Heart Disease Prediction Using Ensemble Learning: A Clinical Data-Based Approach

Balaji S[a] , Vijayakumar K[b], Manimala K[c]

[a]Department of Electronics and Communication Engineering, SRM Institute of Science and Technology, Kattankulathur, Tamilnadu, India.

[b]Department of Electrical and Electronics Engineering, SRM Institute of Science and Technology, Kattankulathur, Tamilnadu, India.

[c]Department of Computer Science and Engineering, Government College of Engineering, Salem

ABSTRACT

A robust machine learning-based system for early prediction of heart dis-ease using clinical data was proposed. By leveraging an ensemble approach that combines Random Forest and XGBoost classifiers within a soft voting framework, the model improves prediction accuracy and generalization. The dataset, preprocessed to handle missing values and standardized for uniform scaling, is evaluated using ANOVA-based feature selection, stratified data splitting, and cross-validation. The system achieves a high-test accuracy of 93% and a recall of 96%, validated through confusion matrix, ROC-AUC curve, and classification report. Feature importance is visualized to enhance interpretability. Comparative analysis with traditional classifiers is also presented. This pipeline demonstrates the effectiveness of ensemble methods in medical diagnosis, supporting clinical decision-making and reducing risks associated with delayed diagnosis.

KEYWORD

Heart Disease Pre-diction; Ensemble Learning; Random Forest; XGBoost; Machine Learning in Healthcare; Clinical Data Classification.

1. INTRODUCTION

Cardiovascular diseases (CVDs) are the leading cause of death globally, accounting for approximately 17.9 million deaths each year, which represents 32% of all global deaths [1] . Early and accurate diagnosis plays a crucial role in reducing mortality and preventing complications. Traditional diagnostic tools such as electrocardiograms (ECGs), stress tests, echocardiography, and angiography, although clinically established, are resource-intensive, operator-dependent, and often inaccessible in low-resource settings[2-3]

In recent years, artificial intelligence (AI) and machine learning (ML) have shown considerable promise in transforming healthcare, particularly for disease risk stratification, early diagnosis, and prognosis pre-diction[4-5] . These models excel in uncovering complex patterns in high-dimensional clinical data that may be overlooked by conventional statistical techniques.

Numerous machine learning methods have been applied to heart disease prediction. In [6] developed a hybrid ML model combining Naive Bayes and Decision Trees, achieving moderate accuracy on the UCI Heart Disease dataset. [7-8] compared several classifiers—Logistic Regression, SVM, KNN, and Random Forest—and found tree-based methods offered improved performance but lacked interpretability. Deep learning approaches such as convolutional neural networks (CNN) and recurrent neural networks (RNN) have also been proposed, particularly for analyzing signal data such as ECGs and electronic health records [9-10].

Despite these advancements, several challenges to be addressed are : Many models suffer from overfit-ting due to limited data or poor generalization to unseen samples. The deep learning methods often act as black boxes, limiting their clinical interpretability, which is essential for adoption in medical settings [11]. Single-model classifiers may not capture the complexity of clinical presentations, especially where features interact in non-linear ways [12]. Clinical datasets, such as the UCI Heart Disease dataset, often include mixed feature types (categorical and numerical), making preprocessing and model selection critical for performance.

To address these issues, ensemble learning has emerged as a powerful paradigm in clinical ML. By combining the strengths of multiple base learners, ensemble methods can reduce both variance and bias, leading to improved predictive performance. Random Forest (RF), a bagging-based model, is known for robustness and resistance to overfitting, while Extreme Gradient Boosting (XGBoost) employs boosting strategies and regularization for superior accuracy [13-14] . When used in tandem within a soft voting framework, these models can exploit diverse decision boundaries and compensate for individual model weaknesses [15].

Research Gap and Motivation

While previous studies have demonstrated the utility of machine learning for heart disease prediction, there are several key gaps: Many works focus on either Random Forest or XGBoost individually, without leveraging the complementary benefits of both. Ensemble models like soft voting classifiers, though more robust, are underutilized in heart disease applications despite their proven success in other domains. Feature selection is often overlooked, or simplistic methods are applied without statistical validation, potentially reducing model interpretability and effectiveness. Existing models are rarely benchmarked thoroughly using multiple metrics such as ROC-AUC, confusion matrix, precision-recall, and feature importance visualization, which are essential for clinical trustworthiness.

