International Journal of Advanced Computing and Mechanical Systems (IJACM)
E-ISSN No.: xxxx-xxxx www.ijacm.com
Volume Number: 2 **Issue Number:** 1 **pp:** 10 - 23 **Publication Month:** Jan-2026

Classification of Heart Failure Using Machine Learning: A Comparative Study

N. Hema Sai Reddy[1,*], Muli Lokeshwari[1], Adavala Dhanush[1], S. Dilli[1]

[1]UG Student, Department of CSE, Siddartha Institute of Science and Technology, Puttur, Andhra Pradesh, India

***Autor1 E-Mail:** hemasaireddy242@gmail.com
Autor2 E-Mail: chinnarimlokeshwari@gmail.com
Autor3 E-Mail: adavaladhanush@gmail.com
Autor4 E-Mail: sagirajudilli2005@gmail.com

ABSTRACT

Heart disease continues to be one of the leading causes of mortality worldwide, making early and accurate diagnosis essential for effective treatment and prevention. In recent years, machine learning techniques have gained significant attention in healthcare applications due to their ability to analyze complex medical data and support clinical decision-making. This paper presents a comprehensive heart disease prediction framework based on a comparative analysis of seventeen machine learning models, including AdaBoost, Random Forest, Support Vector Machine, Extra Trees, Logistic Regression, Ridge Classifier, XGBoost, Passive Aggressive Classifier, Linear Discriminant Analysis, Gradient Boosting, Stacking, Bagging, Naïve Bayes, k-Nearest Neighbors, Stochastic Gradient Descent, Quadratic Discriminant Analysis, and Decision Tree.

Keywords: Heart Disease Predication, Machine Learning, Ensemble Models, Classification.

I. INTRODUCTION

Cardiovascular diseases remain among the foremost contributors to illness and death worldwide. Early detection and accurate diagnosis of these conditions pose a significant challenge within modern healthcare. These diseases commonly referred to as heart diseases comprise a wide spectrum of disorders that affect the heart and its network of blood vessels. They include conditions involving the coronary arteries, heart valves, cardiac muscles, and the heart's electrical conduction system, each of which can severely impair cardiovascular function. Heart failure (HF) is a condition in which the heart is unable to pump blood effectively, leading to inadequate delivery of oxygen and nutrients throughout the body. It commonly results from underlying issues such as coronary artery disease, long-standing hypertension, or structural damage to the heart muscle. Patients may experience symptoms like severe fatigue, breathlessness, and swelling in the limbs. Diagnosis is often complex because HF can manifest in diverse ways and mimic other medical conditions, making advanced imaging techniques and biomarker assessments essential for precise evaluation and appropriate management.

Distinguishing individuals with heart failure from those without the condition holds considerable clinical importance. Early identification of cardiovascular abnormalities allows healthcare providers to initiate timely interventions, which can significantly enhance patient outcomes. Prompt treatment not only minimizes the risk of serious complications such as myocardial infarction or progressive heart failure but also helps reduce the overall healthcare burden associated with these disorders. Moreover, correctly identifying individuals at risk of developing heart failure is vital for implementing effective prevention and management strategies. These measures may involve promoting healthy lifestyle habits such as balanced nutrition and regular physical activity as well as controlling key risk factors like hypertension and high cholesterol. Accurate classification ensures that medical attention and preventive resources are appropriately targeted to those who stand to benefit the most.

doi: https://doi.org/10.5281/zenodo.18266319

In this context, machine learning (ML) algorithms have become powerful tools for addressing the complexities of heart disease classification. This study utilizes the "Heart Failure" dataset available on the Kaggle platform, which contains roughly ten essential clinical features relevant to cardiovascular assessment. These features capture critical aspects of a patient's heart health and form the basis for applying supervised learning techniques aimed at accurately distinguishing individuals with heart failure from those without the condition. The novelty of this work lies in its use of machine learning algorithms to predict heart failure by integrating advanced preprocessing methods such as outlier removal and systematic hyperparameter optimization to enhance model performance. The modular structure of the framework allows it to be easily adapted to datasets involving other medical conditions, increasing its value as an educational and analytical resource. Furthermore, the use of a well-curated dataset, along with performance metrics like specificity, AUC, and Brier score, strengthens the study's relevance and potential utility in clinical decision-making.

Achieving strong performance metrics such as specificity, AUC, and Brier score further underscores the value of machine learning in early detection and informed clinical decision-making. Beyond simply predicting heart failure with reliable accuracy, the model can be incorporated into a Clinical Decision Support System (CDSS) to flag high-risk patients in real time, prioritize urgent cases, and optimize resource allocation, ultimately improving hospital workflow and patient outcomes.

II. LITERATURE REVIEW

This section presents a comprehensive review of the existing literature, emphasizing studies relevant to the current research focus. It examines a range of methodologies and approaches reported in previous works. The machine learning techniques applied across these studies are summarized and compared in Table 1. In this study, six data mining platforms—Orange, Weka, RapidMiner, KNIME, MATLAB, and Scikit-Learn were evaluated using six different machine learning algorithms: logistic regression, SVM, KNN, ANN, naïve Bayes, and random forest. The classification of heart disease was performed on a dataset comprising 13 input features, one target variable, and 303 instances, of which 139 corresponded to patients with cardiovascular disease and 164 to healthy individuals.

