**RESEARCH PAPER**

# A Hybrid Feature Ranking and Ensemble Learning Framework for Early Cardiovascular Disease Prediction

**Deokar Anuradha S[1]* and Pradhan Madhavi A[2]**
*Computer Engineering Department, AISSMSCOE, Kennedy Road, SPPU University, Pune, India*
*[1]asdeokar@aissmscoe.com and [2]mapradhan@aissmscoe.com*

***Received: 24th April, 2026; Revised: 20th May 2026; Accepted: 30th June, 2026; Available Online: 10th July, 2026***

**ABSTRACT**
Cardiovascular disease must be identified early in order to reduce mortality and assist healthcare providers in quickly diagnosing and treating patients. However, achieving high predictive accuracy is challenging due to issues like data imbalance and presence of high –dimensional clinical features. To address these problems, this work provides a strong and understandable machine learning framework based on advanced ensemble boosting techniques. An efficient method for addressing dada imbalance and ensuring equitable representation of both healthy and affected individuals is to use random sampling techniques. A detailed feature analysis identifies the most crucial clinical parameters, and the model is made interpretable by using SHAP (Shapely Additive Explanations), which provides lucid insights into feature contributions. The model combines balanced sampling, feature refinement, and boosting techniques to provides a robust and clinically reliable diagnostic tool. Additionally, SHAP ensures transparency in model predictions by emphasizing the significance of critical clinical characteristics such as age, blood pressure, cholesterol, and smoking status. The finding confirms that combining ensemble learning with balanced datasets and optimal features selection significantly improves accuracy 98.20 % and interpretability.

**Keywords:** Cardiovascular disease; Ensemble learning; Machine learning, Feature selection; Feature selection; SHAP

**How to cite this article:** Deokar AS, Pradhan MA, A Hybrid Feature Ranking and Ensemble Learning Framework for Early Cardiovascular Disease Prediction Int J Drug Deliv Technol. 2026;16(6): 756-765. DOI: 10.25258/ijddt.16.6.114
**Source of support:** Nil.
**Conflict of interest:** None

## I. INTRODUCTION

Cardiovascular disease is becoming major cause of the death in all over the world. According to the WHO, cardiac conditions are responsible for about twelve million deaths worldwide each year. More than half of all deaths in the US and a number of other nations are caused by cardiovascular illnesses [1]. Cardiovascular illness, which includes heart and blood vessel disorders, is the world's biggest cause of death, with the exception of Africa.

Low- and middle-income nations have significantly higher mortality rates from cardiovascular illnesses, accounting for about 80% of all cardiovascular disease-related deaths. According to current estimates, nearly 23 million fatalities per year could be attributed to cardiovascular disorders by 2030 [2].

Historical evidence indicates that cardiovascular diseases existed in prehistoric times, with research on these conditions dating back at least to the 18th century. Efforts to understand the causes, prevention, and treatment of CD remain a major focus of biomedical research, with numerous scientific studies published each week [4].

Cardiovascular diseases are a common public health problem in many countries, posing a significant burden on patients and healthcare systems. Within the United Kingdom and the United States, cardiovascular illness is among the leading origins of hospitalization, and despite recent advances, outcomes remain poor.

In medical terms, the various methods mentioned in the topic can be effectively implemented. The suggested system can minimize many work the medical authorities do for forecast of heart and vascular diseases. The duration of the result is also reduced, and the correctness of the forecast can be improved with the help of increasingly advanced ML algorithms [4]. Together, Technology and Medical Science would have a considerable effect on the human race, improving overall inter-discipline between the Healthcare and IT domains. The detection of cardiovascular problems is made more effective and efficient by this study. A list of the most important contributions made by this paper is as follows:

Dataset preprocessing is part of this activity, and Feature Selection is used to increase computing time efficiency. Implemented a dual-stage feature selection process to reduce dimensionality while preserving relevant features, thereby enhancing model interpretability and reducing computational complexity. Integrated predictions from multiple base classifiers using ensemble techniques to improve model generalization, reduce variance, and achieve superior prediction stability across different datasets.

To compare the suggested method with well-established ML algorithms.

The ability to anticipate and identify cardiac disease in its early stages is a task that is both essential and difficult for medical professionals to do. The prevalence of heart disease continues to be a big issue, and it is frequently ascribed to factors such as the intake of alcohol and tobacco, as well as a lack of physical activity. The utilization of these tools makes it easier to analyses

**Author for Correspondence: asdeokar@aissmscoe.com*

complex medical data and to recognize patterns that may not be immediately obvious to human practitioners. Both of these processes are made possible by the implementation of these tools. Because of this, they make a contribution to the diagnosis and treatment of heart disease at an earlier stage

To establish a foundation for the proposed system, establish a foundation for the proposed system, this section reviews recent advancements in machine learning models data preprocessing techniques, feature selection methods, and ensemble classification strategies relevant to the problem domain.

