American International University Bangladesh (AIUB)

Department of Computer Science

Faculty of Science & Technology (FST)

MACHINE LEARNING

Group -08

Section: E

Supervised By:

DR. MD. ASRAF ALI

Submitted By:

Student Name	*Student ID*
ABRAR SHAKIL OISHIK	*22-46257-1*
FOYSAL AHMED FAHAD	*22-47076-1*
ABDULLAH RAHMAN IRFAN	*22-46877-1*

Submission Date: 27-09-2025

Using machine learning to predict cardiovascular disease type based on genetic, lifestyle, and clinical data.

ABRAR SHAKIL OISHIK, 22-46257-1@student.aiub.edu [1],

FOYSAL AHMED FAHAD, 22-47076-1@student.aiub.edu[2],

ABDULLAH RAHMAN IRFAN, 22-46877-1@student.aiub.edu[3],

Abstract:

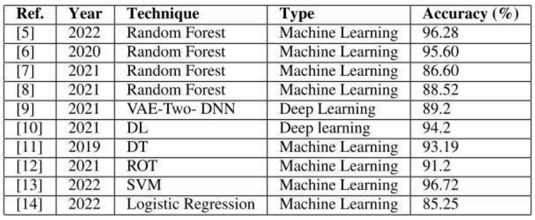
Heart failure is a critical chronic condition that affects millions of people worldwide, and early detection is essential for effective treatment and improved patient outcomes. This study presents a machine learning-based approach to predict heart failure using patient health parameter data. Seven classification algorithms were implemented and compared, including logistic regression, decision tree, support vector machine, random forest, naive Bayes, k-nearest neighbors, and XGBoost. The performance of each model was rigorously evaluated using key metrics such as accuracy, precision, recall, and F1-score. Among all the algorithms tested, XGBoost achieved the highest accuracy of 99%, demonstrating superior overall performance. Confusion matrix analysis further confirmed the robustness of these models in correctly classifying heart failure cases. The results of this study highlight the potential of machine learning techniques to support early and accurate heart failure diagnosis, offering meaningful contributions toward the development of intelligent healthcare support systems.

INDEX TERMS Machine learning, heart failure, cross validations, feature engineering.

I. INTRODUCTION

Cardiovascular disease (CVD) is the biggest concern in the medical sector at present. It remains one of the leading causes of death globally, accounting for an estimated 17.9 million deaths annually, representing about 31% of all global deaths [1][2]. From the recent statistics reported by the World Health Organization (WHO), about 20.5 million people die every year due to cardiovascular disease, which is approximately 31.5% of all deaths globally. It is also estimated that the number of annual deaths will rise to 24.2 million by 2030. About 85% of cardiovascular disease deaths are due to heart attacks and strokes [2]. Heart disease encompasses a variety of conditions affecting both the heart and blood vessels, including coronary artery disease (CAD), heart failure, arrhythmias, hypertension, and strokes. Among these, CAD is particularly concerning due to its high association with death. The primary risk factors contributing to heart disease include hypertension, obesity, diabetes, high cholesterol, and smoking, with early diagnosis and intervention being critical in reducing mortality rates [3].

TABLE 1. The performance accuracy achieved in previous studies for heart failure prediction.



Ref.	Year	Technique	Type	Accuracy (%)
[5]	2022	Random Forest	Machine Learning	96.28
[6]	2020	Random Forest	Machine Learning	95.60
[7]	2021	Random Forest	Machine Learning	86.60
[8]	2021	Random Forest	Machine Learning	88.52
[9]	2021	VAE-Two- DNN	Deep Learning	89.2
[10]	2021	DL	Deep learning	94.2
[11]	2019	DT	Machine Learning	93.19
[12]	2021	ROT	Machine Learning	91.2
[13]	2022	SVM	Machine Learning	96.72
[14]	2022	Logistic Regression	Machine Learning	85.25

Early detection and accurate prediction of CVD are critical, as delayed diagnosis often leads to severe complications or premature death [4]. However, the asymptomatic nature of many cardiovascular conditions makes early detection challenging. Machine learning (ML) is a subset of artificial intelligence (AI), has shown promising potential in identifying complex patterns from large datasets to support early diagnosis and effective treatment planning for heart-related conditions [2][5]. Machine learning has emerged as a promising tool in health care, particularly in predicting cardiovascular diseases by analyzing clinical, genetic, and lifestyle data to provide early warnings and improve forecast.[3]

In ongoing advancements, there remain significant gaps in current research on cardiovascular disease prediction. Many existing approaches depend on limited clinical parameters, often ignoring the integration of genetic and lifestyle data which play critical roles in disease onset and progression [3][6]. Moreover, challenges such as imbalanced datasets, limited feature selection, and lack of model generalizability reduce the effectiveness of these ML techniques [1][3][6]. Although studies have employed diverse algorithms such as Decision Trees, K-Nearest Neighbors, Neural Networks, and Support Vector Machines, the variability in datasets and methodologies results in inconsistent accuracy levels and limits clinical applicability [6][7][8]. Additionally, while some research has explored hybrid models and optimization techniques, most are constrained by narrow feature sets or fail to leverage the full potential of cross-domain data, including genetic predispositions, lifestyle factors, and clinical metrics [3][8].

