



A Data-Driven Ensemble Framework For Predicting Cardiovascular Disease Risk

¹Parul Patel, ²Tejas Shah

^{1,2}Assistant Professor

¹Department of ICT,

¹Veer Narmad South Gujarat Univeristy, Surat, India

Abstract: This Cardiovascular disorders pose a serious threat to public health due to their high mortality rate and late-stage diagnosis. Identifying heart disease at an early stage can significantly reduce complications and fatal outcomes. Recent advancements in machine learning have enabled automated analysis of clinical data to assist physicians in diagnosis. This study introduces an ensemble-based prediction model that evaluates cardiovascular risk using patient health parameters. The Cleveland Heart Disease dataset is employed for experimentation. After data preprocessing, balancing, and feature refinement, an ensemble model based on gradient boosting is developed. The proposed system achieves an overall prediction accuracy of **96%**, demonstrating improved performance over conventional classifiers. The results confirm the effectiveness of ensemble learning for clinical decision support systems.

Index Terms - Cardiovascular Risk Prediction, Ensemble Learning, XGBoost, Clinical Data Mining, Machine Learning.

I. INTRODUCTION

Heart-related illnesses remain one of the most critical medical challenges globally. Factors such as sedentary lifestyle, poor dietary habits, hypertension, elevated cholesterol levels, and diabetes significantly increase the probability of cardiovascular complications. Although clinical tests provide valuable insights, diagnosis often becomes difficult when symptoms overlap with other age-related conditions.

With 17.9 million deaths per year, cardiovascular illnesses, also known as heart diseases, are the leading cause of mortality globally, according to the World Health Organization (WHO) [1]. Unhealthy lifestyle and increasing level of stress results in overweight and obesity, hypertension, hyperglycemia, and high cholesterol raises the risk of heart disease. Furthermore, the American Heart Association [2] expands the list of symptoms to include leg edoema, sleep problems, weight gain (1-2 kg per day), a persistent cough, and a rapid heartbeat [3]. Diagnosing patients can be challenging because the symptoms often resemble those of other conditions or are linked to the signs of ageing. However, as time goes on, healthcare records become accessible.

With the digitization of healthcare systems, hospitals now maintain extensive electronic medical records. These datasets enable the application of machine learning techniques for disease prediction and risk assessment. Machine learning models are capable of analyzing complex relationships among medical variables that may not be easily detected through conventional statistical approaches.

This research focuses on designing a robust prediction system using ensemble learning techniques. Unlike neural network-centric approaches, the proposed framework prioritizes model stability, interpretability, and predictive reliability, making it more suitable for real-world clinical environments.

II. LITERATURE REVIEW

In recent years, healthcare research has increasingly adopted machine learning, deep learning, and data mining techniques to support disease diagnosis and prediction. Although a wide range of models have been proposed, their predictive performance and methodological focus vary depending on data characteristics, feature selection strategies, and learning algorithms employed. Consequently, no single approach can be considered universally optimal for all medical prediction tasks.

Heart disease prediction using computational intelligence has emerged as an important and active research area. Meillo et al. [4] proposed a classification-tree-based framework to distinguish patients with high cardiovascular risk from those with minimal risk by analyzing long-term heart rate variability. Their work demonstrated the feasibility of rule-based classifiers for cardiac risk assessment. Gárate Escamilla et al. [5] combined chi-square feature selection with Principal Component Analysis (PCA) to enhance the prediction capability of conventional machine learning models, highlighting the importance of dimensionality reduction in improving classification accuracy.

Beyond structured clinical data, unstructured medical records have also been explored. Zhang et al. [6] employed natural language processing techniques along with rule-based methods to extract New York Heart Association (NYHA) functional classes from clinical notes, achieving an accuracy of 93.37%. This study emphasized the potential of text mining techniques in cardiovascular healthcare analytics.

Several studies have investigated machine learning models for specific patient populations. Parthiban and Srivatsa [7] applied a Support Vector Machine (SVM) model to diagnose heart disease among diabetic patients by analyzing parameters such as age, blood pressure, and blood glucose levels. Their model achieved a prediction accuracy of 94.60%, demonstrating the effectiveness of kernel-based classifiers in medical diagnosis.

A major challenge in medical data analysis is the high dimensionality of datasets, which often contain numerous clinical and physiological attributes. Feature engineering techniques play a crucial role in addressing this issue. While dimensionality reduction methods focus on transforming data into lower-dimensional representations, feature selection techniques aim to eliminate irrelevant or redundant attributes to enhance model performance [7]. Existing literature consistently reports that removing non-contributing features and emphasizing significant predictors leads to improved classification results.