The heart disease dataset described in [7] consists of 70,000 patient records and includes 12 key attributes: age, height, weight, gender, systolic and diastolic blood pressure, cholesterol level, glucose level, smoking habits, alcohol intake, physical activity, and the cardiovascular disease indicator ("cardio," where 1 denotes the presence of disease and 0 indicates a healthy status). The dataset employed in this study incorporated 14 selected features. The referenced works applied various preprocessing techniques, utilizing the same 14 characteristics but organizing them into four distinct feature groups, each containing approximately six variables. These groups were then evaluated using different machine learning models to assess their impact on classification performance. In this study, a cardiovascular disease dataset sourced from Kaggle is utilized. It contains twelve distinct attributes along with a clearly defined target variable. These features cover demographic factors such as age, height, weight, and gender, clinical measurements including blood pressure, cholesterol, and glucose levels, as well as lifestyle indicators such as smoking, alcohol consumption, and physical activity. The target variable specifies whether cardiovascular disease is present or absent. The machine learning techniques applied across these studies include decision trees (DT), naïve Bayes (NB), random forest (RF), k-nearest neighbors (KNN), support vector machines with a linear kernel (SVM), and logistic regression (LR). In addition, gradient boosted trees (GBT) are incorporated to further enhance analytical performance.

doi: https://doi.org/10.5281/zenodo.18266319

TABLE 1. OVERVIEW OF MACHINE LEARNING MODELS USED FOR CLASSIFYING CARDIAC CONDITIONS

Description	Programming Languages / Tools	Model	Precision	Recall	F1 Score	Reference
Six data-mining tools and multiple ML techniques applied	Orange (3.34), Weka (3.8.3), KNIME (4.7), MATLAB (R2023a), Scikit-Learn (1.2.x)	SVM	83.84%	78.83%	–	[1]
		KNN	76.43%	77.37%	–	
		NB	84.51%	81.02%	–	
		RF	84.48%	81.75%	–	
		ANN*	85.86%	83.94%	–	
		LR	83.84%	81.75%	–	
12-feature dataset with GridSearchCV and hyperparameter tuning	Python (3.10)	XGB	88.93%	83.57%	86.16%	[2]
		MLP	88.70%	84.85%	86.71%	
		DT*	89.58%	81.61%	85.42%	
		RF*	89.42%	81.61%	86.32%	
Classification using UCI dataset (14 features)	Python (3.10)	DT	77.80%	75.50%	86.40%	[3]
		NB	85.40%	80.40%	86.50%	
		RF	88.70%	83.20%	87.90%	
		KNN*	91.70%	94.80%	90.80%	
		SVM	90.70%	87.80%	88.50%	
		LR	88.10%	86.10%	86.50%	
Classification with 14 features grouped into 4 categories	Python (3.8)	LR	–	75%	–	[4]
		KNN	–	64%	–	
		SVM	–	70%	–	
		SVM (Linear)	–	75%	–	
		NB	–	78%	–	
		DT*	–	83%	–	
Cardiovascular disease prediction using 12 features	–	NB	70%	34%	46%	[5]
		DT	75%	68%	71%	
		LR	73%	68%	70%	
		RF	71%	71%	71%	
		SVM*	76%	64%	69%	
		KNN	66%	64%	65%	
Cardiovascular disease prediction using 12 features	–	KNN	70%	70%	70%	[6]
		SVM*	73%	71%	71%	
		DT	63%	63%	63%	
		ANN*	73%	73%	73%	
		NB	71%	71%	71%	
		RF	70%	70%	70%	
		LR*	73%	73%	73%	

III. MATERIALS AND METHODS

The heart disease dataset used in this study obtained from the Kaggle platform (San Francisco, CA, USA) and titled “Heart Failure” provides a consolidated resource for examining cardiovascular conditions. It includes roughly ten key features relevant to the classification of heart disease. All data analysis and model development were performed in a collaborative environment using Google Colab (Mountain View, CA, USA), which offered the computational capacity needed for efficient processing and experimentation. Python 3.13 (Python Software Foundation, Wilmington, DE, USA) served as the primary programming language because of its flexibility and the extensive ecosystem of libraries supporting data analysis and machine learning tasks.

doi: https://doi.org/10.5281/zenodo.18266319

3.1 Data Analysis and Preprocessing

The dataset employed in this study integrates five separate heart disease datasets into a unified collection containing 918 records and 11 attributes. These attributes include age, sex, chest pain type, resting blood pressure, cholesterol level, fasting blood sugar, resting ECG results, maximum heart rate achieved, exercise-induced angina, ST depression, and ST slope. As the largest publicly available dataset on heart disease, it offers a robust foundation for analytical and predictive research. A detailed overview of these features is presented in Table 2.

TABLE 2. CHARACTERISTICS OF THE COHORT USED FOR CARDIAC CONDITION CLASSIFICATION

SN	Attribute	Categories (Frequency)	Type	Range	Description	Mean ± SD	Units
1	Age	–	Numerical	28–77	Age of the patient	53.51 ± 9.43	years
2	Sex	Male: 725 (79%), Female: 193 (21%)	Categorical	–	Biological sex	–	–
3	Chest Pain Type	ASY: 496 (54%), NAP: 203 (22%), ATA: 173 (19%), TA: 46 (5%)	Categorical	–	Type of chest pain	–	–
4	Resting BP	–	Numerical	0–200	Resting blood pressure	132.40 ± 18.51	mmHg
5	Cholesterol	–	Numerical	0–603	Serum cholesterol level	198.80 ± 109.38	mg/dL
6	Fasting BS	0: 704 (77%), 1: 214 (23%)	Binary	0–1	Fasting blood sugar (>120 mg/dL)	0.23 ± 0.42	–
7	Resting ECG	Normal: 552 (60%), LVH: 188 (21%), ST: 178 (19%)	Categorical	–	Resting ECG result	–	–
8	Max HR	–	Numerical	60–202	Maximum heart rate	136.81 ± 25.46	bpm
9	Exercise Angina	No: 547 (60%), Yes: 371 (40%)	Binary	–	Exercise-induced angina	–	–
10	Oldpeak	–	Numerical	−2.6–6.2	ST depression induced by exercise	0.89 ± 1.07	–
11	ST Slope	Flat: 460 (50%), Up: 395 (43%), Down: 63 (7%)	Categorical	–	Slope of the ST segment	–	–
12	Heart Disease	1: 508 (55%), 0: 410 (45%)	Binary	0–1	Output class	0.55 ± 0.50	–

The preprocessing phase began with a comprehensive examination of the dataset. Exploratory Data Analysis (EDA) methods—including histograms, box plots, and density plots (Fig. 1) were applied to understand variable distributions and identify potential irregularities. During the data-cleaning process, 16 outliers were detected and removed using the Interquartile Range (IQR) technique, selected for its robustness and ease of implementation in identifying extreme values [7]. Clinical assessment confirmed that these outliers represented erroneous entries or physiologically improbable conditions. Their removal ensured that the dataset reflected realistic clinical patterns, thereby enhancing the reliability and predictive performance of the subsequent machine learning models.

