic approaches.

Ogunpola *et al.* [12] implemented automated feature selection alongside traditional machine learning classifiers to propose practical strategies for clinical workflow integration.

Khan *et al.* [14] researched on the influence of patient history on the performance of predictive models, confirming Random Forest as the best model.

Theerthagiri and V. [15] examined the use of Gradient Boosting in conjunction with Recursive Feature Elimination (RFE), demonstrating that managing model complexity is crucial for performance, especially with medium to small datasets. Further comparative studies have shown that ensemble models, particularly those that integrate Logistic Regression, SVM, and Random Forest, generally outperform individual algorithms, especially in dealing with incomplete or noisy clinical data.

Taking these papers as our primary reference, in our research we considered Decision Tree, Random Forest, and Bagging methods.

These have been perfectly trained and evaluated on a cardiovascular dataset. Our main aim is to minimize overfitting and improve the accuracy of our model so that it is effective and reliable.

3. Proposed Framework

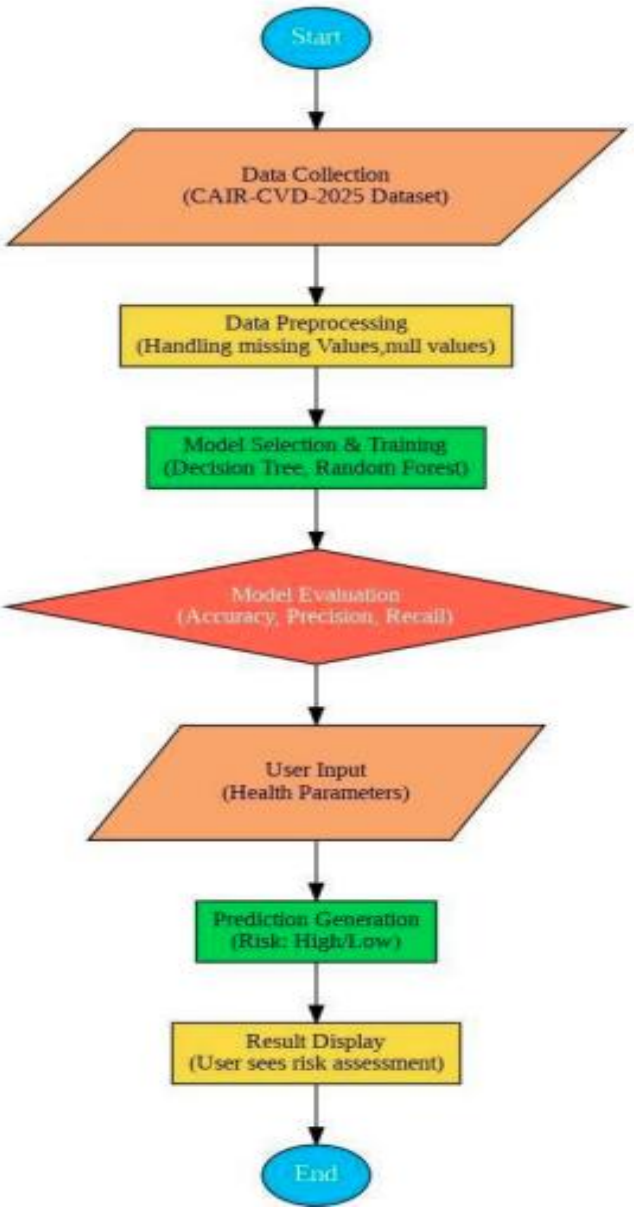

Fig 1: Work flow of proposed method

To enhance early detection of cardiovascular disease (CVD), we applied supervised machine learning algorithms on a curated healthcare dataset. The dataset, derived from CAIRCVD2025, contains critical clinical and physiological attributes such as age, blood pressure, cholesterol, glucose levels, and BMI. Prior to model training, we performed data preprocessing steps including handling missing values and applying normalization techniques to ensure the data was clean and consistent.

We trained our models Decision Tree, Random Forest, and Bagging using optimized hyperparameters to boost predictive accuracy while mitigating the risk of overfitting. For ensemble methods, we fine-tuned parameters such as tree depth, number of estimators, and bootstrap configuration. These changes helped the models to extract key patterns from the data while maintaining the stability of the model.