The objective of this research is to develop a more effective machine learning-based prediction model for cardiovascular diseases by integrating genetic, clinical, and lifestyle data. We evaluate and compare several machine learning algorithms, including Random Forest, Support Vector Machine, Logistic Regression, K-NN, XGBoost,descision tree, Naïve Bayes, to identify the most effective techniques for accurate prediction.

A novel hybrid approach is also proposed to improve model performance and reduce prediction bias. Through feature selection, cross-validation, and performance benchmarking,the research emphasizes optimizing the model for higher accuracy, interpretability, and clinical relevance. Ultimately, this study contributes to the development of more robust, scalable, and personalized predictive tools that can help healthcare professionals in early diagnosis and intervention strategies for cardiovascular disease.

II. RESEARCH METHODOLOGY

2.1 Data Collection

For this research, the Heart Disease Dataset was utilized to develop and evaluate heart disease prediction models. (https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset/data) The study dataset based on heart disease-related features is used to build the applied models in this study. The study dataset has 14 features initially. The dataset contains 1025 patient records of heart disease based on 713 men and 312 women samples. We have checked the dataset datatypes, which show 13 attributes have int type, and one attribute, slope, has float type. We have the dataset for null values, which results in the performance. The analysis shows that the dataset does not contain any null values. Demographic data (like age and sex), clinical measurements (such as blood pressure, cholesterol, and heart rate), and categorical variables (like chest pain type and ECG results).

- Age
- Sex
- Chest pain type
- Resting blood pressure
- Serum cholesterol
- Fasting blood sugar
- Resting electrocardiographic results
- Maximum heart rate achieved
- Exercise-induced angina
- ST depression
- Slope of the peak exercise ST segment

- Number of major vessels colored by fluoroscopy
- Thalassemia

The target variable indicates the presence (1) or absence (0) of heart disease. Before model training, necessary preprocessing steps were carried out, including handling missing data, encoding of categorical variables, and normalizing continuous features to improve model performance and consistency.

Before model training, the dataset was preprocessed to ensure quality and consistency. This included handling missing values, encoding categorical variables where necessary, and normalizing continuous features to enhance model performance. The processed data was then used to train and test multiple machine learning classifiers for comparative analysis.

2.2 Data Preprocessing

To optimize the dataset for model training and improve overall model performance, the following preprocessing techniques were applied:

1. Handling Missing Values

The (Heart) dataset from Kaggle is relatively clean and does not contain missing values. However, a preliminary check was conducted using methods such as isnull().sum() in Python (pandas) to confirm the absence of null entries. If any missing values were present, imputation strategies (mean, median, or mode) would have been applied.

2. Categorical Variables

Some features in the dataset are categorical, such as:

- Sex (1 = male; 0 = female)
- Chest pain type (cp)**:** values 0 to 3
- Resting ECG (restecg)**:** Resting electrocardiographic results (0–2)
- Exercise induced angina (exang)**:** 1 = yes; 0 = no
- Slope of the ST segment (slope)**:** 0 to 2
- Thalassemia (thal)**:** 1 = normal, 2 = fixed defect, 3 = reversible defect; encoded differently depending on version

These categorical variables were encoded using Label Encoding, depending on the algorithm to be used.

3. Feature Scaling (Normalization)

Since the dataset includes features with different ranges (e.g., age in years vs. cholesterol in mg/dL), feature scaling was applied. Min-Max normalization was used to bring all numerical features to a similar scale, especially for algorithms sensitive to feature magnitudes like K-Nearest Neighbors (KNN) or Support Vector Machines (SVM). It might be necessary to scale features such as trestbps (blood pressure), chol (cholesterol), and thalach (maximum heart rate)

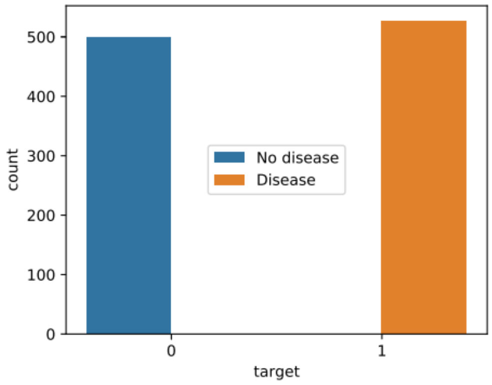


FIGURE 2. The dataset distribution analysis is based on the target column.

4. Outlier Detection

Box plots analysis were used to detect potential outliers in numerical attributes such as cholesterol and maximum heart rate. Any extreme values that could distort model learning were either capped or removed, depending on their impact.

