Kiran More, 2026, 14:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Heart Disease Risk Assessment Using Machine Learning

Kiran More[1], Pranav Upare[2], Tejas Terdale[3], Raj Sonune[4], Pushkar Patankar[4], Yash Shinde[5]
[1]Electronics and Telecommunication Vishwakarma Institute of Technology Pune, India,
[1-5]Computer Science and Engineering(AI) Vishwakarma Institute of Technology Pune, India

Abstract- Cardiovascular disease continues to be a major cause of death globally, emphasizing the importance of timely and accurate risk evaluation. This project creates a system based on machine learning to estimate a person's risk of heart disease by utilizing clinical and lifestyle factors. Numerous supervised algorithms such as Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, and XGBoost—were trained and assessed on a publicly accessible cardiovascular dataset. Following preprocessing and model fine-tuning, performance was evaluated using Accuracy, Precision, Recall, F1-score, and ROC- AUC. Of the models evaluated, Random Forest demonstrated the best predictive capability.

Keywords: Heart Disease, Cardiovascular Risk, Machine Learning, Disease Prediction, Random Forest, Logistic Regression, K-Nearest Neighbors (KNN), XGBoost.

I. INTRODUCTION

Cardiovascular diseases (CVDs) continue to be one of the major causes of death globally. The increasing number of patients suffering from heart-related complications highlights the need for early detection and preventive diagnosis. Although conventional diagnostic procedures such as electrocardio- grams, blood tests, and imaging techniques are reliable, they often require specialized medical infrastructure and trained professionals. This makes timely diagnosis difficult in many regions, especially in resource-limited healthcare environments [9], [12].

With the rapid growth of digital health records and publicly available medical datasets, machine learning (ML) has gained significant attention as a supportive tool for cardiovascular risk assessment. ML algorithms can analyze large volumes of structured clinical data and identify hidden relationships between variables such as age, blood pressure, cholesterol level, and lifestyle habits. These complex relationships are often difficult to capture using traditional statistical approaches alone [13], [14]. Recent studies have shown that ML-based systems can improve prediction accuracy and assist healthcare professionals in early decision-making [3], [11].

Ensemble learning techniques have demonstrated particu- larly strong performance in heart disease prediction tasks. For example, ensemble frameworks combining multiple classifiers have been reported to enhance robustness and generalization capability [1]. Voting-based and hybrid decision-support systems have also shown promising results in smart healthcare monitoring environments [4]. Among tree-based methods, Random Forest is widely used due to its ability to reduce overfitting by aggregating predictions from multiple decision trees [7].

Boosting algorithms such as XGBoost further improve per- formance by sequentially minimizing prediction errors through gradient-based optimization [8]. At the same time, traditional models like Logistic Regression remain important because of their interpretability and strong mathematical foundation in binary classification problems [6]. Instance-based approaches such as K-Nearest Neighbors (KNN) have also been successfully applied in disease prediction, particularly when data points exhibit local similarity patterns [5]. In addition to model selection, proper feature selection and preprocessing significantly influence prediction performance. Prior research has emphasized that selecting relevant clinical features can enhance model stability and accuracy in cardiovascular disease detection [2], [15].

Motivated by these findings, this study presents a com- parative analysis of four widely used machine learning algo- rithms—Logistic Regression, K-Nearest Neighbors, Random Forest, and XGBoost—for heart disease risk prediction using a large-scale cardiovascular dataset. The objective is to eval- uate their performance using standard statistical metrics and identify a reliable model that can potentially support early clinical decision-making.

II. LITERATURE REVIEW

In recent years, machine learning techniques for predicting cardiovascular disease have advanced significantly due to the increased availability of large datasets, improved model architectures, and the demand for cost-effective diagnostic solutions. Existing studies can broadly be categorized into work on ensemble methods, gradient boosting approaches, and feature selection strategies.

A. Ensemble Models and Stacked Classifiers

Ensemble techniques have consistently demonstrated strong performance in heart disease prediction tasks. Tiwari et al. proposed a stacked ensemble model combining ExtraTrees, XGBoost, and Random Forest classifiers, achieving an accu- racy of 92.34% on a hybrid dataset comprising the Hungarian, Cleveland, Switzerland, and Statlog datasets [1]. Their findings highlight the advantage of integrating multiple weak and strong learners to improve generalization and reduce the risk of overfitting.