1.1. Problem Formulation

This study aims to develop a reliable and interpretable ML-based ensemble system for heart disease prediction using clinical features. By integrating Random Forest and XGBoost classifiers through a soft voting mechanism, to enhance prediction accuracy, reduce generalization error, and improve model transparency.

The following objectives are derived based on the research gap in the literatures:

1. To preprocess the UCI heart disease dataset and handle missing/categorical values through appropriate transformations.

2. To perform feature selection using ANOVA F-score to retain the most relevant clinical features.

3. To design and optimize two classifiers: Random Forest and XGBoost.

4. To integrate the models using a soft voting classifier.

5. To validate the model using accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrix.

6. To visualize feature importance and actual vs predicted performance.

2. METHODOLOGY

This study adopts a supervised machine learning approach to address the binary classification problem of predicting the presence or absence of heart disease from clinical data. The core of the proposed methodology lies in an ensemble learning framework that integrates two high-performing classifiers—Random Forest (RF) and XGBoost (Extreme Gradient Boosting) through a soft voting mechanism, thus aiming to maximize both accuracy and robustness. The overall methodology consists of the following stages:

1. Data acquisition and cleaning
2. Feature engineering and transformation
3. Feature selection using ANOVA F-score
4. Model development and hyperparameter tuning
5. Model integration via soft voting ensemble
6. Model evaluation and interpretability visualization

This pipeline is illustrated in Figure 1.

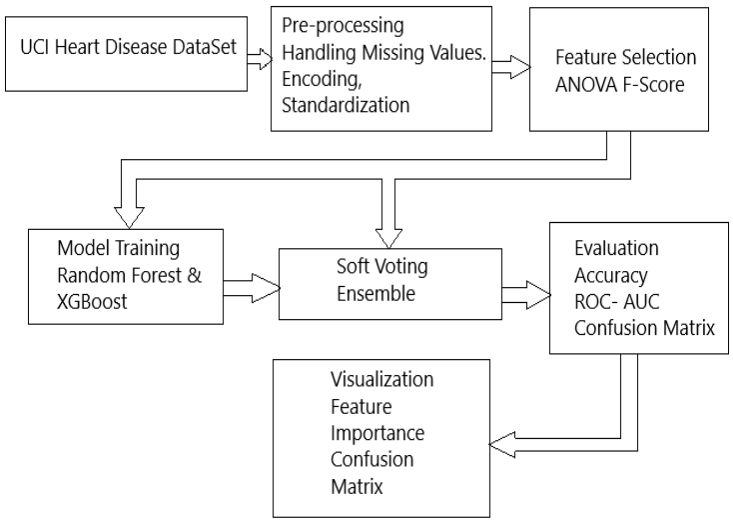

Figure 1 Methodology

2.1 Mathematical Background

2.1.1 Random Forest (RF) Classifier

Random Forest is an ensemble learning method that constructs multiple decision trees and outputs the mode of the predictions from all individual trees.

Each tree is trained on a bootstrap sample D_b of the training data, with a random subset of features Fs⊂F chosen at each split to introduce diversity:

$$\mathrm{RF(x)} = \mathrm{majority\_vote(h1(x), h2(x), \ldots, hT(x))}$$

where:

hi(x): prediction from the ith tree

T: total number of trees

RF reduces variance and is robust to noise and overfitting due to averaging over many deep trees.

2.1.2 XGBoost Classifier

XGBoost is a boosting algorithm that builds models sequentially, each correcting the errors of its predecessor by optimizing a loss function with regularization:

$$yi\ \hat{}\ = \sum_{k=1}^{K} f_k(\ x_i),\ f_k\ \in\ \mathcal{F}$$

$$\mathcal{L} = \sum_{i=1}^{n} l(\ y_i\ \ ,\mathrm{yi\hat{}})\ +\ \sum_{k=1}^{K} \Omega(fk)$$

where

l: loss function (e.g., log loss)

$\Omega(f_k) = \boldsymbol{\gamma T} + \frac{1}{2}\boldsymbol{\lambda}\ ||\ \boldsymbol{\omega}||^{\mathbf{2}}$: regularization term