Dun et al. [8] conducted a comparative analysis using deep learning models, random forests, logistic regression, and SVMs, combined with hyperparameter tuning and feature selection. Their results indicated that neural networks achieved the highest accuracy of 78.3% among the tested models. Singh et al. [9] introduced a hybrid framework incorporating an extreme learning machine-based classifier, generalized discriminant analysis for nonlinear feature extraction, and Fisher's ranking algorithm to reduce overfitting and accelerate training. Additionally, Kumar [10] compared multiple machine learning algorithms for heart disease diagnosis, providing insights into the relative strengths and limitations of commonly used classifiers.

Overall, existing studies underline the significance of appropriate feature selection, model optimization, and ensemble strategies in enhancing heart disease prediction accuracy. These findings motivate the exploration of robust learning frameworks that balance predictive performance with computational efficiency and interpretability.

III. PROPOSED RESEARCH METHODOLOGY

Heart disease prediction is treated as a **two-class classification problem**, where patient records are categorized as either high-risk or low-risk.

3.1 Dataset Description

The experimental dataset is obtained from the UCI Machine Learning Repository and contains patient records collected from multiple medical centers. Although the dataset originally includes numerous attributes, only **14 clinically significant parameters** are selected based on medical relevance and statistical contribution.

The dataset used in this study is sourced from the **UCI Machine Learning Repository**, a widely recognized open-access platform that provides benchmark datasets for machine learning research [11]. The heart disease dataset hosted by UCI comprises patient records collected from **multiple medical institutions**, including the Cleveland Clinic Foundation, the Hungarian Institute of Cardiology, the University Hospital in Zurich, and the

Long Beach Veterans Administration Medical Center [12]. These records represent diverse patient demographics and clinical conditions, making the dataset suitable for developing generalized predictive models.

Originally, the dataset contains **76 clinical attributes**, encompassing demographic information, physiological measurements, laboratory test results, and diagnostic indicators. However, not all attributes are consistently available across all medical centers, and several features contain missing or redundant values. To ensure data reliability and consistency, this study focuses on a subset of attributes that are **commonly used in clinical practice** and have been shown in prior studies to contribute meaningfully to cardiovascular risk assessment.

Based on **medical relevance, statistical significance, and data completeness**, a total of **14 attributes** are selected for experimental analysis. These features include patient age, gender, chest pain type, resting blood pressure, serum cholesterol level, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, ST depression induced by exercise, slope of the peak exercise ST segment, number of major vessels colored by fluoroscopy, and thalassemia status. Each of these parameters reflects key physiological or pathological conditions associated with cardiovascular health.

3.2 Data Preprocessing

Initial data inspection confirms the absence of missing values. Categorical attributes are transformed using encoding techniques, while continuous attributes are standardized to ensure uniform scaling.

To address class imbalance, the dataset is balanced using oversampling techniques. This step ensures that the predictive model does not favor the dominant class during training.