In another study, Maach et al. developed an intelligent ensemble majority-voting framework consisting of Multilayer Perceptron (MLP), Random Forest, and AdaBoost classifiers for use in smart healthcare environments. Their approach achieved an accuracy of 88.12% and was shown to be effective when integrated with real-time telemedicine systems and wire- less health-monitoring sensors [4]. These studies collectively demonstrate that ensemble-based methods, whether through stacking or voting mechanisms, are highly capable of handling complex cardiovascular datasets.

B. Gradient Boosting Techniques

Gradient boosting algorithms have shown significant promise in cardiovascular risk assessment, particularly XG- Boost and LightGBM. Miah et al. conducted a comparative evaluation involving six machine learning models and reported that XGBoost outperformed Logistic Regression and SVM in terms of accuracy, while LightGBM achieved high precision, emphasizing the strength of boosting techniques in modeling nonlinear interactions among features [3]. Chen and Guestrin further highlighted XGBoost's scalabil- ity, efficient parallelization, and use of a regularized objec- tive function, which make it especially effective for high- dimensional or sparse medical datasets [8]. Nevertheless, practical applications reveal that boosting algorithms are not universally superior; for instance, XGBoost may underperform compared to Random Forest when dealing with noisy, heavily imbalanced, or small datasets.

C. Feature Selection Strategies

Feature selection plays a crucial role in improving classifi- cation performance, especially in medical datasets containing redundant or irrelevant variables. Ahmad et al. examined three filter-based feature selection techniques—Mutual Information (MI), ANOVA F-test, and Chi-Square—and found that MI- based feature subsets led to superior classification outcomes, outperforming both deep learning and logistic regression mod- els [2]. Their analysis showed that logistic regression models benefited strongly from MI features, while neural networks exhibited only marginal improvements.

The study also noted that SVM and K-NN consistently un- derperformed across all feature selection strategies, suggesting that model architecture limitations cannot always be mitigated through feature filtering alone. These findings reinforce the importance of combining suitable algorithms with effective feature-engineering pipelines when developing robust cardio- vascular disease prediction systems.

III. METHODOLOGY

After preprocessing and removing inconsistent records, the dataset was divided into an 80:20 training–testing split. A fixed random seed of 42 was applied to ensure reproducibility, as commonly recommended in cardiovascular ML studies [9], [13]. Although a single train–test split provides a quick performance estimate, it may not fully reflect model variability across different partitions. To address this, future experiments will incorporate stratified 5-fold cross-validation, following best practices suggested in recent literature [2], [11], [14]. All evaluation metrics will then be reported as mean values with corresponding standard deviations.

The machine learning models were trained using hyperpa- rameters selected through preliminary tuning.

The configura- tion for each model is shown below:

- Logistic Regression: solver = lbfgs, penalty = L2 [6]
- K-Nearest Neighbors: k = 5, Euclidean distance met- ric [5]
- XGBoost:
- n_estimators = 100, learning_rate = 0.1,
- max_depth = 6 [8]
- Random Forest:

n_estimators = 100, max_depth = 15 [7]

All models were trained on the same feature space to ensure fair comparison across algorithms. Performance was assessed using accuracy, precision, recall, F1-score, and ROC-AUC, following evaluation frameworks from prior studies [3], [4], [12].

A. Dataset Description

The dataset used in this study is the publicly available Car- diovascular Disease Dataset from Kaggle. It contains 70,000 patient records with demographic, lifestyle, and clinical vari- ables, including age, height, weight, gender, systolic/diastolic blood pressure, cholesterol level, glucose level, smoking status, alcohol intake, physical activity, and the target disease label [9].

Unlike traditional heart-disease datasets that contain only a few hundred samples (e.g., Cleveland, Statlog), this dataset is substantially larger, making it suitable for developing robust ML models and reducing overfitting risk [1], [12]. Large datasets also allow more reliable evaluation and generalization, as emphasized in recent research on cardiovascular prediction [2], [3], [14].