Following data cleaning, categorical variables were examined to detect categories with low occurrence. A 5% frequency threshold was established to flag infrequent categories; however, none of the categories in the evaluated columns fell below this cutoff. Consequently, all original categories were preserved. For preprocessing, ordinal encoding was applied to variables with an inherent order, while

doi: https://doi.org/10.5281/zenodo.18266319

one-hot encoding was used to convert nominal categorical variables into numerical formats appropriate for machine learning algorithms.

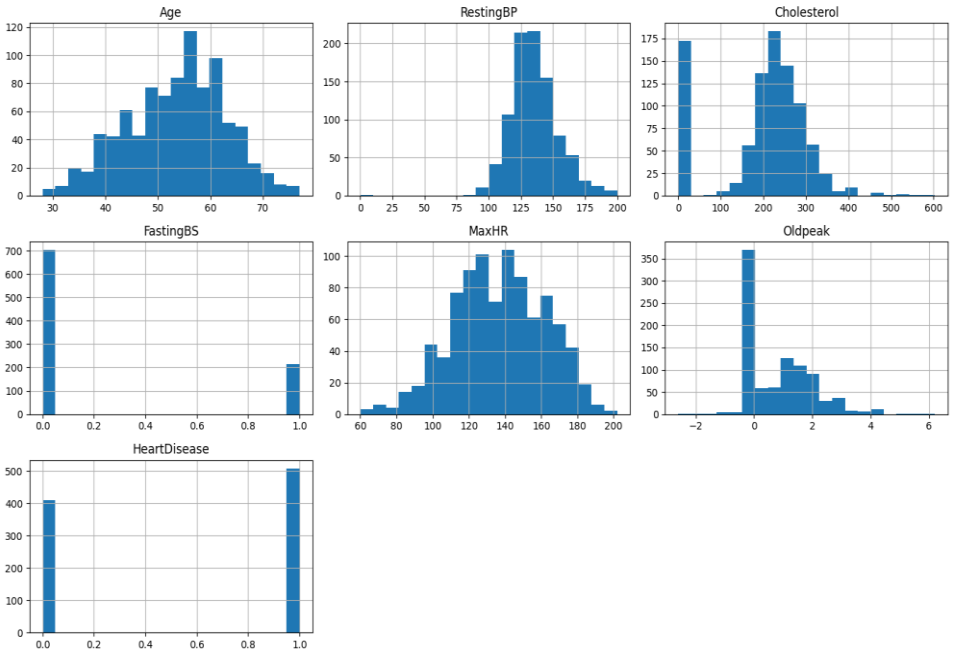

Figure 1(a): Histograms depicting the distribution of age (40–60 years), resting blood pressure (120–140 mmHg)

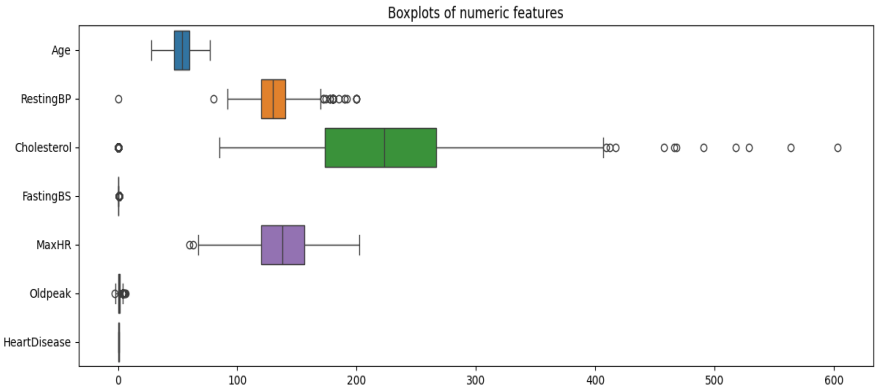

Figure 1(b): Box plots illustrating the spread, central tendency, and presence of potential outliers for the same variables

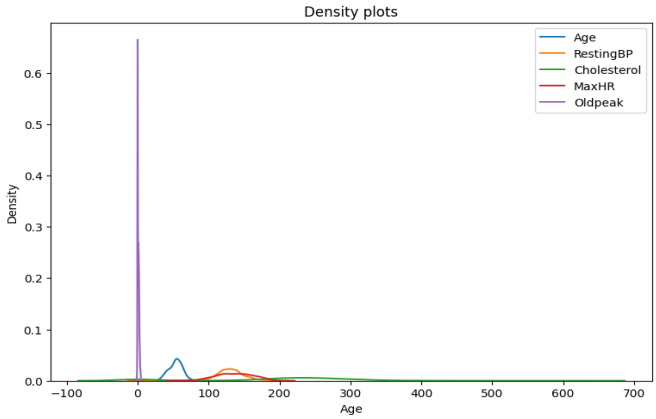

Figure 1(c): Density plots showing the underlying distribution patterns and smooth probability curves for each variable.

doi: https://doi.org/10.5281/zenodo.18266319

The binary variables produced during encoding were consolidated to streamline the dataset and ensure consistent representation. All Boolean columns were subsequently standardized into binary values (0 and 1), promoting uniformity across features. After completing these transformations, a correlation matrix (Fig. 2) was generated to examine the relationships among the variables in the preprocessed dataset. This evaluation revealed clearer patterns and stable associations between features, confirming the effectiveness of the preprocessing workflow. Notably, no significant multicollinearity was detected, which is particularly advantageous for models such as logistic regression. Each cell in the matrix reflects the correlation between a pair of variables, expressed on a scale ranging from −1 to 1. A correlation value of 1 denotes a perfect positive relationship, −1 indicates a perfect negative relationship, and 0 represents the absence of any linear association.