To evaluate model performance, we employed standard classification metrics: accuracy, precision, recall, and F1-score. Among the models trained and tested, Random Forest produced accurate predictions on the test set. This model has

shown the best results over decision trees and when extra randomness was introduced the model outperformed.

4. Experimental Results

When the models were evaluated on the dataset decision trees and random forest were prone to overfitting. We focused on evaluating the performance of the models and bagging classifier for cardiovascular disease prediction so that our prediction is trustworthy. The cardiovascular dataset we chose consists of 809 patient records with key health parameters like BMI, age, gender, blood pressure. The dataset was loaded, preprocessed, trained and tested. Notably, the Bagging model enhanced with additional randomness demonstrated the highest generalization capability and consistency across data splits. While the Decision Tree and Random Forest classifiers achieved perfect training accuracy (100%), both exhibited signs of overfitting during testing phases. The Bagging model achieved a test accuracy of 93. 5% but was prone to overfitting, while the one with extra randomness scored 92. 5% and was effective. The Bagging model achieved a precision of 91. 13%, recall of 96. 58%, and an F1-score of 93. 78%, alongside a near-perfect Area Under the ROC Curve (AUC ≈ 1. 0). These findings suggest that Bagging is a more reliable option for clinical applications, offering stable and accurate predictions with reduced variance and improved robustness.

Table1: Performance Comparison of Machine Learning Models for CVD Prediction

Metric	Decision Tree	Random Forest	Bagging	Bagging (Extra Randomness)
Train Accuracy	100%	100%	100%	95.1%
Test Accuracy	94%	95.5%	93.5%	92.5%
Precision	90.2%	91.6%	91.13%	90.4%
Recall	95.7%	96.2%	96.58%	95.1%
F1-Score	92.8%	93.8%	93.78%	92.7%
ROC AUC Score	~0.98	~0.99	~1.0	~1.0

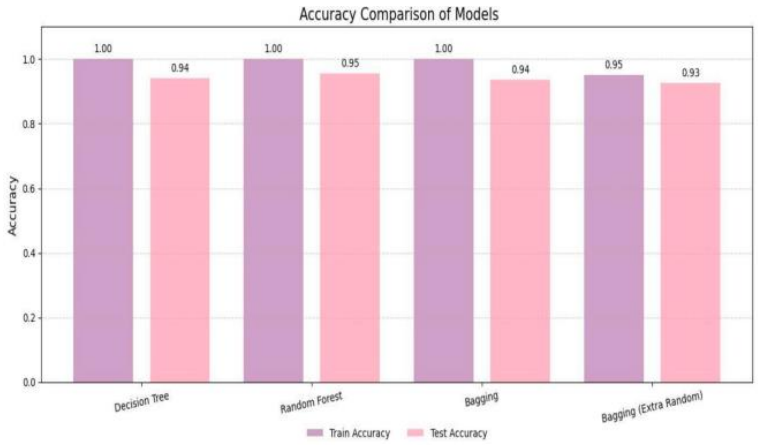

Fig 2: Accuracy comparison of machine learning models

The bar chart compares the prediction accuracies of four different machine learning algorithms tested on a cardiovascular dataset. Although all models attained 100% training accuracy (save Bagging with Extra Randomness at 95%), test accuracy was somewhat lower, suggesting possible overfitting in Decision Tree and Random Forest.