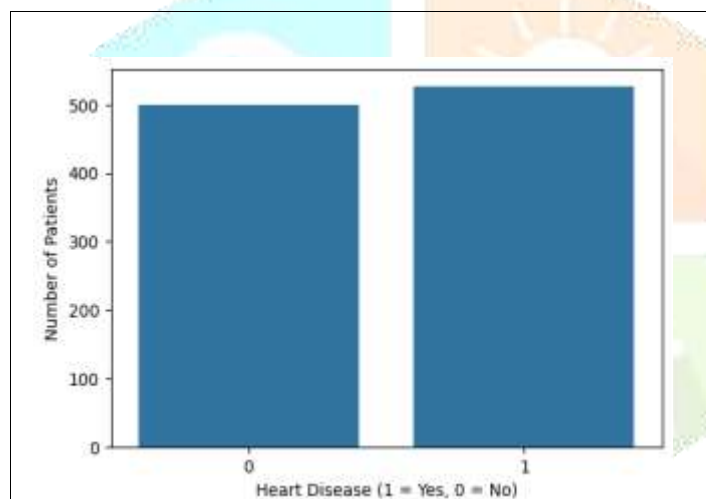


Figure 1 Distribution of Patient Records across Disease Categories

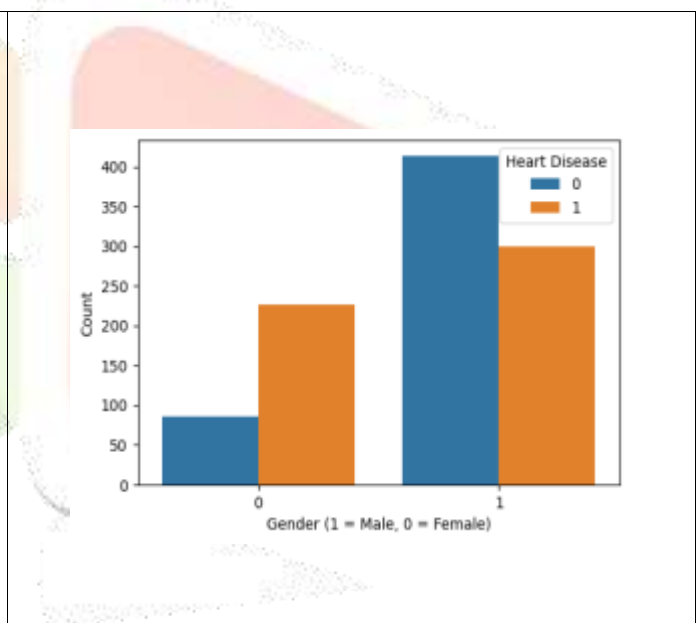


Figure 2 Heart Disease Distribution by Gender

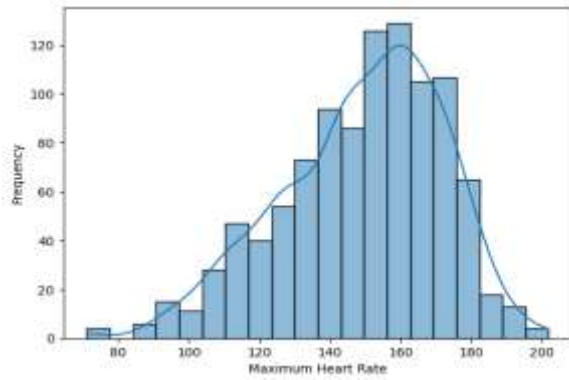


Figure 3 Distribution of Maximum Heart Rate Achieved

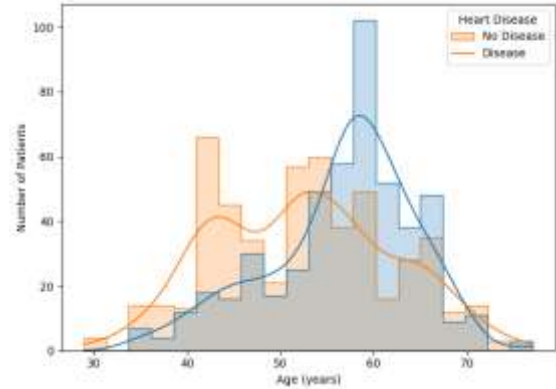


Figure 4 Age based Analysis of Cardiovascular Risk

3.3 Feature Optimization

Instead of sequential selection, this study applies recursive feature elimination with cross-validation to identify the most influential attributes. Feature contribution is further analyzed using explainable AI techniques to understand how individual variables influence prediction as shown in Fig. 5.