[4], suggested a framework for predicting cardiovascular disease using ML techniques. A variety of classification algorithms, including Random Forest combined with ADA Boost, were employed, resulting in an accuracy of 86% on the Framingham dataset. For the purpose of enhancing the accuracy of forecasts, the study emphasizes the need of resolving the issues of missing data as well as class imbalance. The system used the Framingham coronary heart disease dataset to assess the effectiveness and generalizability of eight machine learning classification techniques. Unique double discriminant scoring method was put forth, and it outperformed other algorithms in variety of training and testing scenarios. The results highlight how well it reduces biases in predictive modeling [2] .[3]placed a high priority on creating a reliable model to forecast cardiovascular diseases. In order to increase prediction accuracy, it also assessed classifiers and identified important contributing factors.

Using the framingham coronary heart disease dataset,the study investigated the generaliz`ability of ML classification techniques[4].It evaluated a number of algorithms predictive ablility and emphasized the importance of selecting appropriate probalility thresholds,offering useful details regarding their suitability for different popilations[6] offer valuable insights regarding the efficacy and validation of cardiovascular prediction classification algorithms,particularly with the framinham dataset.when develping trustworthy prediction models ,they stress the importance of managing data,selecting algorithms and considering generalizability[5].These days ,it is difficult to detect and determine the primary cause of CVD at an early stage[8].A novel Quine McClusky Binary Classifier mdel was proposed,using seven models.The top 10 features are identified and as ubset of the dataset is created using the chi square ANOVA method.Nine prime components are found using CA with an accuarcy of 98.36%

The methods for predicting cardiovascular disease are in this section .Fifure 1 shows the recommended six phase strategy.A crucial step in obtaining informatiion for analysis is data collection .

To ensure that the data is suitable for analysis,preprocessing steps such as cleaning , transforming and encoding are frequently included in the process.Data cleaning is crucial step in the data preprocessing process because it ensures that the datasets are reliable,accurate and consistent.The identification and elimination of duplicate records which can disrupt the analysis and produce erroneous results ,is one of the methods several essential steps[10].

While data reduction techniques like PCA or t-SNE assist in elimination redundant features or instances,data transformation transforms uncommon information into a format that can be analyzed .Feature scaling allows for better comparison by standaradizing a range of variables,such as blodd presure,cholesterol etc.For ML to produce accurate results,effrctive preprocessing is essential.The probablility of cardiovascular disease is predicted using a variety of machine learning algorithm[12].

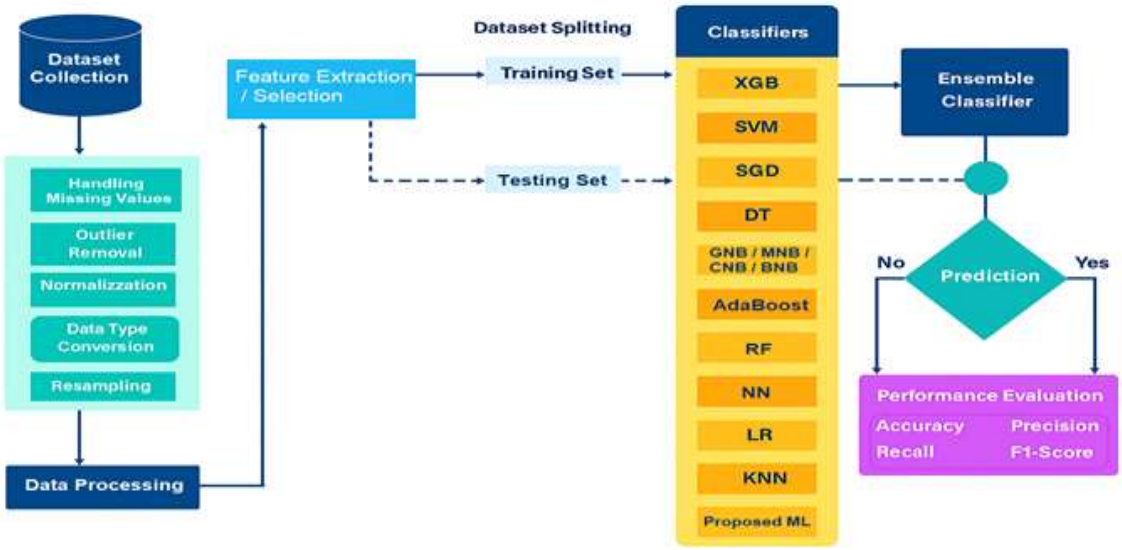


**Fig. 1.** Proposed architecture

Framingham Heart Study dataset [21], includes features 15,4240 instances and 1 target variable.,

**A. Correlation analysis and preprocessing of data**

Data preprocessing is crucial in machine learning (ML) to assess data quality and extract key features that affect model performance. Before a model is trained, the data must be processed to ensure that the results are reliable and consistent. Key elements like missing values, feature scaling, and standardization are handled during this phase.:

- It is advised to use preprocessing methods such as outlier mitigation and detection to improve the model's performance and achieve a reliable degree of disease prediction accuracy. Furthermore, uniformity across features is ensured by using a standardized data scaler, which improves model training. The way labels are organized within the dataset affects how well machine learning models work. By converting categorical target variables into numerical values label encoding enhances disease diagnosis by enabling machines to process them effectively.
- The following changes were made to the dataset in order to improve the identification of critical factors influencing cardiovascular disease:
- Body Mass Index Calculation: By combing a person's height and weight into a single metric, the BMI calculates the proportion of their body that is made up of fat.
- Gender and Age Transformation: Age, which was originally recorded in days, was changed to years for easier interpretation, and gender values were converted into binary format for better representation. Before a model is trained, the data must be processed to ensure that the results are reliable and consistent. Key elements like missing values, feature scaling, and standardization are handled during this phase:
- Feature Selection: To cut out extraneous variables and concentrate on pertinent features, some columns were eliminated.
- Outlier Detection and Removal: To improve data quality, rows containing outlier were removed after the datasets extreme values were found using quartile analysis.

• Encoding Categorical Variables: Creating numerical representation of categorical characteristics that are appropriate for machine learning algorithms.

By optimizing the dataset structure, these preprocessing techniques guarantee increased model accuracy and better predictions for the diagnosis of cardiovascular disease [11].

A. Analysis of Correlation

As seen in Figure 2, the Pearson Correlation Coefficient (PCC), which measures the relationship between various features in the dataset, is computed using:

$$r = \frac{n\sum xy-(\sum x)(\sum y)}{\sqrt{[n\sum x^2-(\sum x)^2][n\sum y^2-(\sum y)^2]}} \quad (1)$$

**Where r:** Pearson correlation coefficient

**n:** Number of data pairs.

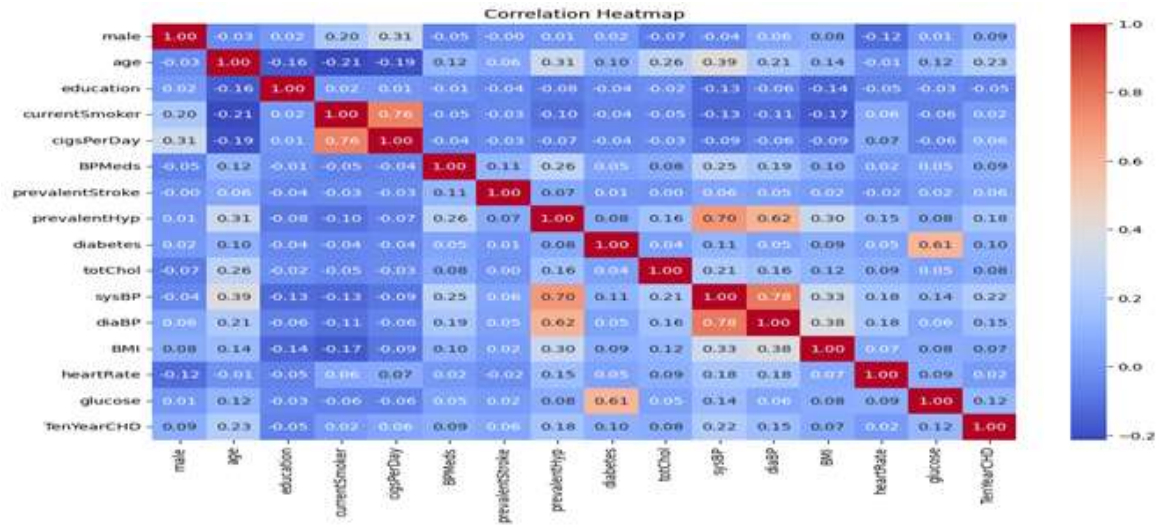


**Fig 2.** Correlation Matrix

x,y: Paired values from two datasets.

∑x,∑y: Sum of all x-values and y-values.

∑xy: Sum of the product of each pair xiyi.

$\sum x^2$, $\sum y^2$: Sum of squared values for x and y.

This helps eliminate irrelevant, redundant, or highly correlated features that do not significantly contribute to model performance

**The correlation coefficient ranges from -1 to 1:**

- A value near -1 shows a strong negative correlation between features.

- A value close to 1 signifies a strong positive correlation, meaning the features are closely related and can impact the model's accuracy.
- To refine feature selection, a correlation threshold of 0.85 was established.
- Features with a correlation value above this threshold were excluded to prevent redundancy.
- Features with a correlation value below the threshold were retained.

Feature Selection: Picking out the most important characteristics that contribute to classification is an absolute necessity [13]. The purpose of this is to reduce the number of dimensions by locating the function that provides the most information and can assist models in performing more effectively. We found that the approach to the selection of RF elements determines the main characteristics [14]. This is also a notable approach to feature selection, remove redundant or less significant features to avoid overfitting in ML techniques, and they work well in real-world scenarios.

