International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 04, April 2026)

# Comparative Analysis of Machine Learning Models for Heart Disease Prediction

Akash Hiremath[1], Adarsha A M[2], Chandrashekar Y[3], Abhinav Shreyas K[4], Hema M S[5]

[1,2,3,4]*Department of Computer Science and Engineering (Students), RV Institute of Technology and Management, Bangalore – 560076, Karnataka, India*

[5]*Head of Department, Department of Computer Science and Engineering, RV Institute of Technology and Management, Bangalore – 560076, Karnataka, India*

***Abstract*—Heart disease has become one of the leading causes of mortality worldwide, making early prediction essential for effective healthcare intervention. This paper presents a comparative analysis of four machine learning models—Logistic Regression, Decision Tree, Random Forest, and XGBoost—using the UCI Heart Disease dataset. A structured pipeline involving data preprocessing, feature encoding, normalization, and stratified data splitting was implemented. The models were evaluated using multiple performance metrics including Accuracy, Precision, Recall, F1-score, and ROC-AUC to ensure a comprehensive assessment. Furthermore, stratified cross-validation techniques such as 5-fold and 10-fold validation were employed to improve robustness and consistency. To address overfitting concerns, a simplified "honest" Random**

**Forest model was introduced to evaluate real-world generalization. Experimental results demonstrate that ensemble methods outperform traditional models, offering higher accuracy and stability. The study provides a reliable, reproducible, and scalable framework for heart disease prediction in modern healthcare systems.**

***Keywords*— Heart Disease Prediction, Machine Learning, Random Forest, XGBoost, Classification, Cross-Validation, Healthcare Analytics**

## I. Introduction

Cardiovascular diseases (CVDs) are among the primary causes of death globally, posing significant challenges to healthcare systems. Early detection and preventive strategies are crucial in reducing mortality rates. Factors such as sedentary lifestyles, unhealthy dietary habits, stress, and lack of regular medical checkups have contributed to the increasing prevalence of heart disease.

Traditional diagnostic methods such as electrocardiograms (ECG), echocardiography, and clinical evaluations are effective but often time-consuming, expensive, and reliant on expert interpretation. With advancements in data-driven technologies, machine learning (ML) has emerged as a powerful tool for predicting diseases using historical clinical data.

Machine learning models can identify complex relationships between multiple risk factors and assist healthcare professionals in making faster and more accurate decisions. However, many existing studies focus on limited models or rely on a single evaluation metric, which does not provide a complete understanding of model performance. Additionally, issues such as overfitting and poor generalization are often overlooked.

This study addresses these gaps by comparing four widely used machine learning models—Logistic Regression, Decision Tree, Random Forest, and XGBoost—under a unified experimental framework. Multiple evaluation metrics and stratified cross-validation are used to ensure reliability, while a simplified baseline model is introduced to assess generalization and prevent overfitting.

## II. Related Work

Several research studies have explored the application of machine learning techniques for heart disease prediction. Traditional algorithms such as Logistic Regression and Decision Trees are widely used due to their simplicity and interpretability. However, these models often struggle to capture complex nonlinear relationships present in medical datasets.

Ensemble learning techniques such as Random Forest and XGBoost have demonstrated improved predictive performance. Random Forest, introduced by Breiman, reduces overfitting by combining multiple decision trees, while XGBoost enhances accuracy through gradient boosting and sequential learning.

Many studies using the UCI Heart Disease dataset report high accuracy levels, but they often rely on limited validation techniques or single performance metrics. This can result in overly optimistic outcomes that may not generalize well to real-world scenarios. In contrast, this study emphasizes a balanced evaluation using multiple metrics and robust cross-validation strategies, along with the introduction of an "honest" model for realistic performance assessment.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 04, April 2026)

## III. Methodology

### A. Dataset

The study utilizes the UCI Heart Disease dataset, which includes clinical attributes such as age, sex, chest pain type, resting blood pressure, cholesterol levels, fasting blood sugar, and maximum heart rate. The target variable indicates the presence or absence of heart disease.