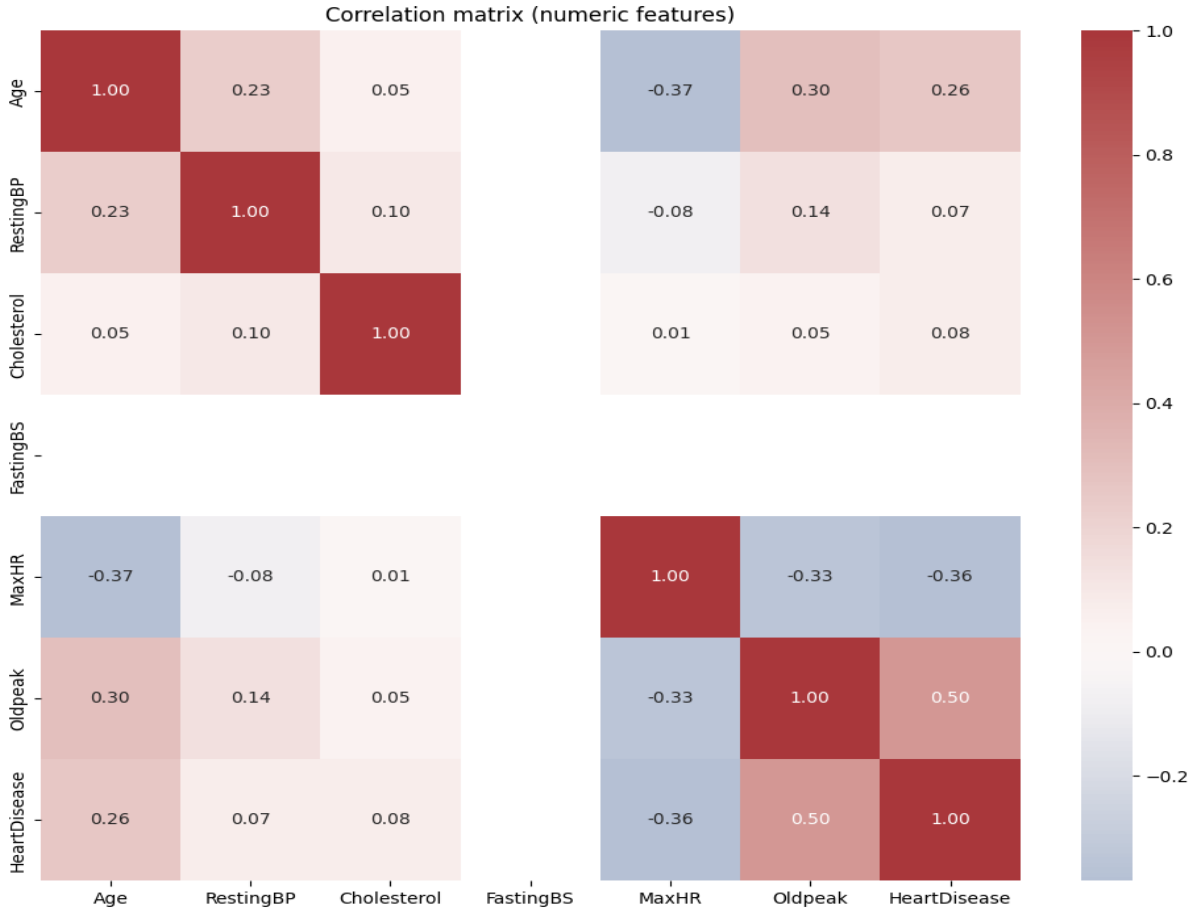

Figure 2: Correlation matrix.

In this matrix, age shows a positive correlation with heart disease, indicating that the likelihood of developing the condition increases with advancing age. Cholesterol levels exhibit a negative correlation with heart disease, suggesting that lower cholesterol values are generally associated with reduced risk. Other variables such as resting blood pressure, fasting blood sugar, and maximum heart rate achieved also display positive correlations with heart disease. Conversely, exercise-induced angina and ST-segment depression show negative correlations, indicating that higher values of these variables correspond to a lower likelihood of heart disease in this dataset. Overall, the correlation matrix indicates that factors commonly associated with increased health risks such as advanced age, elevated blood pressure, high cholesterol levels, and diabetes also exhibit positive associations with a higher likelihood of heart disease.

doi: https://doi.org/10.5281/zenodo.18266319

3.2 Machine-Learning Methods

This section provides a concise overview of the machine learning classification algorithms used in the study, which include Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), k-Nearest Neighbors (k-NN), and Multilayer Perceptron (MLP).

A. Logistic Regression (LR)

Logistic Regression (LR) is a multivariable statistical technique commonly used to model the relationship between several independent variables and a categorical dependent variable. It is particularly suited for predicting binary outcomes, such as determining whether an individual is healthy or diseased, or making other yes-or-no decisions. LR is widely applied in medical diagnosis and decision-making contexts. From a statistical perspective, it addresses binary classification problems by estimating the probability of class membership using the logistic (sigmoid) function [8].