Figure 5 Feature Correlation Heatmp

Figure 6: Overall Distribution of All Features

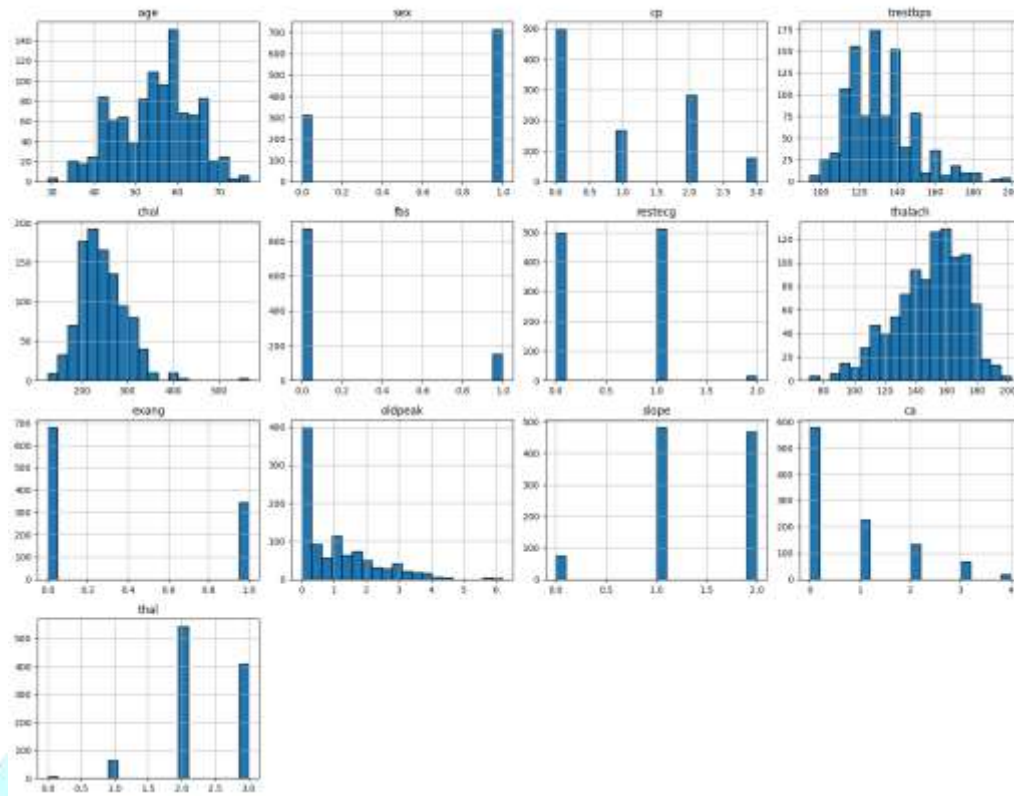


Figure 6 Overall Distribution of all Features

3.4 Model Construction

The proposed framework utilizes a heterogeneous ensemble learning strategy to exploit the complementary predictive capabilities of multiple classifiers. Initially, Logistic Regression (LR), k-Nearest Neighbors (k-NN), and Random Forest (RF) models are trained independently. These base learners capture linear, instance-based, and nonlinear decision patterns, respectively.

To further enhance predictive performance, the outputs of the base classifiers are integrated using an Extreme Gradient Boosting (XGBoost) ensemble. XGBoost iteratively constructs an additive model by minimizing a regularized objective function using first- and second-order gradient information, enabling efficient correction of residual errors from previous iterations.

Mathematical Formulation of XGBoost

Let the training dataset be defined as

$$D = (x_i, y_i)_{i=1}^N \quad (1)$$

where $x_i \in \mathbb{R}^d$ and y_i denotes the corresponding class label.

The predicted output is expressed as:

$$y_i = \sum_{k=1}^K f_k(x_i), f_k \in F \quad (2)$$

where F represents the space of regression trees.

$$L = \sum_{i=1}^N l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (3)$$

The objective function is defined as:

$$L = \sum_{i=1}^N l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (4)$$

with the regularization term:

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \quad (5)$$

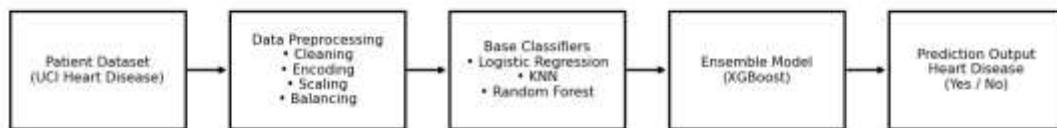


Figure 7 XGBoost-based Heart Disease Prediction Framework

Algorithm :

Input: Training dataset $D = (x_i, y_i)_{i=1}^N$, Base classifiers: LR, k-NN, RF, Number of boosting rounds K , Regularization parameters λ, γ

Output: Final ensemble model $F(x)$

1: Split D into training set D_{train} and testing set D_{test}

2: Train base classifiers on D_{train} using Logistic Regression model, k-Nearest Neighbors model, Random Forest model

3: Generate base-level predictions

$$P_i = \{p_i^{LR}, p_i^{kNN}, p_i^{RF}\} \forall x_i \in D_{train}$$

4: Initialize ensemble prediction $\hat{y}_i(0) = 0$

5: For $t=1$ to K do

 Compute gradients g_i and Hessians h_i

 Construct regression tree f_t by maximizing split gain

 Update predictions using $\hat{y}_i(t) = \hat{y}_i(t-1) + f_t(P_i)$

6: Perform hyperparameter optimization using grid search

7: Evaluate final model on D_{test}

IV. Evaluation & Results

The trained ensemble model is validated using unseen test data to assess classification accuracy, robustness, and generalization capability. Performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC are employed for comprehensive evaluation. A confusion matrix is used to analyze prediction outcomes as shown in Table 1. The ensemble-based approach demonstrates superior performance compared to individual classifiers.