### B. Data Preprocessing

Data preprocessing involved handling missing values using median and mode imputation techniques. Categorical variables were transformed using one-hot encoding, and feature scaling was performed using StandardScaler to normalize the data.

### C. Data Splitting

The dataset was divided into training and testing sets in an 80:20 ratio using stratified sampling to maintain class distribution and avoid imbalance.

### D. Model Configuration

- Logistic Regression: C = 0.5
- Decision Tree: Max depth = 5, Min samples split = 10
- Random Forest: 200 estimators, Max depth = 8
- XGBoost: 200 estimators, Max depth = 4, Learning rate = 0.05

### E. Evaluation Metrics

Performance was evaluated using Accuracy, Precision, Recall, F1-score, and ROC-AUC. F1-score was prioritized due to its importance in balancing false positives and false negatives in medical diagnosis.

### F. Validation Strategy

Stratified 5-fold and 10-fold cross-validation techniques were used to ensure model robustness and stability.

### G. Overfitting Detection

Training and testing accuracies were compared to detect overfitting. A simplified Random Forest model was introduced to evaluate generalization.

### H. Feature Importance

Tree-based models were used to identify the most influential clinical features affecting predictions.

## IV. Implementation

The proposed system was implemented using Python in a Jupyter Notebook environment.

Libraries such as NumPy, Pandas, Scikit-learn, XGBoost, Matplotlib, and Seaborn were used for data processing, model training, and visualization.

The dataset was loaded, explored, and preprocessed before training the models. Predictions were generated using the test dataset, and evaluation metrics were calculated. Confusion matrices and ROC curves were plotted to visualize model performance.

A structured evaluation pipeline ensured reproducibility and consistency. Additionally, an "honest" Random Forest model was implemented to validate real-world applicability and reduce overfitting.

## V. Results And Discussion

### A. Model Performance Comparison

| Model | Accuracy | Precision | Recall | F1 Score | ROC-AUC | CV Accuracy |
|---|---|---|---|---|---|---|
| Logistic Regression | 0.8683 | 0.8545 | 0.8952 | 0.8744 | 0.9439 | 0.8707 ± 0.0239 |
| Decision Tree | 0.8878 | 0.8942 | 0.8857 | 0.8900 | 0.9241 | 0.8976 ± 0.0151 |
| Random Forest | 0.9902 | 0.9813 | 1.0000 | 0.9906 | 0.9996 | 0.9768 ± 0.0124 |
| XGBoost | 0.9902 | 0.9813 | 1.0000 | 0.9906 | 1.0000 | 0.9805 ± 0.0146 |