5. Splitting the Dataset

The dataset was split into: Training set (70%), Testing set (30%). This split was done randomly using stratified sampling to ensure the class distribution remains balanced in both

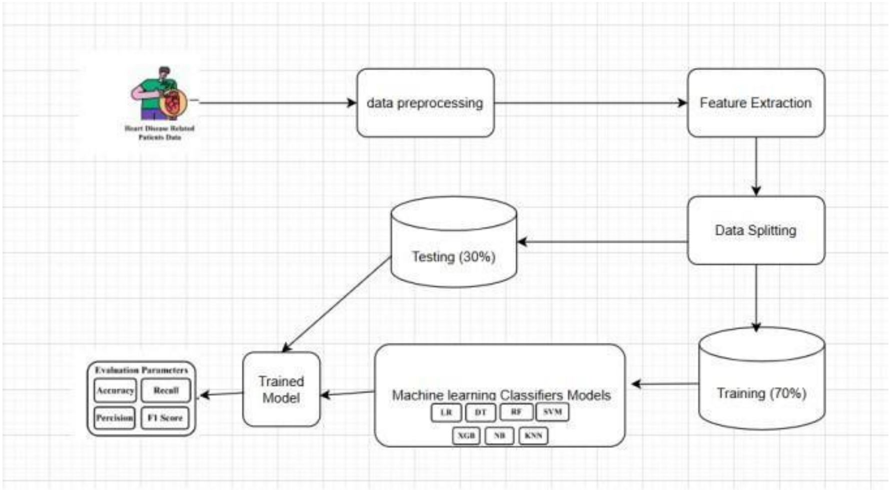
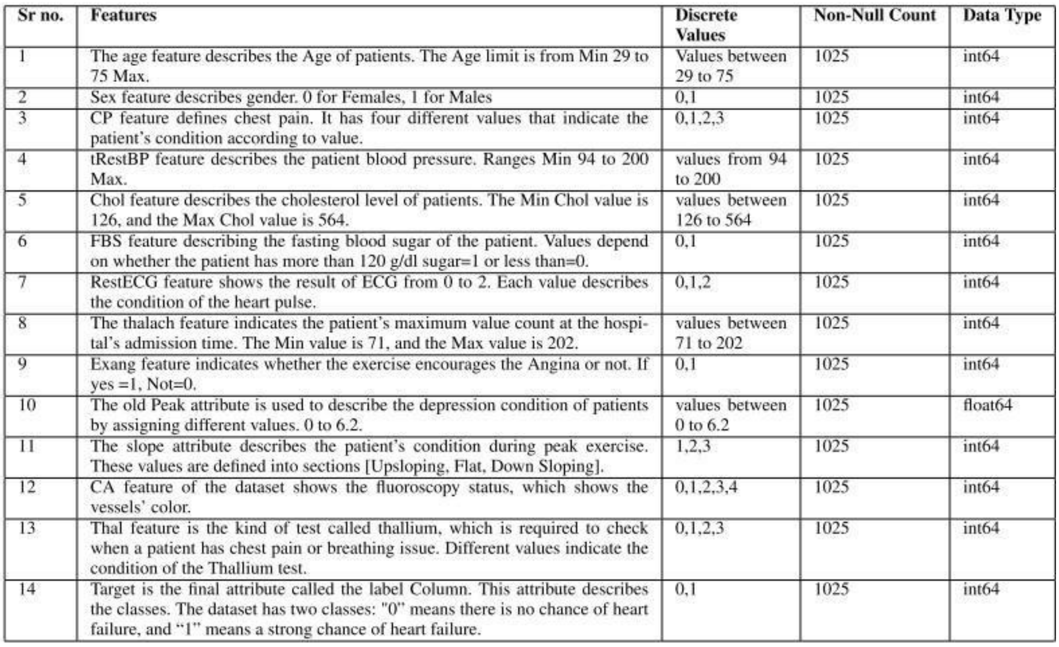


FIGURE 1. The proposed study methodology analysis for heart failure prediction.

TABLE 2. The study heart-related dataset features analysis.