**MODEL INTERPRETABILITY**

- SHAP (SHapley Additive exPlanations) values provide a systematic framework for interpreting the predictions of machine learning models by quantifying the contribution of each feature to an individual outcome[15].For cardiovascular disease (CVD) prediction, SHAP values shown in figure 3, facilitate an in-depth understanding of how specific risk factors—such as age, blood pressure, cholesterol levels, and smoking status—influence the decision-making process of the model for each patient. This level of interpretability not only enhances the transparency and reliability of the predictive system but also assists clinicians in recognizing the most influential determinants of diagnosis, thereby supporting evidence-based and informed medical decision-making.
- Train-Test split: Split the dataset into training and testing sets (e.g., 80%-20%).
- Cross-validation techniques are applied to ensure that the model generalizes effectively across unseen data by reducing both bias and variance during training.Additionally.entropy-based measures are incorporated to evaluate the uncertainty and information gain within the dataset, thereby supporting more reliable feature selection and enhancing the overall robustness of the learning process.

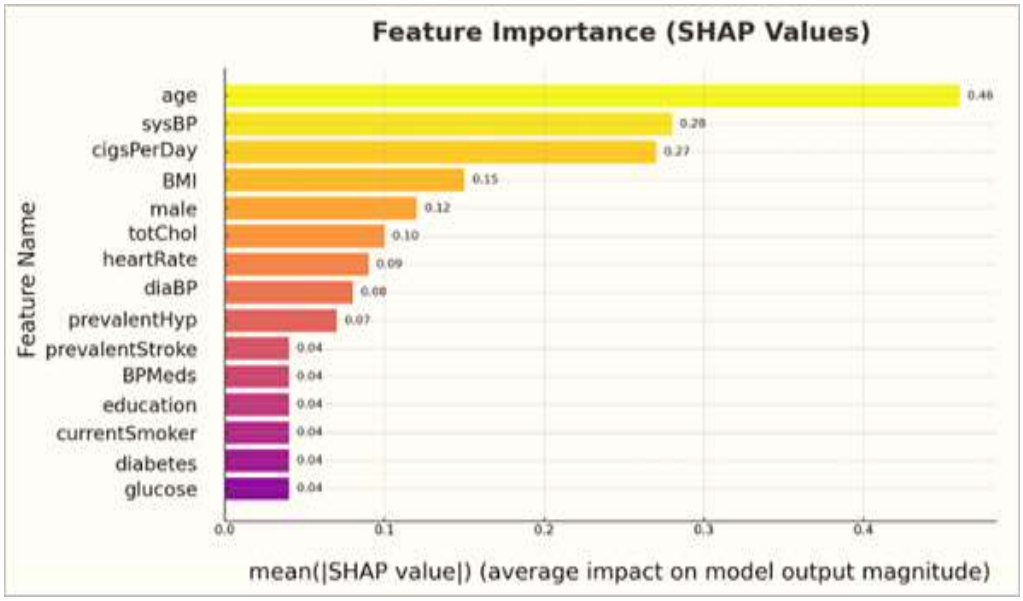


**Fig. 3.** SHAP values Graph

**C. Classification Models**

The framework presented integrates ML and DL models to enhance classification performance. It consists of an ensemble-based ML system combining XGBoost AdaBoost and Gradient Boosting along with a multilevel model incorporating both ML and DL techniques. This section demonstrates the effectiveness of using ensemble classifiers strategies in identifying cardiovascular disease while evaluating the success of the proposed approach.

**Algorithm**

**Algorithm 1:** Proposed Algorithm (Heart Disease Classification Algorithm)

**Input:** Heart Disease Dataset (both balanced and unbalanced)

**Output:** Final classification results with evaluation metrics.

1. **Start**

**2. Exploratory Data Analysis (EDA):**

a. Load the dataset.

b. Examine dataset structure(number of samples,features,data types).

c. Analyse the distribution of features and target classes.

d. Identify correlations among variables.

e. Compare feature distributions across categories (disease vs. no disease).

**3. Data Pre-processing:**

a. Handle missing values using suitable imputation techniques.

b. Encode categorical variable as needed.

c. Normalize or standardize numerical features if required.

d. Perform feature selection /importance analysis.

e. Prepare training and testing sets (apply balancing if necessary).

**4. Model Development and Training:**

a. Initialize candidate machine learning model .

b. For each model:

i Train the model using the training set.

ii. Validate on the testing set.

iii. Compute performance metrics.

iv. Adjust decision boundary iteratively.

- While (error>acceptable error) :

- Update model parameter.

- Recomputed error.

c. Record performance of all models.

**5. Model Selection and Final Evaluation:**

a. Identify the best -performing models based on validation metric.

b. Generate final predictions using the selected models.

c. Evaluate classification results with chosen metrics.

d. Report final performance outcomes.

**6. End**

The algorithm systematically refines the decision boundary by adjusting the classification threshold until the error rate meets the predefined acceptance criterion or the permissible threshold range is exhausted. This iterative process enables the model to achieve an optimal separation between high risk and low risk cardiovascular disease cases, thereby improving classification performance while adhering to the desired accuracy requirements.

**ML Ensemble Classifier**

Ensemble learning has been widely utilized to address various ML challenges by combining multiple model to enhance classification accuracy. Since datasets often contain diverse features and patterns, ensemble methods are employed to differentiate between diseases and classify data more effectively. Each classifier in the ensemble assigns a class label to new data points, and the final prediction is determined by the label that receives the majority vote.