B. Decision Tree (DT)

Decision Trees (DTs) are sequential, rule-based models that combine a series of simple decision tests; each test evaluates either a numerical attribute against a threshold or a categorical attribute against a specific set of values. DTs function as flexible tools for prediction and classification by recursively partitioning the dataset into smaller subsets based on the values of input variables or predictors. This recursive splitting produces branches and terminal nodes (leaves), where each leaf contains data instances with similar target values. As one moves down the tree, the resulting leaves exhibit increasingly distinct class characteristics. This hierarchical structure enables DTs to adapt effectively to a wide range of scenarios, making them valuable instruments in decision-making and predictive analytics [9].

C. Random Forest (RF)

Random Forest (RF) is an ensemble machine learning method that integrates multiple individual Decision Trees (DTs) to produce a more robust and accurate predictive model. During training, the algorithm constructs a large number of trees and generates the final output by aggregating their results using the majority vote for classification tasks or the average prediction for regression tasks. The incorporation of randomness in both the selection of training samples and the subset of features considered at each split strengthens the model's generalization capability and reduces the risk of overfitting. As a result, RF offers improved performance and greater reliability when making predictions on unseen data [10].

D. K-Nearest Neighbor (KNN)

k-Nearest Neighbors (k-NN) is a versatile machine learning algorithm used for both classification and regression tasks. In classification, k-NN assigns a class label to a new data point based on the majority class among its closest neighbors in the feature space. For regression, it predicts a numerical value by averaging the values of these nearest neighbors. The parameter k denotes the number of neighbors considered, and selecting an appropriate value is crucial for achieving optimal model performance. k-NN is particularly effective for datasets with nonlinear decision boundaries or complex underlying structures that cannot be easily captured using explicit mathematical models [11].

doi: https://doi.org/10.5281/zenodo.18266319

E. The Multi-Layer Perceptron (MLP)

A Multilayer Perceptron (MLP) is a type of artificial neural network designed for supervised learning tasks. It is composed of multiple layers of interconnected nodes, including an input layer that receives the initial data, one or more hidden layers that process and transform this information, and an output layer that generates the final prediction or classification. Each node, or perceptron, applies a mathematical activation function to its inputs and passes the resulting output to the next layer. MLPs can learn complex, nonlinear patterns and relationships within data, making them suitable for tasks such as classification, regression, and pattern recognition. During training, algorithms such as backpropagation are used to iteratively adjust the weights of the connections between nodes, enabling the network to improve its predictive performance over time [12].

F. SVM_RBF

The SVM with RBF kernel (SVM_RBF) demonstrates excellent classification capability, effectively capturing nonlinear relationships in the data. Its RBF kernel allows the model to project the input features into a higher-dimensional space, enabling the separation of complex patterns that linear models cannot capture. As reflected in the confusion matrix, the model achieves high performance, correctly classifying most of both positive and negative cases with minimal misclassifications. This highlights the RBF kernel's strength in producing robust decision boundaries, making SVM_RBF a strong candidate for reliable heart disease prediction [13].

G. XGBoost

XGBoost (Extreme Gradient Boosting) is an advanced gradient-boosting algorithm designed for high performance, speed, and efficiency. It builds an ensemble of decision trees sequentially, where each tree focuses on correcting the errors of the previous ones. XGBoost incorporates regularization, optimized tree growth, and efficient handling of missing values, which reduces overfitting and enhances generalization. Due to its ability to model complex nonlinear patterns and deliver strong predictive accuracy, XGBoost is widely used in classification tasks, including medical diagnosis and heart disease prediction.

IV. EXPERIMENTAL RESULTS

In this section, five machine learning techniques were employed to perform effective feature extraction and classification for heart failure diagnosis. These methods include Logistic Regression (LR), Random Forest (RF), Decision Tree (DT), k-Nearest Neighbors (k-NN), and the Multilayer Perceptron (MLP) classifier. The dataset was divided into two subsets: 80% for training and 20% for testing. Model performance was assessed using several evaluation metrics. High specificity (≥ 0.90) was required to minimize false positives, while an excellent AUC (≥ 0.90) indicated strong discrimination between classes. A low Brier score (≤ 0.05) reflected highly accurate probability predictions. The Matthews Correlation Coefficient (MCC) was also considered, with values approaching +1 representing outstanding performance, values near 0 indicating random prediction, and negative values signaling systematic misclassification. Models falling below these thresholds were regarded as performing poorly.

doi: https://doi.org/10.5281/zenodo.18266319

4.1 Model Building and Performance Evaluation

The logistic regression demonstrate that the model achieves an F1-score of 0.85 for both classes, indicating a well-balanced trade-off between precision and recall. The overall accuracy of 85% further shows that the model correctly classifies most samples in a dataset with nearly equal class support (102 for class 0 and 98 for class 1). Additionally, the close alignment between precision and recall values across both classes suggests that the model does not exhibit bias toward any class. An AUC of 0.94 confirms that the model possesses excellent discriminatory power between the two classes. The Brier score of 0.10, together with an MCC value of 0.70, indicates that the predicted probabilities are well calibrated and that there is a strong positive correlation between the actual and predicted labels.

The decision tree results presented show an F1-score of 0.90 for class 0 and 0.89 for class 1, reflecting a well-balanced model capable of effectively distinguishing between both classes. An overall accuracy of 90% further demonstrates the model's strong performance, correctly classifying most samples with nearly equal support across classes (102 for class 0 and 98 for class 1). The close alignment between precision and recall values indicates that the model does not exhibit bias toward either class, reinforcing its balanced predictive behavior. An AUC of 0.89 indicates that the model demonstrates strong discriminatory capability between the two classes. The Brier score of 0.10 and an MCC value of 0.79 further confirm that the predicted probabilities are well calibrated and that there is a strong positive relationship between the true and predicted classifications. Additionally, the high specificity value of 0.94 underscores the model's effectiveness in correctly identifying negative cases, minimizing false positives.