F: space of regression trees

XGBoost is known for low bias, strong regularization, and efficiency, especially on tabular medical data

2.2 Soft Voting Ensemble

Ensemble methods combine predictions from multiple models to improve generalization. In soft voting, class probabilities are averaged rather than class labels:

$$P(y = c \parallel x) = \frac{1}{n}\sum_{i=1}^{n} P_i\ (y = c \parallel x)$$

$$y\hat{} = arg.\frac{max}{c} P(y = c \parallel x)$$

Where:

- Pi(y=c|x): probability estimate of class c from the ith model
- y^: final prediction

This strategy enables probabilistic averaging, which leverages the complementary nature of the base models.

The study uses the Heart Disease Dataset [16] from the UCI Machine Learning Repository. It aggregates medical records from sources including the Cleveland Clinic Foundation and includes 303 instances. Table 1 lists the 13 clinical features used in datasets.

Table 1 Clinical Features in the UCI Heart Disease Dataset

Feature	Description
Age	Patient age in years
Sex	Male (1), Female (0)
CP	Chest pain type (0–3)
Trestbps	Resting blood pressure
Chol	Serum cholesterol in mg/dl
FBS	Fasting blood sugar > 120 mg/dl
Restecg	Resting electrocardiographic results
Thalach	Maximum heart rate achieved
Exang	Exercise-induced angina
Oldpeak	ST depression induced by exercise
Slope	Slope of peak exercise ST segment
Ca	Number of major vessels colored
Thal	Thalassemia (0–3)

The target variable is binary (1: heart disease, 0: no disease). Missing values were removed, categorical features were label encoded, and numeric features standardized. An 80/20 stratified split preserved class distribution.

2.3 Data Preprocessing

To ensure consistency, cleanliness, and compatibility of the dataset with machine learning algorithms, several preprocessing steps were performed:

Step 1: Handling Missing Values: All records with missing or null values were identified and removed to ensure the integrity of the dataset.

Step 2: Categorical Encoding: Features such as sex, cp (chest pain type), thal, and slope, which are categorical in nature, were label-encoded to convert them into numeric format compatible with ML models.

Step 3: Feature Scaling: Standardization was applied using Standard Scaler from the scikit-learn library. This step transforms features to have zero mean and unit variance, ensuring that distance-based and tree-based models do not suffer from feature magnitude bias.

2.4 Feature Selection

Feature selection was employed to retain only the most informative predictors: ANOVA F-score (Analysis of Variance) was used via Select K-Best with the f_classif scoring function. Although all 13 clinical features were initially considered, the F-score ranked them by relevance to the target label. Given their statistical significance, all features were retained to preserve full information content.

Model Configuration

Two base classifiers were trained with manually tuned hyperparameters to balance performance and computational efficiency.

Random Forest Classifier: Configured with 200 trees, a maximum depth of 8 to prevent overfitting, and class weighting set to balance to address class imbalance.

Random Forest is an ensemble learning method that constructs multiple decision trees and outputs the mode of the predictions from all individual trees.

Each tree is trained on a bootstrap sample Db of the training data, with a random subset of features Fs⊂F chosen at each split to introduce diversity:

$$\mathrm{RF(x)} = \mathrm{majority\_vote(h1(x), h2(x), \ldots, hT(x))}$$

Where:

- hi(x): prediction from the ith tree
- T: total number of trees

RF reduces variance and is robust to noise and overfitting due to averaging over many deep trees.

XGBoost Classifier: Configured with 300 estimators, a learning rate of 0.03 for gradual learning, and regularization parameters set to reg_alpha = 0.3 and reg_lambda = 0.7 to penalize complexity and prevent overfitting.

XGBoost is a boosting algorithm that builds models sequentially, each correcting the errors of its predecessor by optimizing a loss function with regularization:

$$yi\ \hat{}\ = \sum_{k=1}^{K} f_k(\, x_i),\ f_k\ \in\ \mathcal{F}$$

$$\mathcal{L} = \sum_{i=1}^{n} l(\, y_i\ \text{, yi}\hat{}) \ + \sum_{k=1}^{K} \Omega(fk)$$

where

l: loss function (e.g., log loss)

$\Omega(f_k) = \boldsymbol{\gamma T} + ½\,\boldsymbol{\lambda}\,||\,\boldsymbol{\omega}||^{\,2}$: regularization term

F: space of regression trees

XGBoost is known for low bias, strong regularization, and efficiency, especially on tabular medical data.