Sr no.	Features	Discrete Values	Non-Null Count	Data Type
1	The age feature describes the Age of patients. The Age limit is from Min 29 to 75 Max.	Values between 29 to 75	1025	int64
2	Sex feature describes gender. 0 for Females, 1 for Males	0,1	1025	int64
3	CP feature defines chest pain. It has four different values that indicate the patient's condition according to value.	0,1,2,3	1025	int64
4	tRestBP feature describes the patient blood pressure. Ranges Min 94 to 200 Max.	values from 94 to 200	1025	int64
5	Chol feature describes the cholesterol level of patients. The Min Chol value is 126, and the Max Chol value is 564.	values between 126 to 564	1025	int64
6	FBS feature describing the fasting blood sugar of the patient. Values depend on whether the patient has more than 120 g/dl sugar=1 or less than=0.	0,1	1025	int64
7	RestECG feature shows the result of ECG from 0 to 2. Each value describes the condition of the heart pulse.	0,1,2	1025	int64
8	The thalach feature indicates the patient's maximum value count at the hospital's admission time. The Min value is 71, and the Max value is 202.	values between 71 to 202	1025	int64
9	Exang feature indicates whether the exercise encourages the Angina or not. If yes =1, Not=0.	0,1	1025	int64
10	The old Peak attribute is used to describe the depression condition of patients by assigning different values. 0 to 6.2.	values between 0 to 6.2	1025	float64
11	The slope attribute describes the patient's condition during peak exercise. These values are defined into sections [Upsloping, Flat, Down Sloping].	1,2,3	1025	int64
12	CA feature of the dataset shows the fluoroscopy status, which shows the vessels' color.	0,1,2,3,4	1025	int64
13	Thal feature is the kind of test called thallium, which is required to check when a patient has chest pain or breathing issue. Different values indicate the condition of the Thallium test.	0,1,2,3	1025	int64
14	Target is the final attribute called the label Column. This attribute describes the classes. The dataset has two classes: "0" means there is no chance of heart failure, and "1" means a strong chance of heart failure.	0,1	1025	int64

sets. splitting is essential if you're training and evaluating a machine learning model to predict cardiovascular disease or any health outcome based on this data. used to train the model, usually 70% of the data is used to test the model's performance on data it hasn't seen, usually 30%.

6. *Balancing the Dataset (if needed)*

Although the Heart dataset is relatively balanced, if any class imbalance had been observed, techniques like SMOTE (Synthetic Minority Over-sampling Technique) or undersampling could be applied to balance the target classes. The difference between the two classes is minimal (less than 3%), so you do not need to apply resampling techniques like SMOTE or undersampling for balancing.

This preprocessing pipeline ensures that the dataset is clean, standardized, and properly formatted, which is crucial for training reliable machine learning models.

2.3 Classification Model

This study conducted a comparative analysis of several machine learning algorithms to develop an accurate model for predicting cardiovascular disease (CVD) using genetic, clinical, and lifestyle data. The selected algorithms were chosen for their proven effectiveness in medical diagnostics, interpretability, and ability to manage complex, multi-dimensional data. Logistic Regression served as a simple, interpretable baseline. SVM was used for its performance with high-dimensional data, while Random Forest and XGBoost were chosen for their ensemble learning strengths and ability to model non-linear relationships.

Naïve Bayes was selected for its probabilistic approach and efficiency with categorical data. KNN was included for detecting local patterns, despite its sensitivity to feature scaling. Decision Tree was used for its intuitive, interpretable structure and ability to handle diverse data types, though it may overfit without pruning. Several machine-learning techniques used for heart failure prediction are studied in this section. The operational principles and basic terminology of machine learning models are explained. Our proposed research evaluates ten advanced machine-learning models for predicting heart failure.

1. Logistic Regression (LR)

Logistic Regression (LR) is a widely used supervised machine learning technique suitable for both regression and classification tasks, though it is primarily employed for binary classification problems. In the context of this heart disease dataset, LR is used to predict the likelihood of a patient having heart disease, where the target variable is binary—'0' indicating no risk and '1' indicating the presence of heart disease. The model works by estimating probabilities using a logistic (sigmoid) function, which maps any input value to a range between 0 and 1. The probability of the positive class (heart disease) is calculated as

$$P(y = 1|x) = \frac{1}{1 + e^{-z}} \quad \text{where} \quad z = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \qquad (1)$$

where $P(y = 1|x)$ represents the probability of a binary outcome variable y taking the value 1, given the predictor variable(s) x. The function e-z is the logistic function, which maps any real value z to the range [0,1].

2. Decision Tree (DT)

The most common supervised approach to solving classification tasks is a Decision Tree (DT). The tree-like structures are made in the DT technique [10]. The DT often includes multiple levels of nodes during tree construction. The root or parent nodes are at the top level, and the others are called child nodes. Large medical data is commonly handled via decision trees because they are easy to utilize. The data is organized as a tree, with internal nodes representing inside dataset attributes, branches for decision-making processes, and leaf nodes representing target results. The DT data is divided between the nodes using the Gini index and entropy functions.

$$f(\mathbf{x}) = \sum_{m=1}^{M} c_m \mathbf{1}(\mathbf{x} \in R_m), \qquad (2)$$

where $f(\mathbf{x})$ is the predicted output for input $\mathbf{x}$, M is the number of leaf nodes in the tree, Rm is the region of input space corresponding to the m-th leaf node, and cm is the prediction value associated with the m-th leaf node.

3. Random Forest (RM)

Random Forest (RF) is another supervised machine learning method commonly used to solve classification and regression problems [11]. The RF method combines several decision trees to address challenging issues and boost efficiency. An RF classifier averages many decision trees on various subsets of a given dataset to increase the predictive accuracy. The maximum number of trees in RF gives the highest accuracy. The RF prevents overfitting and leads toward high performance.