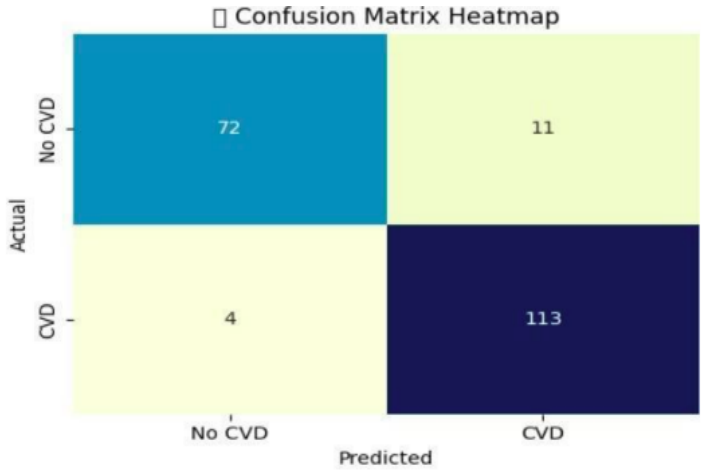

Fig 3: Confusion Matrix Analysis of Bagging with Extra Randomness

The model predicted 113 positive cases correctly and 72 negatives as negative . With only 11 false positives and four false negatives, few misclassifications were found. These results shows how efficient our model is in predicting cardiovascular risk . The ensures trustworthiness of the prediction.

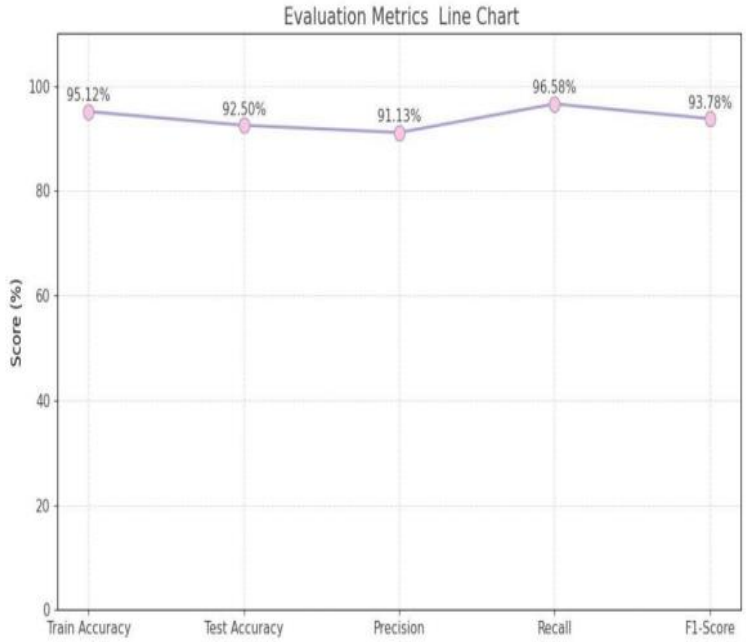

Fig 4: Evaluation of Precision, Recall, and F1-Score for Bagging with Extra Randomness Model

High recall guarantees almost all genuine instances are found; high precision implies less false alarms. These results confirmed that the Bagging with Extra Randomness model is efficient, dependable and reliable for CVD prediction.

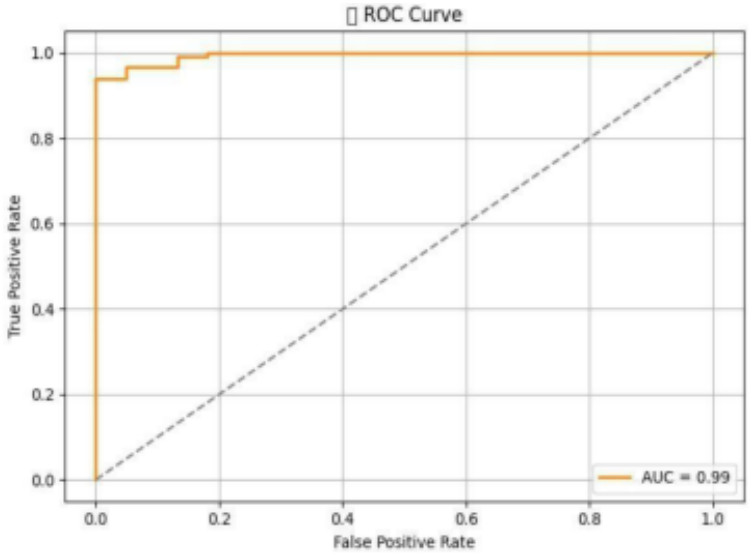

Fig 5: ROC Curve for Bagging with Extra Randomness Model

When we look at the ROC curve (which is like a report card for the model), we can see it hugs the top-left corner really tightly. This means two great things: First, the model catches almost all true cases of CVD (very few sick people slip through). Second, it hardly ever raises false alarms by mistaking healthy people for sick ones. The near-perfect score of 1. 0 for the area under the curve (that's what AUC stands for) confirms that this model performs exceptionally well overall.