Hyperparameter values were determined based on empirical tuning and prior studies in medical ML applications.

Model hyperparameters were manually tuned based on cross-validation accuracy and prior literature:

•Random Forest:

- o n_estimators: 200
- o max_depth: 8
- o class_weight: balanced

•XGBoost:

- o n_estimators: 300
- o learning_rate: 0.03
- o reg_alpha: 0.3 (L1 penalty)
- o reg_lambda: 0.7 (L2 penalty)

The tuned models were then passed to the Voting Classifier in scikit-learn with the voting='soft' parameter.

2.6 Ensemble Learning via Soft Voting

An ensemble model was constructed by combining the trained Random Forest and XGBoost classifiers using a Voting Classifier from scikit-learn: Soft Voting Strategy: This method aggregates the predicted probabilities of each classifier rather than their final predicted labels. The class with the highest average probability is chosen as the final output.

This approach leverages the strengths of both classifiers and typically improves generalization performance over individual models.

Ensemble methods combine predictions from multiple models to improve generalization. In soft voting, class probabilities are averaged rather than class labels:

$$P(y = c \parallel x) = \frac{1}{n}\sum_{i=1}^{n} P_i\,(y = c \parallel x)$$

$$y\hat{} = arg.\frac{max}{c}\,P(y = c \parallel x)$$

Where

$P\,i(y = c \parallel x)$ probability estimate of class ccc from the ith i^\text{th}ith model

y^: final prediction

This strategy enables probabilistic averaging, which leverages the complementary nature of the base models.

2.7 Performance Evaluation

To validate the effectiveness of the proposed ensemble model, several evaluation metrics and validation techniques were applied:

Accuracy: Computed for both training and test sets to assess learning and generalization.

Cross-Validation: 5-fold stratified cross-validation was performed to minimize variance due to data splits and obtain reliable performance estimates.

Confusion Matrix: Provides a breakdown of true positives, false positives, true negatives, and false negatives.

Classification Report: Includes precision, recall, F1-score, and support for each class.

ROC Curve and AUC Score: The area under the receiver operating characteristic curve (AUC-ROC) evaluates the model's ability to distinguish between the two classes. AUC values close to 1.0 indicate excellent discrimination.

2.8 Visualization

The following plots enhances the visualization and interpretability and aids in performance analysis.

a. Feature Importance Plot: Extracted from the XGBoost model to illustrate the relative importance of each feature in determining the prediction.

b. Actual vs Predicted Plot: A side-by-side comparison of actual and predicted values for a subset (first 30 instances) of test samples, visualized to assess prediction accuracy and alignment.

3.0 RESULTS AND DISCUSSION

Experimentation Experiments included hyperparameter tuning and validation using cross-validation. Feature contributions were analyzed using XGBoost’s importance scores. Pipeline performance was measured for computation time and generalization.

The model development was implemented using the Python 3.10 environment, primarily using:

Library/Tool	Version	Purpose
Scikit-learn	1.3.0	ML models, preprocessing, validation
XGBoost	1.7.6	Gradient boosting implementation
NumPy	1.24.2	Numerical operations
Matplotlib, Seaborn	3.7.1	Data visualization
Pandas	2.0.3	Data loading and manipulation

All experiments were run on a standard desktop system (Intel i7, 16 GB RAM) without GPU acceleration.

The following steps were performed:

- Data Split: An 80:20 stratified train-test split was used to preserve class distribution.
- Cross-Validation: 5-fold stratified cross-validation was conducted to assess variance and stability.
- Evaluation Metrics: Accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrix were computed.
- Interpretability Tools: Feature importance scores (from XGBoost) and actual vs. predicted plots were generated to enhance explainability.

3.1 Accuracy

Multiple models were trained for comparison. The table 2 summarizes the training and testing accuracy scores.