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} f_t(x) \qquad (3)$$

where $y\hat{}$ represents the predicted output for a given input vector x. The prediction is made by taking the average of the outputs of T decision trees, denoted by $ft(x)$.

4. Support Vector Machine

The Support Vector Machine (SVM) is a well-liked supervised learning technique [12] that can be utilized to overcome classification and regression problems. Making proper decision thresholds is the aim of SVM. The SVM divides the n-dimensional space into classes using an ideal decision boundary known as a hyperplane. The hyperplane makes SVM simple to assign a new data point to the appropriate category. The hyperplane is created by SVM by choosing extreme support vectors. This technique is known as a support vector machine due to the support vectors.

$$w^{\rightarrow} \cdot {}^{\rightarrow}x + b = 0 \qquad (4)$$

where $w^{\rightarrow}$ is the weight vector, $x^{\rightarrow}$ is the input vector, and b is the bias term.

5. Extreme Gradient Boosting

Extreme Gradient Boosting (XGB) is a supervised ensemble machine learning model used for classification and regression analysis [13]. The ensemble learning algorithms combine multiple machine learning algorithms for a better outcome. The XGB technique combines numerous decision trees. A pair of shallow decision trees is iteratively trained by XGB, which fits the next model with each iteration, utilizing the prior model's error residuals. The final prediction output is the weighted average of each prediction in the tree. The XGB method reduces underfitting and boosts bias. The loss score in XGB is determined using the gradient descent algorithm.

$$yi\hat{} = \mathrm{X}\ k = 1\ K fk\ (xi), fk \in F \quad (5)$$

where $y\hat{}i$ represents the predicted value for the i-th instance, fk represents the k-th weak learner added to the ensemble.

6. Naïve Bayes

Naive Bayes (NB) is a supervised machine learning model to solve classification problems. The NB method is a very straightforward and effective technique capable of making accurate predictions. The NB technique is based on probability, which means the classifier makes predictions based on the likelihood of dataset variables. To predict the target class for each record, NB uses the values for the independent variables. The NB model is commonly used in medical research.

$$P(C_k|X) = \frac{P(X|C_k)P(C_k)}{P(X)} \qquad (6)$$

$P(Ck\ |X)$ represents the probability of a sample belonging to class Ck given the input features X. $P(X|Ck\ )$ is the likelihood of observing the input features X.

7. K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a supervised machine learning technique primarily used for classification and regression tasks [14]. The KNN algorithm is non-parametric, which means it doesn't make any assumptions about the underlying data. The KNN algorithm places the new instance into a category comparable to the available classes, assuming that the new and available cases are similar. Most sample information is retrieved using the Euclidean distance metric in KNN.

$$yq = \mathrm{mode} yi1\ , yi2\ , \ldots , yik \qquad (7)$$

where yq represents the predicted label for a given query point, and $i1$, $i2,...,$ ik represent the indices of the k nearest neighbors of the query point.

2.4 Evaluation Metrics

The performance of the model was assessed using the following evaluation metrics to ensure a comprehensive understanding of its accuracy and classification ability:

1. ConfusionMatrix:

A confusion matrix will be used to visualize model predictions, showing the distribution of true positives, true negatives, false positives, and false negatives. This aids in understanding where the model makes correct or incorrect classifications.

2. Accuracy:

Accuracy measures the proportion of total correct predictions over all instances in the dataset. While useful for initial assessment, accuracy alone may not be sufficient if class imbalance is present.

Accuracy=(TP+TN)/(TP+TN+FP+FN) (1)

3. Precision:

Precision evaluates the proportion of true positive predictions out of all positive predictions made by the model. In medical diagnostics, high precision ensures that a positive diagnosis (e.g., presence of CVD) is likely to be correct.

Precision=TP/(TP+FP) (2)

4. Recall(Sensitivity):

Recall measures the ability of the model to correctly identify all actual positive cases. It is especially important in healthcare applications, where failing to detect a disease (false negative) can lead to severe consequences.

Recall=TP/(TP+FN) (3)

5. F1-Score:

The F1-score is the harmonic mean of precision and recall, providing a balanced measure of a model's performance, especially when there is a trade-off between false positives and false negatives.

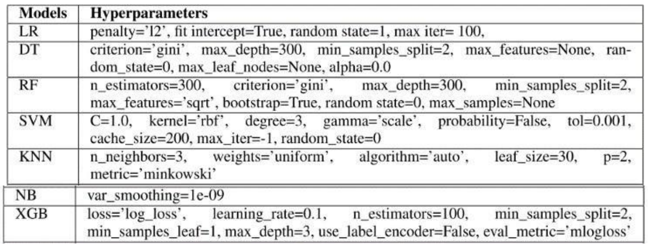
F1=2(Precision·Recall)/(Precision+Recall)
(4)

The goal is to find a model that not only performs well in terms of accuracy but also minimizes critical diagnostic errors, particularly false negatives.