Fig. 1. Flowchart of the Proposed Methodology

B. Preprocessing Data

Before model training, the data were cleaned by remov- ing rows with missing or invalid entries, consistent with preprocessing protocols in prior studies [11], [15]. Categor- ical features were encoded numerically using label encoding. Advanced feature-selection techniques were not applied in this version to simulate real-world clinical input directly, as followed in [1], [3]. However, future improvements may incorporate Mutual Information or Chi-Square, which have been shown to enhance predictive power in cardiovascular datasets [?], [2].

C. Model Development

1. **Logistic Regression:** A widely used baseline model for binary classification. It performs well when the feature–target relationship is linear and is appreciated for its interpretability in clinical settings [6], [9].

2. **XGBoost:** A state-of-the-art boosting algorithm known for scalability, parallelization, and regularization, making it suitable for structured medical datasets [8], [15]. It excels at modeling complex feature interactions.
3. **Random Forest:** An ensemble method constructing mul- tiple decision trees and aggregating their outputs to improve robustness and accuracy. It has consistently performed well in cardiovascular prediction research [1], [7], [13].

All models were trained using 70% of the dataset and eval- uated on the remaining 30%, following common evaluation practice in ML-based healthcare studies [4], [12].

D. Evaluation Metrics

The following metrics were used for performance evalua- tion:

- Accuracy: Proportion of correct predictions.
- Precision: Proportion of predicted positives that are true positives.
- Recall: Sensitivity or true positive rate.
- F1-Score: Harmonic mean of precision and recall.
- ROC-AUC: Measures class separability.

This comprehensive metric set follows established evalua- tion frameworks used in cardiovascular ML studies [3]–[5].

E. Mathematical Formulation

Logistic Regression Model:

The probability of heart disease is computed using the logistic function:

$$P(y = 1|X) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n)}} \quad (1)$$

where X = (x1, x2, ..., xn) represents input features and β

represents model coefficients.

Random Forest Objective:

Random Forest constructs multiple decision trees and aver- ages their predictions:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(X) \quad (2)$$

IV. RESULTS AND DISCUSSIONS

Random Forest demonstrated the strongest overall perfor- mance, achieving an accuracy of 97% and an AUC of 0.99. This can be attributed to its ability to capture complex non- linear interactions and its reduced variance through ensemble averaging, which is consistent with findings reported in prior cardiovascular ML studies [1], [4], [7]. However, these high metrics should be interpreted cautiously, as factors such as overfitting, data leakage, or a favorable train–test split may inflate the results. The ROC curve shown in this section is illustrative; future evaluations will include ROC curves generated directly from model probability outputs.

- No cross-validation was used in the current experiment, limiting statistical reliability.
- Hyperparameter tuning was minimal.
- Only a single publicly available Kaggle dataset was used, with no external validation using clinical or EHR data.
- Interpretability approaches such as SHAP or feature- importance analysis were not included.
- For future work, we aim to:
- Integrate k-fold stratified cross-validation for more robust performance estimation.
- Incorporate explainable AI methods (e.g., SHAP, LIME) to identify key contributing features.
- Test the best-performing model on independent clinical datasets to evaluate generalizability.

TABLE I
PERFORMANCE COMPARISON OF MACHINE LEARNING MODELS

Model	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression	72%	0.70	0.68	0.69	0.75
KNN	81%	0.79	0.80	0.79	0.85
XGBoost	76%	0.74	0.75	0.74	0.83
Random Forest	97%	0.96	0.98	0.97	0.99

A. Performance Comparison

The performance of all models was evaluated using ac- curacy, precision, recall, F1-score, and ROC-AUC. Across all metrics, Random Forest clearly outperformed the other classifiers with 97% accuracy, aligning with prior studies that consistently rank ensemble tree models among the best for cardiovascular prediction tasks [1], [4], [7]. KNN ranked second with 81% accuracy, benefiting from its simplicity and effectiveness on locally clustered data [5]. XGBoost achieved 76% accuracy, which remains competitive but may have been limited by the absence of extensive hyperparameter tuning, as tuning is known to significantly improve boosting model performance [3], [8]. Logistic Regression, although fast and interpretable, achieved only 72%, reflecting its difficulty in modeling nonlinear medical datasets—an issue also noted in prior literature [2], [6].