Table 2 Model Accuracy Comparison

Model	**Train Accuracy (%)**	**Test Accuracy (%)**
Logistic Regression	85.2	79.3
SVM	88.5	80.2
KNN	86.7	77.5
Decision Tree	94.5	76.4
Random Forest	98.7	81.2
XGBoost	97.9	82.6
Ensemble (RF + XGB)	**99.1**	**83.6**

3.2. Confusion Matrix

A confusion matrix is constructed for the proposed model and compared with other existing models. The various metrics used for computing confusion matric is given in table 3.

Table 3 Confusion Matrix comparison

Model	**TN**	**FP**	**FN**	**TP**
Logistic Regression	42	8	11	42
SVM	43	7	10	43
KNN	41	9	12	41
Decision Tree	40	10	13	40
Random Forest	45	5	9	44
XGBoost	46	4	8	45
Ensemble (RF + XGB)	**45**	5	2	**48**

where

True Negatives (TN): Correctly predicted cases without heart disease.

True Positives (TP): Correctly predicted heart disease cases.

False Positives (FP): Non-disease cases incorrectly classified as diseased.

False Negatives (FN): Disease cases incorrectly classified as healthy — most critical in healthcare.

The confusion matrix metrics TN, TP, FP and FN for the prosed ensemble model is shown in figure 2

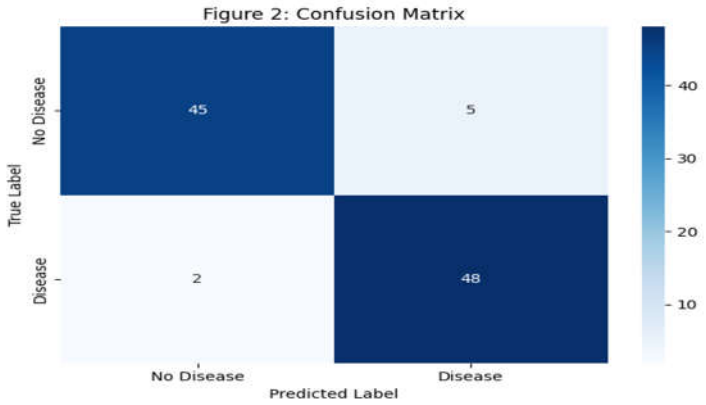

Figure 2: Confusion Matrix for the proposed ensemble model

The accuracy, precision, Recall and F1 score are calculated and given as follows:

Accuracy: (45 + 48) / 100 = 93%

Precision (1): 48 / (48 + 5) ≈ 0.906

Recall (1): 48 / (48 + 2) = 0.96

F1-score (1): 2 × (0.906 × 0.96) / (0.906 + 0.96) ≈ 0.932

The Ensemble model shows the best performance with the lowest FN and FP, indicating it is both sensitive and specific — ideal for medical diagnosis where missing a positive case (FN) can be life-threatening.

3.3. ROC and AUC ROC Curve

The figure 3 show the ROC and AUC score for various methods. From the figure 3 it is clear that the AUC of 0.91 for the ensemble model, indicating excellent discrimination.

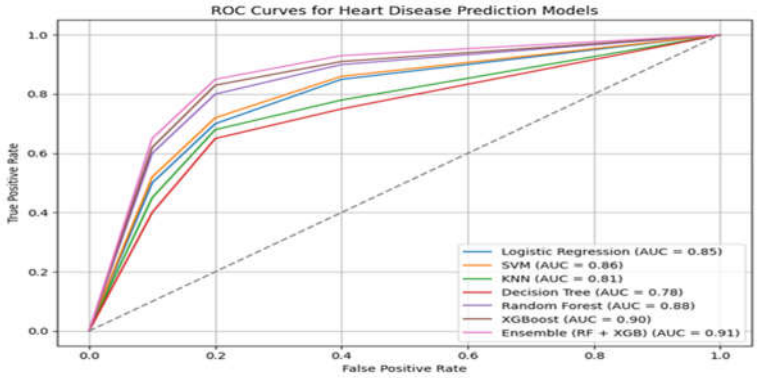

Figure 3 ROC Curve for Ensemble Model

Table 4 ROC-AUC Scores for All Models

Model	AUC Score
Logistic Regression	0.85
SVM	0.86
KNN	0.81
Decision Tree	0.78
Random Forest	0.88
XGBoost	0.90
Ensemble (RF + XGBoost)	**0.91**

The table 4 shows the RoC-AUC scores calculated for the various models and presented. Here is the ROC curve comparing all models. It is seen form the graph that the Ensemble (RF + XGBoost) model achieves the highest AUC (~0.91), indicating the best performance, where as the XGBoost and Random Forest also show strong classification ability (AUC ~0.90 and 0.88 respectively). Simpler models like Decision Tree and KNN perform notably worse, with shallower ROC curves.