5. Conclusion and Future Scope

This project highlights the successful use of machine learning techniques—specifically Decision Tree, Random Forest, and Bagging classifiers—to predict the risk of cardiovascular disease (CVD) using organized patient health data. Among all the models we tested, the ensemble methods stood out, with Bagging with Extra Randomness delivering the best results. This approach helped reduce overfitting and provided consistent performance across various evaluation metrics, including accuracy, precision, recall, and F1-score.

The strong performance of these models suggests they could play a key role in enabling early diagnosis and raising awareness about cardiovascular health. This proactive approach could help lessen the burden on healthcare systems

by allowing for timely medical interventions.

Looking ahead, this work could be expanded to include predictions for specific cardiovascular events, such as heart attacks, arrhythmias, or strokes, which would allow for more detailed diagnostics. To make the system more user-friendly, we could integrate these features into mobile or web applications that offer interactive evaluations and personalized health insights. Wearable devices could also be used for automatic monitoring and timely notifications, which would enhance the system's functionality. These improvements would make the system more valuable for proactive healthcare and drive innovation in digital health.

6. References

1. Akanbi OB. Prediction of heart disease risk among patients in Federal Medical Centre, Abeokuta using Naïve Bayes. Asian J Probab Stat 2024;26(10):46–63.
2. Biswas N, Al-Zahrani F, Moni MA, *et al*. Machine learning-based model to predict heart disease in early stage employing different feature selection techniques. BioMed Res Int 2023;2023:6864343.
3. Khan A, Qureshi M, Daniyal M, *et al*. A novel study on machine learning algorithm-based cardiovascular disease prediction. Health Soc Care Community 2023;2023:1406060.
4. Osei-Nkwantabisa AS, Ntumy R. Classification and prediction of heart diseases using machine learning algorithms. University of Texas Rio Grande Valley and University of Ghana Research Publication; 2023.
5. Dritsas E, Trigka M. Efficient data-driven machine learning models for cardiovascular diseases. Sensors 2024;24(18).
6. Sarra RR, Dinar AM, Mohammed MA, *et al*. Enhanced heart disease prediction based on machine learning and χ^2 statistical optimal feature selection model. Designs 2022;6(5):87.
7. Baghdadi NA, Abdelaliem SMF, Malki A, *et al*. Advanced machine learning techniques for cardiovascular disease early detection and diagnosis. J Big Data 2023;10:144.
8. Muibideen M, Prasad K. A fast algorithm for heart disease prediction using Bayesian network model. arXiv 2020;2012.09429.
9. Ahmad U, Tawiah K, Albarrak AM, *et al*. Feature selection strategies for optimized heart disease diagnosis using ML and DL models. arXiv 2025;XXXX.XXXXX.
10. Garcia-Ordaz MT, Villanueva OI, Saeed F, *et al*. Heart disease risk prediction using deep learning techniques with feature augmentation. arXiv 2024;2402.05495.
11. Arunachalam SK, Rekha R. A novel approach for cardiovascular disease prediction using machine learning algorithms. Concurr Comput 2022;34(19):e7027.
12. Ogunpola A, Saeed F, Basurra S, *et al*. Machine learning-based predictive models for detection of cardiovascular diseases. Diagnostics 2024;14(2):144.
13. Paulino-Moreno C, Iparraguirre-Villanueva O. Classification and prediction of heart disease using machine learning models: a promising approach for medical diagnosis. Int J Educ Pract Eng 2024;1(1):1–8.
14. Khan A, Qureshi M, Daniyal M, *et al*. A novel study on machine learning algorithm-based cardiovascular disease prediction. Health Soc Care Community 2023;2023:1406060.
15. Theerthagiri P, J V. Cardiovascular disease prediction using recursive feature elimination and gradient boosting classification techniques. arXiv 2021;2106.08889.