III. RESULT ANALYSIS

This section discusses our proposed research results and scientific validity. The machine algorithms are developed using the python programming language-based skit-learn library module. Our study performance measures are the runtime computation, accuracy, precision, recall, and f1 scores. The performance indicators of our research models are evaluated for scientific results validation.

TABLE 3. The optimal hyperparameters analysis for applied machine learning models is analyzed.



Models	Hyperparameters
LR	penalty='l2', fit intercept=True, random state=1, max iter= 100,
DT	criterion='gini', max_depth=300, min_samples_split=2, max_features=None, random_state=0, max_leaf_nodes=None, alpha=0.0
RF	n_estimators=300, criterion='gini', max_depth=300, min_samples_split=2, max_features='sqrt', bootstrap=True, random state=0, max_samples=None
SVM	C=1.0, kernel='rbf', degree=3, gamma='scale', probability=False, tol=0.001, cache_size=200, max_iter=-1, random_state=0
KNN	n_neighbors=3, weights='uniform', algorithm='auto', leaf_size=30, p=2, metric='minkowski'
NB	var_smoothing=1e-09
XGB	loss='log_loss', learning_rate=0.1, n_estimators=100, min_samples_split=2, min_samples_leaf=1, max_depth=3, use_label_encoder=False, eval_metric='mlogloss'

TABLE 4. The results comparison analysis of the applied machine learning models on test data

Models	Runtime computation(second)	Accuracy%	Precision%	Recall	F1 Score
LR	0.044	82.05%	0.789	0.89	0.84
DT	0.005	98.01%	1.00	0.96	0.98
RM	1.113	98.01%	1.00	0.96	0.98
SVM	0.096	85.03%	0.816	0.92	0.86
NB	0.002	83.49%	0.836	0.84	0.83
KNN	0.005	70.13%	0.712	0.70	0.70
XGB	0.119	99%	1.00	0.98	0.99

Table 4 presents a comprehensive comparison of various machine learning models based on their performance on the test dataset. The evaluation metrics include accuracy, precision, recall, F1-score, and runtime computation.

Among all the models evaluated,**XGBoost (XGB)** classifiers outperformed the others with the highest accuracy scores of **99%.**

precision (**1.00**) and a high recall value of **0.98**, leading to an excellent **F1-score of 0.99**. These results indicate that ensemble-based models are highly effective for this classification task due to their robustness and ability to reduce overfitting.

The **Logistic Regression (LR)** model demonstrated a balanced performance with an accuracy of **82.05%**, precision of **0.789**, and the highest recall value among non-ensemble models (**0.92**). The **Naive Bayes (NB)** scores of **83.4%** and **Decision Tree (DT)** models also performed relatively well with accuracy score **98.01%**, respectively. Despite their relatively lower precision, both models exhibited competitive recall and F1-scores. Importantly, Naive Bayes was the fastest model in terms of execution time (**0.002 seconds**), making it well-suited for applications where computational efficiency is critical.

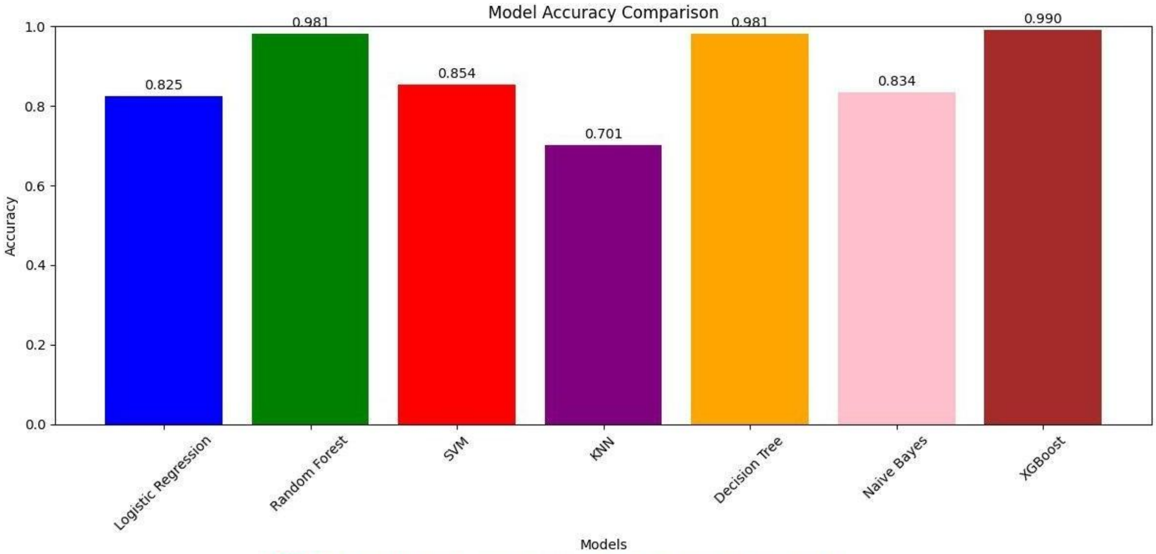


FIGURE 6. The bar chart-based results comparison analysis of the applied machine learning models using the PCHF technique.