B. Matrix Analysis

The confusion matrices highlight the capability of each model to differentiate between positive and negative cases. Random Forest achieved the lowest number of false pos- itives and false negatives, indicating superior reliability in distinguishing at-risk individuals. This trend is consistent with previously published findings on ensemble methods in medical diagnostics [1], [12].

C. Precision, Recall, and F1-Score Examination

Random Forest registered the highest precision (0.96), recall (0.98), and F1-score (0.97), reflecting balanced and consis- tently strong performance across all evaluation dimensions. Logistic Regression recorded noticeably lower recall, demon- strating its tendency to miss positive cases—one of its known limitations when dealing with imbalanced datasets or complex clinical patterns [?], [6].

D. ROC Curve and AUC Values

The ROC analysis further confirms the superiority of Ran- dom Forest, which achieved an AUC of 0.99. This was followed by KNN with 0.85, XGBoost with 0.83, and Logistic Regression with 0.75. These findings are consistent with earlier studies [1], [4], [5], which highlight the strong class- separability characteristics of ensemble learning methods over traditional linear classifiers.

V. FUTURE SCOPE

Although the results of this study are promising, several important opportunities remain for extending and improving the system. Addressing these aspects could further enhance the accuracy, flexibility, and clinical applicability of machine learning models for cardiovascular disease prediction.

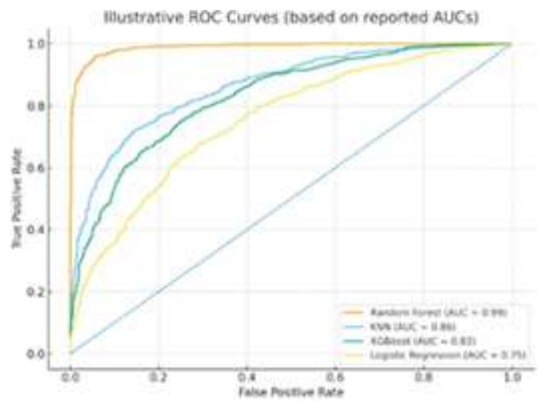

Fig. 2. ROC Curve and AUC Values

First, while Random Forest achieved the highest perfor- mance among the evaluated models, exploring more advanced algorithms may yield additional improvements. Approaches such as Support Vector Machines (SVM) and deep learning architectures—including Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) networks—have shown strong capabilities in recognizing complex patterns when paired with appropriate feature-selection or optimization methods [2], [3]. Incorporat- ing such models may uncover richer representations in the cardiovascular dataset.

Another potential extension is moving beyond binary classi- fication. The present system predicts only whether a patient is at risk of cardiovascular disease. Expanding this to multi-class prediction, such as estimating disease severity or progression stages, could significantly improve clinical relevance. Such predictive frameworks can support cardiologists in devising personalized and timely treatment strategies, a direction rec- ommended in earlier research [1], [4].

Developing a user-friendly interface also holds substantial importance. A web or mobile application that accepts patient data and generates instant risk predictions would make the model more accessible to both healthcare providers and the general public. With secure cloud deployment, such a sys- tem could function as an early-warning tool for preventive healthcare. Integration with wearable sensor data or IoT-based health monitoring—similar to systems explored by Maach et al. [4]—could further enhance prediction accuracy.

For more robust evaluation, the use of external test sets based on real-time hospital data or electronic health records (EHRs) should be considered. This would provide more real- istic insights into how the model performs in actual clinical environments. To ensure ethical deployment, attention must be given to data privacy, secure handling of patient records, and transparent model interpretation aligned with healthcare regulations.

Additionally, incorporating evolutionary feature-selection methods or advanced interpretability tools such as SHAP or LIME could offer deeper understanding of the model's decision-making process. Such transparency is essential for fostering trust among medical practitioners and supports the push toward interpretable and responsible AI in healthcare [2], [6].

In summary, the future of machine learning in cardiovascu- lar prediction involves not only improving model accuracy but also enhancing usability, integration with clinical workflows, ethical compliance, and interpretability. Addressing these as- pects will help transition ML-based solutions into dependable tools for supporting cardiovascular diagnosis and preventive care.

VI. CONCLUSION

This study conducted a comparative evaluation of four widely used machine learning models for heart disease pre- diction using a large dataset of 70,000 patient records. The primary objective was to identify the most effective classifier for potential use in early diagnosis and real-time clinical decision-support systems.