*Best Model: Random Forest*

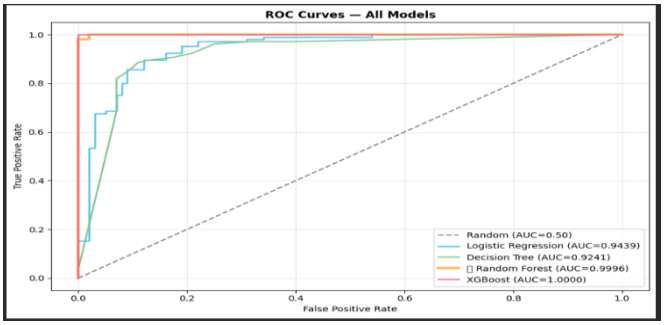
### B. Performance Analysis

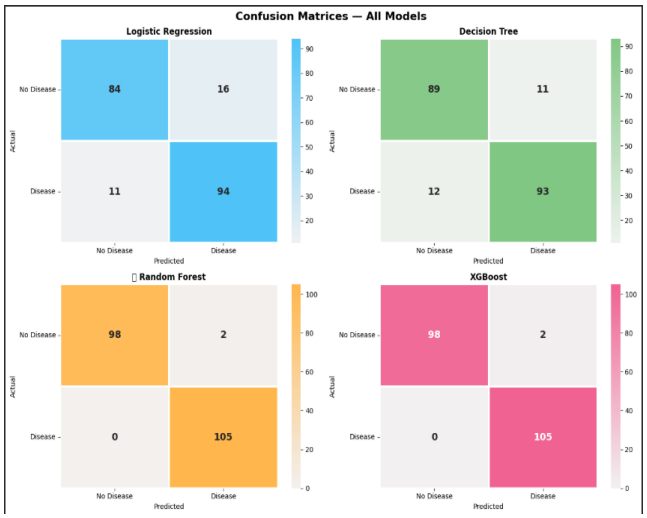
Ensemble models significantly outperformed traditional models across all evaluation metrics. Random Forest and XGBoost demonstrated superior accuracy, stability, and classification ability.

### C. Honest Model Evaluation

| Metric | Value |
|---|---|
| Accuracy | 0.8537 |
| Precision | 0.86 |
| Recall | 0.85 |
| F1 Score | 0.85 |
| CV Accuracy | 0.8653 ± 0.0187 |

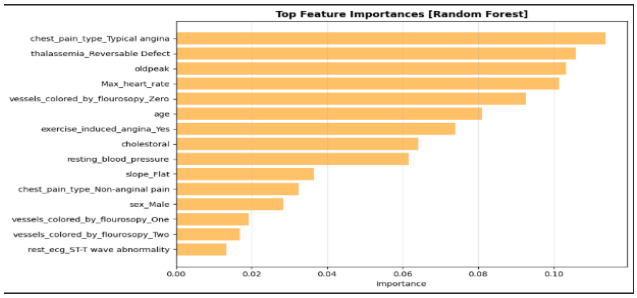
International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435 (Online) Volume 15, Issue 04, April 2026)




### *D. Feature Importance*

Key features influencing predictions include chest pain type, cholesterol levels, and maximum heart rate.




## VI. Final And Future Work

This study presented a comparative analysis of four machine learning models for heart disease prediction. Ensemble methods, particularly Random Forest and XGBoost, demonstrated superior performance and stability.

The inclusion of cross-validation and an honest model enhanced reliability and ensured realistic evaluation. Despite promising results, limitations include the use of a relatively small dataset and lack of real-time clinical data.

Future work may involve the use of larger datasets, deep learning techniques, and explainable AI models such as SHAP and LIME. Additionally, deploying the model as a web or mobile application can improve accessibility and practical use in healthcare environments.

## REFERENCES

[1] D. Dua and C. Graff, “UCI Machine Learning Repository,” 2019.
[2] R. Detrano et al., “Diagnosis of coronary artery disease,” 1989.
[3] J. H. Friedman, “Gradient boosting machine,” 2001.
[4] L. Breiman, “Random forests,” 2001.
[5] T. Chen and C. Guestrin, “XGBoost,” 2016.
[6] F. Pedregosa et al., “Scikit-learn,” 2011.
[7] S. Rajpurkar et al., “Arrhythmia detection,” 2019.
[8] K. Uddin et al., “Heart disease prediction,” 2023.
[9] M. Chicco and G. Jurman, “MCC vs F1,” 2020.
[10] A. Lundberg and S. Lee, “Explainable AI,” 2017.
[11] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016.
[12] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, 2009.
[13] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
[14] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2011.
[15] A. Esteva et al., “A guide to deep learning in healthcare,” *Nature Medicine*, vol. 25, pp. 24–29, 2019.
[16] P. Domingos, “A few useful things to know about machine learning,” *Communications of the ACM*, vol. 55, no. 10, pp. 78–87, 2012.
[17] S. Kotsiantis, “Supervised machine learning: A review of classification techniques,” *Informatica*, vol. 31, pp. 249–268, 2007.
[18] A. Rajkomar, J. Dean, and I. Kohane, “Machine learning in medicine,” *New England Journal of Medicine*, vol. 380, pp. 1347–1358, 2019.