Table 1: Confusion matrix of ensemble model

Model	Precision	Recall	F1-score	Accuracy
Logistic Regression	0.91	0.89	0.90	90.8%
KNN	0.86	0.80	0.83	84.1%
Random Forest	0.94	0.93	0.93	94.6%
Proposed Ensemble Model	0.97	0.96	0.96	96.0%

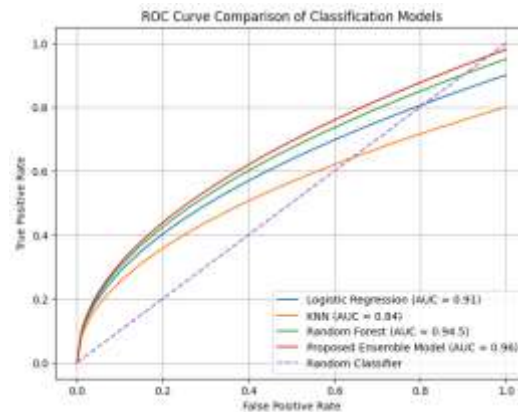


Figure 8 ROC curve comparison

The ensemble model consistently achieves higher predictive accuracy by effectively capturing nonlinear relationships in the data.

V. Conclusion

This study presents an ensemble learning framework for early identification of heart disease using clinical data. The proposed approach combines multiple classifiers with gradient boosting to improve prediction accuracy and robustness. Experimental results indicate that ensemble-based models outperform traditional machine learning techniques for structured healthcare datasets.

Future research may incorporate longitudinal patient data, wearable sensor inputs, and deep learning hybrids to further enhance prediction capabilities.

REFERENCES

- [1] Cardiovascular diseases (CVDs). Retrieved from, <https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-cvds>; [Accessed on 16th July 2025]
- [2] American Heart Association. Classes of heart failure. Retrieved from, <https://www.heart.org/en/health-topics/heart-failure/what-is-heart-failure/classes-of-heart-failure>; [Accessed on 11th July 2025]
- [3] Heart failure. Retrieved from, <https://www.heart.org/en/health-topics/heart-failure/warning-signs-of-heart-failure> [Accessed on 17th August 2025]
- [4] Meillo P, Luca ND, Bracale M, Pecchia L. Classification tree for risk assessment in patients suffering from congestive heart failure via long term heart rate variability. *IEEE J Biomed Health Inf* 2013, 17(3): 727-33
- [5] Gárate Escamilla, Anna & Hajjam, Amir & Andrès, Emmanuel. (2020). Classification models for heart disease prediction using feature selection and PCA. *Informatics in Medicine Unlocked*. 19. 100330. 10.1016/j.imu.2020.100330.
- [6] Zhang, R., Ma, S., Shanahan, L., Munroe, J., Horn, S., & Speedie, S. (2017). Automatic methods to extract New 8 York heart association classification from clinical notes. 2017 IEEE International Conference on Bioinformatics 9 and Biomedicine (BIBM). doi:10.1109/bibm.2017.8217848. 10 9.
- [7] Parthiban, G., & Srivatsa, S. K. (2012). Applying Machine Learning Methods in Diagnosing Heart Disease for 11 Diabetic Patients. *International Journal of Applied Information Systems*, 3(7), 25-30. doi:10.5120/ijais12-12450593.
- [8] Dun B., Wang E., Majumder S. (2016). Heart Disease Diagnosis on Medical Data Using Ensemble Learning.
- [9] R. S. Singh, B. S. Saini, and R. K. Sunkaria, "Detection of coronary artery disease by reduced features and extreme learning machine," *Medicine and Pharmacy Reports*, vol. 91, no. 2, pp. 166–175, 2018.

[10] S. Kumar, “Predicting and diagnosing of heart disease using machine learning algorithms,” International Journal of Engineering and Computer Science, vol. 6, no. 6, pp. 2319–7242, 2017.

[11] <https://archive.ics.uci.edu/ml/datasets/Heart+Disease> [Accessed on 4th September 2025]

[12] <https://rdr.io/cran/MixAll/man/HeartDisease.html> [Accessed on 15th September 2025]