The results of the random forest model, demonstrate strong and consistent performance. Precision, recall, and F1-scores are well balanced across both classes, with class 0 achieving 90% precision, 93% recall, and a 92% F1-score, and class 1 achieving 93% precision, 90% recall, and a 91% F1-score. The overall accuracy of 92%, along with the macro and weighted averages, further confirms that the model performs effectively in a balanced dataset and maintains reliable classification across both classes. The specificity of 93% indicates that the model correctly identifies most negative cases, minimizing false positives. An AUC of 0.97 further highlights the model's excellent capacity to distinguish between the two classes. Moreover, the low Brier score of 0.06 reflects well-calibrated probability estimates, while the MCC value of 0.83 confirms strong overall reliability, even in the presence of slight class imbalance.

The KNN model shows moderate performance. For class 0, it achieves 78% precision, 69% recall, and an F1-score of 73%. In contrast, for class 1, precision decreases to 71%, while recall improves to 80%, resulting in an F1-score of 75%. These variations suggest that the model struggles to consistently distinguish between the two classes. Consequently, its overall accuracy reaches only 74%, indicating limited effectiveness compared with the other evaluated models. A specificity of 69% shows that the model faces difficulty accurately identifying negative cases. The AUC value of 0.81 indicates a reasonable, though not strong, ability to distinguish between classes. However, the Brier score of 0.18 suggests that the predicted probabilities are not well calibrated. Additionally, the Matthews Correlation Coefficient (MCC) of 0.48 reflects moderate overall performance, particularly in situations involving slight class imbalance.

The MLP model demonstrates strong performance, achieving an overall accuracy of 94%. For class 0, it attains an F1-score of 84%, although the lower recall of 75% indicates some difficulty in correctly identifying all positive cases. Conversely, class 1 exhibits a high recall of 95%, but with reduced precision (79%), resulting in an F1-score of 86%. These variations highlight the trade-off in the model's

doi: https://doi.org/10.5281/zenodo.18266319

ability to balance sensitivity and precision across classes, with the overall performance reflected in an average accuracy of 85%. Additionally, the model exhibits a moderate specificity of 75%, indicating some challenges in correctly identifying negative cases. Despite this, the high AUC value of 0.95 and a low Brier score of 0.11 demonstrate the model's strong ability to discriminate between classes and produce well-calibrated probability estimates. Furthermore, the MCC of 0.72 reflects a robust overall performance, even in the presence of class imbalance.

TABLE 3. EVALUATION METRICS OF ML MODELS

Model	Accuracy	Precision	Recall	F1	Specificity	AUC	Brier	MCC
1	MLP	0.965	0.979381	0.95	0.964467	0.98	0.9842	0.038400
2	RandomForest	0.960	0.960000	0.96	0.960000	0.96	0.9963	0.028233
3	DecisionTree	0.955	0.978947	0.93	0.953846	0.98	0.9550	0.045000
4	KNN	0.900	0.850877	0.97	0.906542	0.83	0.9611	0.079200
5	LogisticRegression	0.845	0.816514	0.89	0.851675	0.80	0.9240	0.111373

The SVM with an RBF kernel demonstrates strong sensitivity, achieving a recall of 0.95, which indicates its effectiveness in identifying positive cases. However, its specificity of 0.89 shows some difficulty in correctly detecting negative instances. With an accuracy of 92% and an F1-score of 0.922, the model provides balanced performance, though not as strong as ensemble-based approaches. The AUC value of 0.9498 confirms that the classifier is capable of distinguishing between classes effectively, while the Brier score of 0.0756 suggests moderate calibration of predicted probabilities. The MCC of 0.84 further reflects solid but not exceptional predictive reliability.

XGBoost achieves the highest performance among all evaluated models, with an accuracy of 97.5% and an F1-score of 0.9748, demonstrating excellent predictive capability. The model maintains very high precision (0.9798) and recall (0.97), indicating that it performs consistently well in identifying both classes. Its specificity of 0.98 shows exceptional ability to correctly recognize negative cases. Additionally, the AUC of 0.9931 highlights outstanding class-separation capability, while the low Brier score (0.0309) reflects highly calibrated probability estimates. The MCC of 0.95 further confirms XGBoost as the most reliable and robust model in the comparison (Table 4).

TABLE 3. EVALUATION METRICS OF ENSEMBLE MODELS

Model	Accuracy	Precision	Recall	F1	Specificity	AUC	Brier	MCC
1	XGBoost	0.975	0.979798	0.97	0.974874	0.98	0.9931	0.030880
2	SVM_RBF	0.920	0.896226	0.95	0.922330	0.89	0.9498	0.075626

4.2 Statistical Significance Analysis

To statistically validate differences in classifier performance, McNemar's test was applied to the paired predictions of all models. The resulting significance matrix (Figure 10) highlights clear contrasts among classifiers. Random forest demonstrated statistically superior performance compared to logistic regression ($p = 0.0005$) and K-nearest neighbors ($p < 0.0001$), which is consistent with its strong evaluation metrics (AUC = 0.97; specificity = 0.93). In contrast, KNN exhibited the weakest performance, showing significant differences against every other model ($p < 0.01$). Meanwhile, logistic regression displayed no statistically significant differences when compared with the decision tree ($p = 0.5716$) or

doi: https://doi.org/10.5281/zenodo.18266319

the MLP classifier (p = 0.1796), indicating comparable effectiveness in these specific classification settings.