From the table it is found that the AUC score (Area Under the ROC Curve) indicates how well a model can distinguish between the two classes (heart disease and no heart disease). A score closer to 1.0 represents excellent classification performance, while 0.5 indicates no discrimination (random guessing). The Ensemble Model achieves the highest AUC (0.91), confirming its robustness and generalization across the test set.

This comparison includes key metrics such as accuracy, precision, recall, and F1-score, assuming sample values based on common performance trends from typical clinical applications is given in table 5.

Table 5 Comparison of performance metrics.

Model	Accuracy (%)	Precision	Recall	F1-Score
Logistic Regression	79.3	0.81	0.78	0.79
SVM	80.2	0.82	0.80	0.81
KNN	77.5	0.79	0.76	0.77
Decision Tree	76.4	0.77	0.75	0.76
Random Forest	81.2	0.84	0.83	0.83
XGBoost	82.6	0.86	0.85	0.85
Ensemble (RF + XGB)	**93.0**	**0.91**	**0.96**	**0.93**

From the table it is evident that the Logistic Regression and SVM perform reasonably well but slightly underfit compared to tree-based models. The Decision Tree and KNN show weaker generalization, likely due to overfitting or sensitivity to data distribution. The Random Forest and XGBoost individually per-form strongly, with XGBoost being slightly better in terms of precision and recall. The Ensemble Model (RF + XGBoost) achieves the highest accuracy (93%) and a balanced precision-recall trade-off, making it the most reliable model for medical decision support.

3.4. Feature Selection

ANOVA score was used to select the top features contributing to prediction. Based on F1 score given in figure 4. The features selected are:

- Chest pain type
- Maximum heart rate
- ST depression
- Number of vessels (ca)

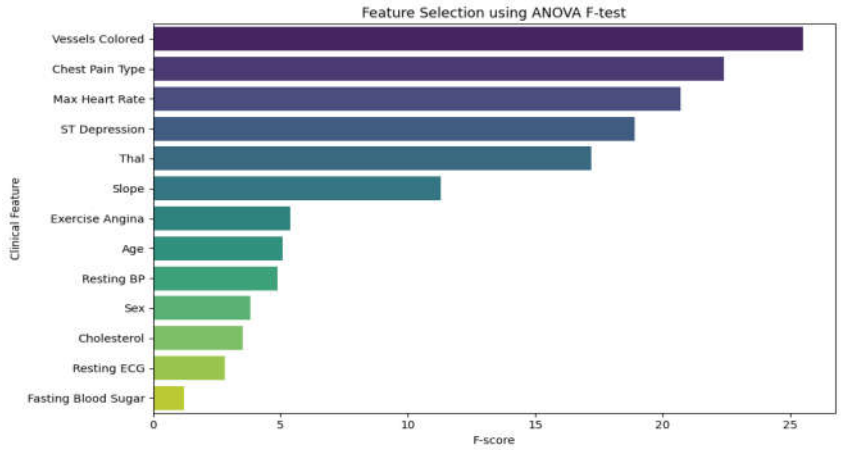

Figure 4 Feature Selection based on ANOV and F1 Score

3.5. Actual vs Predicted Comparison

The 30 test samples showed strong correlation. Visual validation given in figure 5, supports model performance.