To improve performance, ensemble classifiers integrate multiple ML models and utilize different voting mechanisms. These strategies outperform traditional single-model approaches by enhancing generalization and predictive accuracy. Through the utilization of a weighted majority voting mechanism, the suggested ensemble classifier is able to combine predictions from several classifiers, hence guaranteeing improved outcomes following the improvement of the model.

An ensemble learning technique called XGBoost (eXtreme Gradient Boosting) constructs several decision trees one after the other.

**Gradient Boosting**

This technique minimizes a loss function by using gradient descent. Cross-entropy loss generates a probability value between 0 and 1 to assess the effectiveness of a classification model:

$$Cross - Entropy = -\frac{1}{n}[y_i \log(y_i) + (1 - y_i)\log(1 - y_i)] \quad (2)$$

**Where:**

- $y_i$ is the actual label (0 or 1).
- $y_i$ the positive class's expected probability.
- n is the number of samples.

For implementation, we utilized the default settings provided by the Scikit-learn (sklearn) library, ensuring a standardized and efficient approach to classification.

**RESULTS AND DISCUSSION**

Performance Parameters: Details of Performance Parameters shown in table 2 [1].

**Dataset Details and Processing**

For each type of dataset, it checks dataset is unbalanced or balanced. In this case, Class 1-644 (Heart disease), Class 0-3596 (No Heart disease), figure 4 shows, it is imbalanced. For balancing dataset, SMOTE

technique is used. SMOTE generate synthetic samples Fig.4**.** Class distribution of target variable (Imbalance data)

to balance the dataset, which increases the quantity of training samples. Figure5, shows balanced dataset [20].

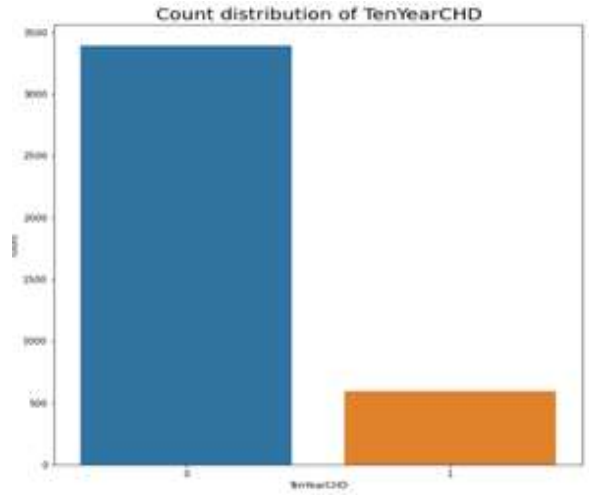


**Fig.4.** Class distribution of target variable (Imbalance data)

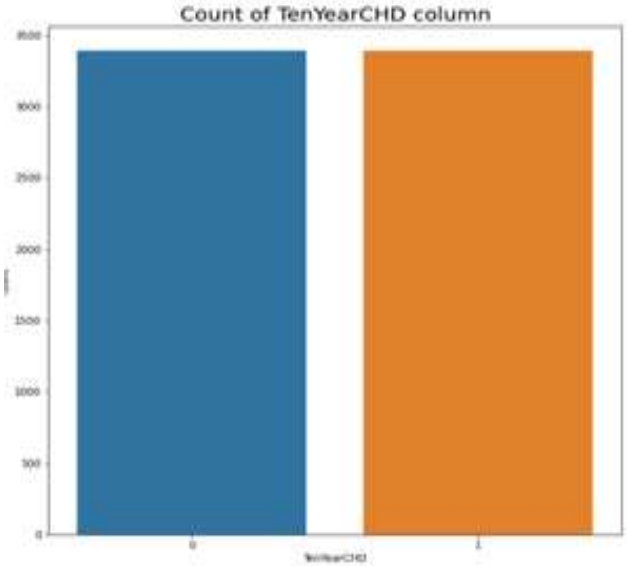


**fig.5.** Class distribution of target variable (balance data)

**Table Iii.** The Performance Measures Acquired From

| S No | Classifier | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|---|
| 1 | XGBoost Classifier | 97.15 | 97.2 | 97.15 | 97.15 |
| 2 | Support Vector Classifier | 67.45 | 67.73 | 67.45 | 67.36 |
| 3 | Stochastic Gradient Descent Classifier | 62.35 | 68.22 | 62.35 | 59.32 |
| 4 | Decision Tree | 96.75 | 96.95 | 96.75 | 96.75 |
| 5 | Gaussian Naive Bayes | 63.2 | 66.69 | 63.2 | 60.95 |
| 6 | Multinomial Naive Bayes | 54.65 | 54.59 | 54.65 | 54.17 |
| 7 | Complement Naive Bayes | 54.55 | 54.59 | 54.55 | 54.17 |
| 8 | Bernoulli Naive Bayes | 64.35 | 64.58 | 64.35 | 64.14 |
| 9 | K-nearest neighbors | 91.65 | 92.34 | 91.65 | 91.62 |
| 10 | Logistic Regression | 66.80 | 66.8 | 66.8 | 66.8 |
| 11 | Gradient Boosting classifier | 71.75 | 71.81 | 71.75 | 71.74 |
| 12 | Adaboosting ML | 70.85 | 70.9 | 70.85 | 70.84 |
| 13 | Neural Network | 67.45 | 67.73 | 67.45 | 67.36 |
| 14 | Proposed Decision boundary updating ensemble ML model | 98.10 | 98.14 | 98.10 | 98.10 |