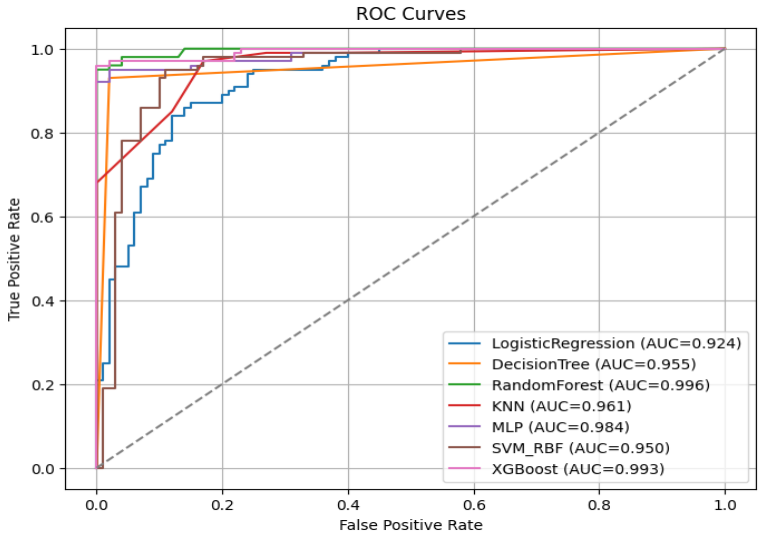

Figure 3: ROC and AUC Curves

In this study, the Heart Failure Prediction dataset was utilized to evaluate and identify the most effective machine-learning classification algorithm for accurate feature extraction and heart failure diagnosis. Multiple models were examined to assess their respective strengths and limitations. Logistic regression (LR) demonstrated balanced performance, achieving an F1 score of 0.85 and an AUC of 0.94, reflecting solid discriminative capability. However, despite its stability across classes, LR falls short compared to the more advanced models in this analysis, which exhibit superior predictive power and calibration.

The decision tree (DT) model demonstrates stronger performance than logistic regression, achieving an accuracy of 90% and maintaining a well-balanced relationship between precision and recall. Its F1 score for class 0 reaches 0.90, and the high specificity of 0.94 confirms its effectiveness in correctly identifying negative cases. However, its AUC of 0.89 is slightly lower than that of LR and several other advanced models, indicating that while DT performs reliably, its discriminative capability still leaves room for improvement.

The k-nearest neighbors (k-NN) classifier shows notable limitations, achieving a relatively low F1 score of 0.75 and a reduced specificity of 0.69, which indicates difficulty in correctly identifying negative cases and managing class imbalance. Although its AUC of 0.81 suggests a reasonable ability to distinguish between classes, its performance remains noticeably weaker compared to the other models analyzed.

The multilayer perceptron (MLP) demonstrates strong overall performance, highlighted by an AUC of 0.95 and an accuracy of 85%. However, the noticeable gap in recall 75% for class 0 versus 95% for class 1 indicates inconsistent sensitivity across classes. This imbalance suggests challenges in uniformly identifying positive instances, which ultimately results in a lower F1 score compared with the superior performance achieved by the random forest model.

Finally, the random forest (RF) model demonstrates outstanding predictive performance, with precision values of 90% for class 0 and 93% for class 1, indicating its strong capability to accurately

doi: https://doi.org/10.5281/zenodo.18266319

distinguish both negative and positive cases. The model also maintains high recall—93% for class 0 and 90% for class 1—ensuring that most actual instances are correctly identified. These strengths contribute to F1 scores of 0.92 and 0.91 for classes 0 and 1, respectively, reflecting the model's balanced and robust effectiveness across both categories.

Moreover, the RF model's specificity of 0.93 and AUC of 0.97 further emphasize its strong discriminative capability, ensuring accurate separation between positive and negative cases while minimizing misclassification. The low Brier score (0.06) demonstrates that the model generates well-calibrated and reliable probability estimates, and the MCC value of 0.83 confirms its robust, balanced performance even under conditions of slight class imbalance. Collectively, these metrics position random forest as the most effective classifier among all the models evaluated, confirming its suitability for reliably identifying cases of heart failure.

In comparison of studies on machine learning models, the proposed approach demonstrates superior performance by achieving the highest precision (92%), a result attributed to rigorous preprocessing and optimized hyperparameter tuning. This contrasts with several previous works that did not account for outliers and consequently reported lower precision values—such as 85% in study [18], 83% in [13], 76% in [27], and 73% in [28]. These discrepancies highlight the importance of proper data preprocessing, particularly the treatment of outliers, which can substantially enhance model accuracy. Supporting this observation, study [26], which also incorporated outlier management, achieved a precision of 89.58%, while study [19] reported a comparable precision of 91.70%. These findings collectively reinforce the conclusion that effective preprocessing plays a crucial role in improving predictive performance in heart failure classification tasks.

A key finding of this study is that, despite utilizing only 12 features, the proposed model achieved a higher accuracy (92%) than several studies that employed a larger number of features. For instance, study reported an accuracy of 91.70% using 14 features, while study obtained 83% with the same feature count. This outcome highlights that effective feature selection, combined with robust preprocessing techniques such as outlier removal, can substantially improve model performance even when fewer features are used. These results emphasize the importance of data quality and preprocessing over the mere quantity of features in achieving superior predictive accuracy (Table 4).

TABLE 4. HEART DISEASE CLASSIFICATION RESULT

Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0
39	M	NAP	120	339	0	Normal	170	N	0.0	Up	0

The McNemar test significance matrix offers a robust statistical foundation for evaluating comparative model performance. The strong superiority of the random forest classifier ($p < 0.001$ against major models) reinforces the strengths of ensemble learning, where aggregating multiple decision trees reduces overfitting and enhances the ability to model complex feature interactions. Conversely, the consistently poor performance of KNN ($p < 0.01$ in all pairwise tests) highlights its vulnerability in clinical

doi: https://doi.org/10.5281/zenodo.18266319

datasets, where class imbalance, noise, and overlapping feature distributions undermine its distance-based decision mechanism. Meanwhile, the absence of significant differences between logistic regression, decision tree, and MLP ($p \geq 0.05$) suggests that these models may converge toward similar decision boundaries in this dataset, reflecting comparable sensitivity to its underlying linear and nonlinear patterns.