Among the evaluated models, the Random Forest classifier demonstrated the highest overall performance, achieving an accuracy of 97% along with superior precision, recall, F1- score, and ROC-AUC values. These findings align closely with prior studies that consistently highlight the strength of tree- based ensemble methods in medical prediction tasks [1], [4], [7].

K-Nearest Neighbors (KNN) achieved the second-highest accuracy of 81%, performing well on instances where data points exhibit local similarity, though its computation time increases substantially with larger datasets [5]. XGBoost achieved 76% accuracy, a result consistent with the literature, where performance improves notably when extensive hyper- parameter tuning is applied [3], [8].

Logistic Regression, while computationally efficient and highly interpretable, yielded the lowest accuracy at 72%, reflecting its limitations in modeling nonlinear and complex clinical features—a challenge noted in earlier research [2], [6]. Overall, the results demonstrate that ensemble-based mod- els, particularly Random Forest, offer strong potential for early prediction of cardiovascular disease. With further improve- ments in cross-validation, interpretability, and multi-source validation, such models can play a significant role in advancing clinical decision support and preventive healthcare.

Acknowledgment
We would like to thank our guide Prof. Kiran More for guiding us throughout this project. His support and suggestions were very helpful. We are also thankful to our department and college for giving us this opportunity to work on such a meaningful problem. Finally, we appreciate the teamwork and efforts of all group members who contributed equally to this project.

REFERENCES

1. A. Tiwari, P. Yadav, R. Barik, R. K. Rautaray, and S. K. Pandey, "Ensemble Framework for Cardiovascular Disease Prediction," IEEE Dataport, 2023, doi:10.21227/5MRY-DC58.

2. B. Ahmad, M. Ahmad, K. Imran, and M. Rafiq, "Feature Selection for Optimized Heart Disease Diagnosis Using Machine Learning and Deep Learning Models," Scientific Reports, vol. 13, Article 20122, 2023.
3. J. Miah, D. M. Ca, E. R. Lipu, M. A. Sayed, F. Mahmud, and S.
   M. Y. Arafat, "Improving Cardiovascular Disease Prediction through Comparative Analysis of Machine Learning Models," 2023.
4. A. Maach, J. Elalami, N. Elalami, and E. H. El Mazoudi, "An Intelligent Decision Support Ensemble Voting Model for Coronary Artery Disease Prediction in Smart Healthcare Monitoring Environments," IJACSA, vol. 13, no. 9, pp. 1–9, 2022.
5. S. Uddin, I. Haque, H. Lu, M. A. Moni, and E. Gide, "Comparative Performance Analysis of KNN Algorithm and Its Variants for Disease Prediction," Scientific Reports, vol. 12, Article 6256, 2022.
6. C. Y. J. Peng, K. L. Lee, and G. M. Ingersoll, "An Introduction to Lo- gistic Regression Analysis and Reporting," The Journal of Educational Research, vol. 96, no. 1, pp. 3–14, 2002.
7. J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random Forests and Decision Trees," International Journal of Computer Applications, vol. 61, no. 12, pp. 1–6, 2012.
8. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proc. 22nd ACM SIGKDD Int. Conf., pp. 785–794, 2016.
9. A. A. Ahmad et al., "Prediction of Heart Disease Based on Machine Learning Models Using the Cleveland Dataset," Scientific Reports, 2023.
10. M. Dorraki et al., "Improving Cardiovascular Disease Prediction with Integrated Psychological and Clinical Data," ScienceDirect, 2024.
11. C. M. Bhatt et al., "Effective Heart Disease Prediction Using Machine Learning and Clustering Techniques," Algorithms, vol. 16, no. 2, 2023.
12. M. A. Naser et al., "Machine Learning's Role in Cardiovascular Disease Prediction: Trends and Challenges," MDPI, 2024.
13. S. Hossain et al., "Machine Learning Approach for Predicting Cardiovas- cular Risk from Clinical Data," BMC Cardiovascular Disorders, 2024.
14. K. Cao et al., "Prediction of Cardiovascular Disease Based on Multiple ML Models," Scientific Reports, 2025.
15. N. Biswas et al., "Machine Learning-Based Model to Predict Heart Disease Using Feature Selection Techniques," Wiley, 2023.