## II. RESULT ANALYSIS

Table 3 and Figure 6 display the performance measures obtained from the various classifiers.

*a.* Measure Computational Efficiency

- Evaluate runtime usage of proposed algorithm during: Training of model
- Feature Selection to Improve Accuracy and Decrease Training Time

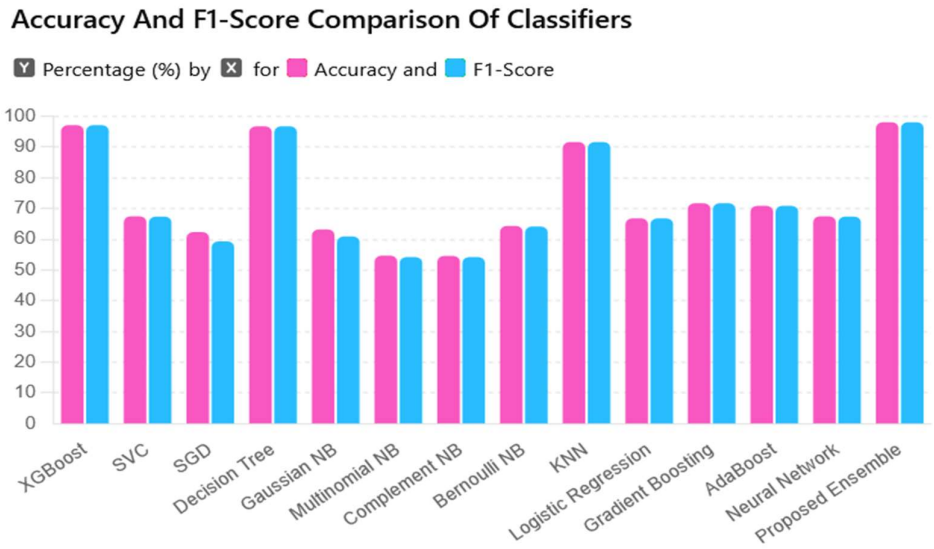


**Fig.6.** Performance Comparison of various classification models

**Table 4.** Training Time Prior to and Following Feature Selection for Imbalanced Dataset

| Imbalanced Dataset | Prior to selecting features | Following feature selection |
|---|---|---|
| Training Time | 4.5 seconds | 3.5 seconds |
| No of Features | 15 | 10 |
| Accuracy | 92.75 | 94.11 |
| Precision | 92.79 | 94.15 |

**Table 5.** Training Time Prior to and Following Feature Selection for Balanced Dataset

| Balanced Dataset | Prior to selecting features | Following feature selection |
|---|---|---|
| Training Time | 5.5 seconds | 4.2 seconds |
| No of Features | 15 | 10 |
| Accuracy | 96.40 | 98.10 |
| Precision | 96.44 | 98.14 |

In table [5], training time for balance dataset is higher than imbalance dataset. The dataset initially exhibited an imbalance before feature selection. The training time in table [4], was 4.5 seconds, with 15 features contributing to an accuracy of 92.75% and 94.11%. Following the selection of the most pertinent features, the training time was shortened to 3.5 seconds by reducing the number of features to 10. The model accuracy increased to 94.15% as a result of this optimization, proving that feature selection can improve performance while lowering computational complexity. For the balanced dataset, the training time before and after feature selection increased by 5.5 and 4.2 seconds, respectively, with improved accuracy of 96.40% and 98.10%, as shown in table [5].

You can verify the performance and computational efficiency of your designed algorithm by methodically assessing it using relevant metrics, datasets, and comparisons. A graph that shows the relationship between a model's true positive rate (TPR) and false positive rate (FPR) at different thresholds is known as a receiver operating characteristic curve, or ROC curve. The FPR is shown against the time period on the X-axis, while the TPR is shown on the Y-axis. When assessing a model's performance, one tool used is the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). In this assessment, the true positive rate and false positive rate are plotted against one other, and the area beneath the curve is then calculated. A closer score indicates that the algorithm can distinguish between the two outcome groups more successfully.