The algorithm can be seamlessly integrated with devices that generate static data or with electronic health records (EHRs), which store patients' medical histories, laboratory findings, and diagnostic information. Such integration supports healthcare professionals in making informed clinical decisions and optimizing the management of cardiovascular and other chronic conditions. When connected to EHR systems, the algorithm can analyse historical data and laboratory parameters to proactively identify high-risk patients. In emergency department (ED) settings, models with high sensitivity such as the MLP (94%) can be used to rapidly prioritize critical cases, while highly specific models such as the decision tree (93%) help reduce false positives and prevent unnecessary diagnostic procedures. Additionally, implementing the model within continuous monitoring systems enables the early detection of clinical deterioration, generating timely alerts and improving resource allocation within healthcare facilities.

This study demonstrates encouraging results; however, several limitations should be acknowledged. First, although the dataset offers a strong foundation for model development, expanding it to include additional clinically relevant variables—such as left ventricular ejection fraction (LVEF) or cardiac biomarkers like NT-proBNP could enhance diagnostic precision and improve the model's clinical applicability. Second, while the sample size is adequate for training robust classifiers, incorporating data drawn from more diverse populations and a wider range of comorbid conditions would strengthen the model's generalizability across different demographic and clinical settings.

Importantly, the models were evaluated using a preprocessed static dataset and have not yet been tested within real-world clinical workflows. In practical settings, factors such as variability in real-time data acquisition, heterogeneous measurement conditions, and dynamic changes in patient status may influence model performance. As a result, additional validation in operational clinical environments is essential to confirm the robustness and reliability of the proposed models.

V. CONCLUSION

In summary, this study highlights the strong potential of machine learning techniques in accurately classifying heart failure, with the random forest model emerging as the top performer, achieving an accuracy of 92% and demonstrating exceptional ability in distinguishing between positive and negative cases. The application of advanced preprocessing steps including outlier removal and class balancing played a critical role in enhancing overall model performance. The findings underscore the value of machine learning in managing complex clinical data, supporting early diagnosis, and ultimately contributing to improved patient outcomes and quality of life.

Compared with previous studies, this work achieves superior accuracy as a result of effective data preprocessing and systematic hyperparameter optimization. The findings demonstrate that a well-curated dataset even one with fewer features can outperform approaches in which data quality and preprocessing were not adequately prioritized. Overall, the results reaffirm the growing influence of machine learning in clinical medicine and its potential to transform the diagnosis and management of cardiovascular diseases. One promising direction is the integration of these models with real-time

doi: https://doi.org/10.5281/zenodo.18266319

monitoring devices to support continuous risk assessment. Future work should emphasize prospective validation in hospital environments to evaluate the model's stability within dynamic clinical workflows, particularly in reducing false positives and ensuring timely prioritization of critical cases. Additionally, exploring hybrid architectures such as combining random forest with LSTM networks could enable the capture of temporal patterns in continuous ECG signals, thereby improving diagnostic accuracy in scenarios requiring both static and dynamic feature analysis.

VI. REFERENCES

[1] O. Gaidai, Y. Cao, and S. Loginov, "Global cardiovascular diseases death rate prediction," *Elsevier*, vol. 48, p. 101622, 2023.

[2] B. Alić, L. Gurbeta, and A. Badnjević, "Machine learning techniques for classification of diabetes and cardiovascular diseases," in *Proc. 6th Mediterranean Conf. Embedded Computing (MECO)*, Bar, Montenegro, Jun. 11–15, 2017.

[3] B. Hawashin, A. Mansour, F. Fotouhi, S. AlZu'bi, and T. Kanan, "A novel recommender system using interest extracting agents and user feedback," in *Proc. Int. Conf. Information Technology (ICIT)*, Amman, Jordan, Jul. 14–15, 2021.

[4] W. M. Jinjri, P. Keikhosrokiani, and N. L. Abdullah, "Machine learning algorithms for the classification of cardiovascular disease: A comparative study," in *Proc. Int. Conf. Information Technology (ICIT)*, Amman, Jordan, Jul. 14–15, 2021, pp. 132–138.

[5] F. I. Alarsan and M. Younes, "Analysis and classification of heart diseases using heartbeat features and machine learning algorithms," *Journal of Big Data*, vol. 6, no. 1, p. 81, 2019.

[6] H. Denolin, H. Kuhn, H. P. Krayenbuehl, F. Loogen, and A. Reale, "The definition of heart failure," *European Heart Journal*, vol. 4, no. 7, pp. 445–453, 1983.

[7] D. Tomasoni, M. Adamo, C. M. Lombardi, and M. Metra, "Highlights in heart failure," *ESC Heart Failure*, vol. 6, no. 6, pp. 1105–1127, 2019.

[8] P. Ponikowski *et al.*, "Heart failure: Preventing disease and death worldwide," *ESC Heart Failure*, vol. 1, no. 1, pp. 4–25, 2014.

[9] C. Gutierrez and D. G. Blanchard, "Diastolic heart failure: Challenges of diagnosis and treatment," *American Family Physician*, vol. 69, no. 11, pp. 2609–2617, 2004.

[10] M. Komajda, "Current challenges in the management of heart failure," *Circulation Journal*, vol. 79, no. 5, pp. 948–953, 2015.

[11] S. Panicker and G. P, "Use of machine learning techniques in healthcare: A brief review of cardiovascular disease classification," in *Proc. 2nd Int. Conf. Communication & Information Processing (ICCIP)*, Tokyo, Japan, Nov. 27–29, 2020.

[12] N. Louridi, M. Amar, and B. E. Ouahidi, "Identification of cardiovascular diseases using machine learning," in *Proc. 7th Mediterranean Congress of Telecommunications (CMT)*, Fes, Morocco, Oct. 24–25, 2019.

doi: https://doi.org/10.5281/zenodo.18266319