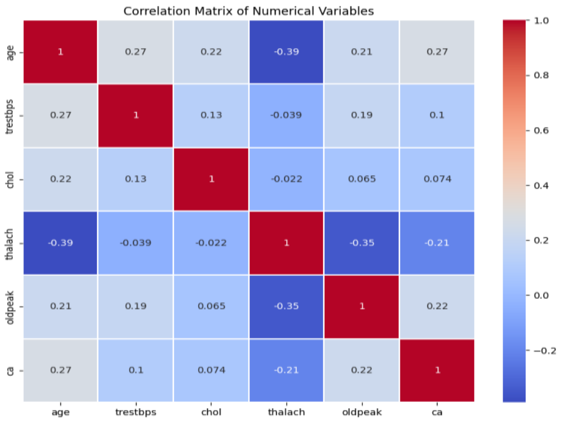
On the other hand, the **Support Vector Machine (SVM)** and **K-Nearest Neighbors (KNN)** models yielded comparatively lower performance metrics. SVM achieved an accuracy of **85.4%** and an F1-score of **0.86**, while KNN obtained an accuracy of **70.13%** and an F1-score of **0.70**. These results suggest that SVM and KNN are less effective for this particular dataset, possibly due to limitations in their capacity to handle complex patterns or class imbalance without parameter tuning.

In terms of runtime performance, the Decision Tree and KNN models were also among the fastest (**0.005 seconds**), while Random Forest required the most time (**1.113 seconds**), which is expected due to the complexity of the ensemble learning process.

In summary, XGBoost demonstrated superior classification performance, albeit at the cost of higher computational time. Logistic Regression offereda strong balance between performance and efficiency. Simpler models like Naive Bayes and Decision Tree proved to be computationally efficient with reasonable accuracy, while SVM and KNN lagged behind in overall effectiveness. These findings highlight the importance of choosing the right model based on application-specific trade-offs between accuracy and computation time.

3.2 Confusion Matrix Analysis

The confusion matrix analysis provides a deeper understanding of each model's classification behavior by highlighting how well the predicted labels match the actual class labels. Models such as Random Forest (RF) and XGBoost (XGB) exhibited nearly perfect confusion matrices, with most values concentrated along the diagonal, indicating accurate predictions across all classes and minimal misclassifications. Logistic Regression (LR) and Naive Bayes (NB) also showed strong performance, though they made some errors in distinguishing between similar classes. The Decision Tree (DT) model achieved a reasonably good distribution but demonstrated occasional misclassifications, particularly in borderline cases. In contrast, Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) showed more dispersed off-diagonal values, suggesting greater difficulty in correctly identifying certain class labels. These misclassifications impacted their overall precision and recall. Overall, the confusion matrix analysis reinforces that ensemble-based models like RF and XGB offer superior and consistent predictive performance, while simpler models may require further tuning or data preprocessing to achieve comparable results.



3.3 Comparison with Existing Machine Learning Models

Ref	Year	Technique	Type	Accuracy%
[5]	2022	Random Forest	Machine learning	96.28
[6]	2020	Random Forest	Machine learning	95.60
[7]	2021	Random Forest	Machine learning	86.60
[9]	2021	Random Forest	Machine learning	88.52
[11]	2019	DT	Machine learning	93.19
[12]	2021	SVM	Machine learning	91.2
[13]	2022	Logistic Regression	Machine learning	85.25
	2025	**Random Forest(proposed)**	Machine learning	**98.02**

proposed RF model outperformed the previously proposed techniques. Our proposed study achieved the highest's scores for heart failure prediction. XGBoost, Random Forest and Decission tree has performed and fully get demonstrated superior classification performance, albeit at the cost of higher computational time. Logistic Regression offered a strong balance between performance and efficiency. Simpler models like Naive Bayes proved to be computationally efficient with reasonable accuracy, while SVM and KNN lagged behind in overall effectiveness. These findings highlight the importance of choosing the right model based on application- specific trade-offs between accuracy and computation time.

IV. CONCLUSION

In this study, machine learning methods were applied to predict heart disease using a dataset comprising 1,025 patient records. A total of eight significant features were selected using a tailored feature engineering approach to improve model performance. Several classification algorithms were implemented and evaluated, including logistic regression, decision tree, support vector machine, random forest, naive Bayes, k-nearest neighbors, and XGBoost. All models were validated using 10-fold cross-validation to ensure robustness and generalization. Among the evaluated models, XGBoost achieved the highest accuracy of 99% with low runtime computation. Confusion matrix analysis revealed strong prediction performance across all classes, highlighting the models' capability to effectively distinguish between different patient conditions. Overall, the proposed approach demonstrates significant potential for early and accurate heart disease detection using machine learning techniques.