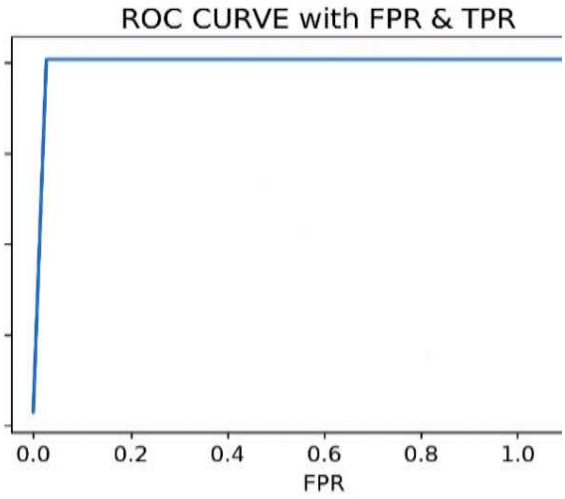


**Fig.7.** ROC curve with FPR and TPR

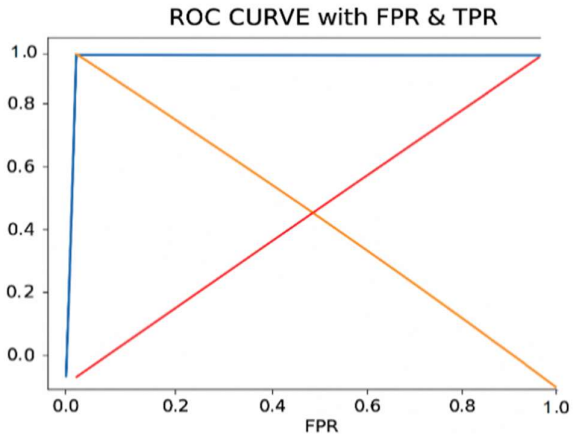


**Fig.8.** ROC curve with FPR and TPR with interception

ROC AUC scores can vary from 0 to 1, with 0.5 indicating random guessing (Figure 8) and 1 indicating flawless performance ((Figure 7). The range of these scores is from 0 to 1.

Figure 7, orange line indicates TPR and red line indicate FPR, Idle intersection is 0.58, in graph intersection of two lines is 0.58, means system is reliable.

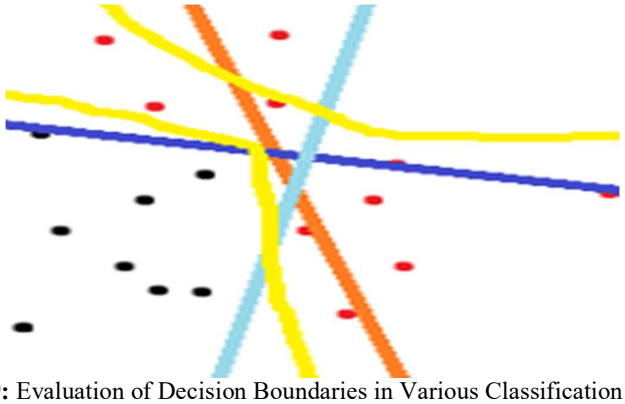

**Figure 9:** Evaluation of Decision Boundaries in Various Classification Systems

The decision boundary analysis (Figure 9) emphasizes how important model selection is to cardiovascular disease prediction performance. When it comes to addressing overlapping region where high risk and low risk patient profile share similar characteristics, Liner classifies comparatively simple separation boundaries are frequently insufficient. Such overlap lowers classification reliability by creating zones of diagnostic ambiguity. On the other, nonlinear classifier. Such as ensemble learning techniques and kernel based system show more flexibility in capturing intricate relationships between clinical features, leading to better prediction accuracy. These findings highlight how crucial it is to use sophisticated nonlinear methods. When creating cardiovascular disease prediction models, especially when dealing with heterogeneous datasets that have complex feature relationships.

## CONCLUSION AND FUTURE SCOPE

The primary goal of this study was to classify cardiovascular disease using range of models and empirical data collection. Key metrics such as accuracy98.20%, precision 98.14%, recall 98.1%, F1-score 98.1%, and computational efficiency have been used to assess the developed classification algorithms efficacy. Its ability to meet performance standards and generate reliable classification results is demonstrating by the analysis. Performance Accuracy: When compared to common benchmarks, the algorithms competitive accuracy levels validate its capacity to correctly identify class labels

- **Stability:** The algorithm exhibits resilience across a range of datasets, ensuring stable performance despite unbalanced or noisy data.
- **Efficiency:** Computational evaluations show that the algorithm performs well, making it appropriate for real time applications as well as large –scale implementations. Future models should incorporate patient –specific factors like age, ethnicity, and comorbidities to allow for more customized and clinically significance predictions.

## REFERENCES


[1] Belayadi, T. M. Rehman, S. Ali, M. Javaid, M. F. Almufareh, M. Humayun, and M. Shaheen, "Novel approach for the effective prediction of cardiovascular disease using applied artificial intelligence techniques," ESC Heart Failure, vol. 11, no. 6, pp. 3742–3756, 2024, doi: 10.1002/ehf2.14942.

[2] Y. Dubey, A. Bhongade, P. Palsodkar, and P. Fulzele, "Efficient explainable models for Alzheimer's disease classification with feature selection and data balancing approach using ensemble learning," Diagnostics, vol. 14, p. 2770, 2024, doi: 10.3390/diagnostics14242770.