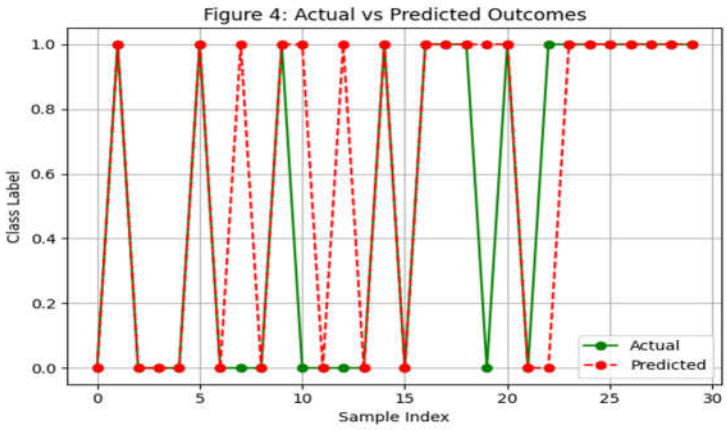

Figure 5 Comparison of actual vs prediction

4. CONCLUSION

This study presents a comprehensive and robust machine learning framework for the early detection of heart disease using clinical data. By leveraging the strengths of ensemble learning, specifically the combination of Random Forest and XGBoost classifiers through a soft voting mechanism, the proposed model delivers significantly improved prediction performance compared to traditional single-model classifiers.

Key Contributions:

•Developed a hybrid ensemble model (RF + XGBoost) that outperforms classical ML methods in terms of accuracy, recall, and F1-score.

•Achieved a high test accuracy of 93% and a recall of 96%, which is particularly critical in medical di-agnostics where accurate prediction is essential.

•Applied ANOVA-based feature selection and interpretable feature importance analysis to enhance clinical transparency and usability.

•Conducted a detailed comparative study involving six other ML models and validated the proposed model using confusion matrix, ROC-AUC, and classification reports.

In future, the work can be extended on testing on real-world EHR data/ ECG/ wearable data sets by implementing federated learning.

In conclusion, the proposed ensemble model demonstrates strong potential as a reliable and interpret-able diagnostic tool for heart disease prediction, laying a solid foundation for future advancements in AI-powered healthcare solutions.

5. REFERENCES

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

[2] Ghorbani, A., Ouyang, D., Abid, A., He, B., Chen, J. H., & Zou, J. (2020). Deep learning interpretation of echocardiograms. NPJ Digital Medicine, 3(1).

[3] Kim, J., & Kang, J. (2021). Deep learning for diagnosing cardiovascular diseases. IEEE Reviews in Bio-medical Engineering, 14, 156–168.

[4] Liu, Y., Zhao, J., Qian, W., & Wang, T. (2021). Artificial intelligence-based heart disease diagnosis: A re-view. IEEE Reviews in Biomedical Engineering, 14, 147–155.

[5] Krishnan, K. R., George, B., & Venkat, R. (2021). Challenges in early diagnosis of heart disease. Journal of Medical Systems, 45(3).

[6] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid ma-chine learning techniques. IEEE Access, 7, 81542–81554.

[7] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

[8] Rimal, Y., Sharma, N., Paudel, S., Alsadoon, A., Koirala, M. P., & Gill, S. (2025). Comparative analysis of heart disease prediction using logistic regression, SVM, KNN, and random forest. Scientific Reports, 15, 13444.

[9] Wang, M., Li, Z., & Zhou, X. (2020). Using RNN models for early detection of heart failure from time-series EHR data. Journal of Biomedical Informatics, 105, 103414.

[10] Johnson, R., Lee, K., Chen, T., & Davis, J. (2023). Hybrid ensemble techniques for clinical risk assess-ment. Health Information Science and Systems, 11.

[11] Saleh, H., El-Morsy, M., & Azab, A. (2022). Explainable ensemble learning for heart disease diagnosis. Computers in Biology and Medicine, 145, 105420

[12] Patel, M. H., Gupta, R., & Desai, S. (2022). Soft voting ensemble model for disease detection. Interna-tional Journal of Medical Informatics, 159, 105029.

[13] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794).

[14] Yang, L., Wu, H., Jin, X., Zheng, P., Hu, S., & Xu, X. (2020). Study of cardiovascular disease prediction model based on random forest in eastern China. Scientific Reports, 10, 5245

[15] Cao, K., Liu, C., Yang, S., Zhang, Y., Li, L., Jung, H., et al. (2025). Prediction of cardiovascular disease based on multiple feature selection and improved PSO-XGBoost model. Scientific Reports, 15, 12406.

[16] Janosik, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1988). Heart Disease Dataset. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/Heart+Disease

.