References

[1] Ogunpola, A., Saeed, F., Basurra, S., Albarrak, A. M., & Qasem, S. N. (2024). Machine Learning-Based Predictive Models for Detection of Cardiovascular Diseases. *Diagnostics*, *14*(2), 144. https://doi.org/10.3390/diagnostics14020144

[2] Reddy, K. V. V., Elamvazuthi, I., Aziz, A. A., Paramasivam, S., Chua, H. N., &Pranavanand, S. (2021). Heart Disease Risk Prediction Using Machine Learning Classifiers with Attribute Evaluators. *Applied Sciences*, *11*(18), 8352. **https://doi.org/10.3390/app11188352**

[3] Bhowmik, P. K., Miah, M. N. I., Uddin, M. K., Sizan, M. M. H., Pant, L., Islam, M. R., & Gurung, N. (2024). Advancing Heart Disease Prediction through Machine Learning: Techniques and Insights for Improved Cardiovascular Health. *British Journal of Nursing Studies*, *4*(2), 3550.

[4] Qadri, A. M., Raza, A., Munir, K., & Almutairi, M. S. (2023). Effective feature engineering technique for heart disease prediction with machine learning. *IEEE Access*, *11*, 56214-56224.

[5] C. A. Uslan, J. Iqbal, R. Irfan, S. Hussain, A. D. Algarni, S. S. H. Bukhari, N. Alturki, and S. S. Ullah, "Effectively predicting the presence of coronary heart disease using machine learning classifiers," *Sensors*, vol. 22, no. 19, p. 7227, Sep. 2022.

[6] R. Katarya and S. K. Meena, "Machine learning techniques for heart disease prediction: A comparative study and analysis," *Health Technol.*, vol. 11, no. 1, pp. 87–97, Jan. 2021.

[7] P. Rani, R. Kumar, N. M. O. S. Ahmed, and A. Jain, "A decision support system for heart disease prediction based upon machine learning," *J. Reliable Intell. Environ.*, vol. 7, no. 3, pp. 263–275, Sep. 2021.

[8] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE access*, *7*, 81542-81554.

[9] N. L. Fitriyani, M. Syafrudin, G. Alfian, and J. Rhee, "HDPM: An effective heart disease prediction model for a clinical decision support system," *IEEE Access*, vol. 8, pp. 133034–133050, 2020.

[10] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. W. Quinn, and M. A. Moni, ‘‘Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison,’’ *Comput. Biol. Med.*., vol. 136, Sep. 2021, Art. no. 104672.

[11] F. S. Alotaibi, "Implementation of machine learning model to predict heart failure disease," *Int.J. Adv. Comput. Sci. Appl.*, vol. 10, no. 6, pp. 1–8, 2019.

[12] D. K. Plati, E. E. Tripoli, A. Bechloulis, A. Ramnos, I. Dimou, L. Lakkas, C. Watson, K. McDonald, M. Ledwidge, R. Pharthi, J. Gallagher, L. K. Michalis, Y. Goletsis, K. K. Naka, and D. I. Fotiadis, "A machine learning approach for chronic heart failure diagnosis," *Diagnosticos*, vol. 11, no. 10, p. 1863, Oct. 2021.

[13] A. Saboor, M. Usman, S. Ali, A. Samad, M. F. Abrar, and N. Ullah, "A method for improving prediction of human heart disease using machine learning algorithms," *Mobile Inf. Syst.*, vol. 2022, pp. 1–9, Mar. 2022.

[14] A. Raza, K. Munir, M. Almutairi, F. Younas, and M. M. S. Fareed, ‘‘Predicting employee attrition using machine learning approaches,’’ *Appl. Sci.*, vol. 12, no. 13, p. 6424, Jun. 2022.

[15] A. Raza, H. U. R. Siddiqui, K. Munir, M. Almutairi, F. Rustam, and I. Ashraf, ‘‘Ensemble learning-based feature engineering to analyze maternal health during pregnancy and health risk prediction,’’ *PLoS ONE*, vol. 17, no. 11, Nov. 2022, Art. no. e0276525.

[16] N. L. Fitriyani, M. Syafrudin, G. Alfian, and J. Rhee, ‘‘HDPM: An effective heart disease prediction model for a clinical decision support system,’’ *IEEE Access*, vol. 8, pp. 133034–133050, 2020.

[17] R. Katarya and S. K. Meena, ‘‘Machine learning techniques for heart disease prediction: A comparative study and analysis,’’ *Health Technol.*, vol. 11, no. 1, pp. 87–97, Jan. 2021.

[18] heart_dataset from Kaggle (https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset/data)