[3] N. Kahouadji, "On the generalizability of machine learning classification algorithms and their application to the Framingham heart study," Information, vol. 15, no. 5, p. 252, 2024, doi: 10.3390/info15050252.

[4] N. Kahouadji, "Comparison of machine learning classification algorithms and application to the Framingham Heart Study," arXiv preprint, Feb. 2024, doi: 10.48550/arXiv.2402.15005.

[5] J. Roopini, B. G. Deepa, and N. G. Pooja, "Assessing and exploring machine learning techniques for cardiovascular disease prediction using Cleveland and Framingham datasets," ResearchGate, 2024, doi: 10.1109/NMITCON62075.2024.10698965.

[6] M. Qadri, A. Raza, K. Munir, and M. Almutairi, "Effective feature engineering technique for heart disease prediction with machine learning," IEEE Access, vol. 11, pp. 56214–56227, 2023, doi: 10.1109/ACCESS.2023.3281484.

[7] A. Jain, K. Kumar, R. G. Tiwari, N. B. Jain, V. Gautam, and N. K. Trivedi, "Machine learning-based detection of cardiovascular disease using classification and feature selection," in Proc. IEEE CSNT, Apr. 2023, doi: 10.1109/CSNT57126.2023.10134672.

[8] K. T. Ragunathan, S. Saleti, T. J. Lakshmi, and M. W. Ahmad, "Heart disease prediction using novel Quine McCluskey binary classifier (QMBC)," IEEE Access, vol. 11, pp. 64324–64336, 2023, doi: 10.1109/ACCESS.2023.3289584.

[9] A. Abdellatif, H. Abdellatef, J. Kanesan, C. Chow, J. H. Chuah, and H. M. Gheni, "An effective heart disease detection and severity level classification model using machine learning and hyper-parameter optimization methods," IEEE Access, 2022, doi: 10.1109/ACCESS.2022.3191669.

[10] P. Ghosh et al., "Efficient prediction of cardiovascular disease using ML algorithms with Relief and LASSO feature selection techniques," IEEE Access, Feb. 2021, doi: 10.1109/ACCESS.2021.3053759.

[11] A. Newaz and S. Muhtadi, "Performance improvement of heart disease prediction by identifying optimal feature sets using feature selection technique," in Proc. IEEE ICIT, Jul. 2021, doi: 10.1109/ICIT52682.2021.9491739.

[12] F. Tasnim and S. U. Habiba, "A comparative study on heart disease prediction using data mining techniques and feature selection," in Proc. IEEE ICREST, Jan. 2021, doi: 10.1109/ICREST51555.2021.9331158.

[13] A. Lakshmanarao, A. Srisaila, and T. S. R. Kiran, "Heart disease prediction using feature selection and ensemble learning techniques," in Proc. IEEE ICICV, pp. 994–998, 2021, doi: 10.1109/ICICV50876.2021.9388482.

[14] S. J. Pasha and E. S. Mohamed, "Novel feature reduction (NFR) model with machine learning and data mining algorithms for effective disease risk prediction," IEEE Access, Oct. 2020, doi: 10.1109/ACCESS.2020.3028714.

[15] R. G. Franklin and B. Muthukumar, "Survey of heart disease prediction and identification using machine learning approaches," in Proc. IEEE ICISS, Dec. 2020, doi: 10.1109/ICISS49785.2020.9316119.

[16] A. Nikam, S. Bhandari, A. Mhaske, and S. Mantri, "Cardiovascular disease prediction using machine learning models," in Proc. IEEE PuneCon, pp. 22–27, 2020, doi: 10.1109/PuneCon50868.2020.9362367.

[17] A. U. Haq, M. Li, M. H. Memon, J. Khan, and S. M. Marium, "Heart disease prediction system using model of machine learning and sequential backward selection algorithm for feature selection," in Proc. IEEE I2CT, Mar. 2019, doi: 10.1109/I2CT45611.2019.9033683.

[18] M. Deviaene, P. Borzée, B. Buyse, D. Testelmans, S. Van Huffel, and C. Varon, "Pulse oximetry markers for cardiovascular disease in sleep apnea," in Computing in Cardiology (CinC), pp. 1–4, 2019, doi: 10.22489/CinC.2019.205.

[19] N. V. Khandait and A. A. Shirolkar, "ECG signal processing using classifier to analyze cardiovascular disease," in Proc. IEEE ICCMC, pp. 855–859, 2019, doi: 10.1109/ICCMC.2019.8819777.

[20] K. G. Dinesh, K. Arumugaraj, K. D. Santhosh, and V. Mareeswari, "Prediction of cardiovascular disease using machine learning algorithms," in Proc. IEEE ICCTCT, pp. 1–7, 2018, doi: 10.1109/ICCTCT.2018.855.

[21] Framingham Heart Study Dataset [Online]. Kaggle, cited Oct. 6, 2025. Available: https://www.kaggle.com/datasets/aasheesh200/framingham-heart-studydataset