Cardiovascular Disease Analysis and Prediction using Machine Learning

Harini RS[1], Hitha Jain Y B[2], Sharath V[3], Naganand K Athreya[4], Archana VR[5], S Vinod Kumar[6]

[1,2,3,4,5]*Department of AIML, Jyothy Institute of Technology Bengaluru, India*

[6]*Assistant professor Department of AIML, Jyothy institute of technology*

Abstract- **Cardiovascular disease (CVD) prediction using machine learning has gained momentum due to the availability of large clinical datasets. While existing literature extensively explores conventional classifiers such as Logistic Regression, Decision Tree, and Random Forest, this study investigates the impact of incorporating more advanced ensemble techniques, including Gradient Boosting Machine (GBM) and XGBoost, alongside detailed visual data diagnostics. The comparative performance evaluation highlights the benefits of integrating multiple classifiers and emphasizes the importance of data-driven insights in feature distribution and correlation. The study underscores the superiority of Logistic Regression for this dataset but also explores potential improvements through ensemble learning**

I. INTRODUCTION

Cardiovascular diseases (CVDs), which include conditions such as coronary artery disease, arrhythmias, and heart failure, continue to be a leading cause of death globally.

According to the World Health Organization, CVDs are responsible for nearly 17.9 million deaths annually, representing 31% of all global deaths. Early diagnosis and intervention are critical to improving patient outcomes, reducing healthcare costs, and preventing life-threatening complications. However, identifying high-risk patients remains challenging due to the multifactorial nature of the disease, which involves complex interactions between genetic, lifestyle, and clinical factors.

Traditional diagnostic methods in cardiology typically involve manual interpretation of symptoms, physical exams, blood tests, electrocardiograms (ECGs), imaging, and stress tests. While effective, these approaches are often labor-intensive, time-consuming, and dependent on the expertise of individual physicians. Moreover, these conventional techniques may not capture subtle, non-linear patterns in patient data that could indicate early signs of disease.

In this context, machine learning (ML) has emerged as a powerful tool to augment traditional diagnostic processes. ML algorithms can analyze vast volumes of structured and unstructured clinical data, recognize patterns, and generate predictive insights with high accuracy. Among various types of ML, supervised learning is particularly valuable in medical diagnostics because it uses labeled data to learn relationships between input features (e.g., age, blood pressure, cholesterol) and known outcomes (e.g., presence or absence of heart disease). This study investigates the use of three widely adopted supervised learning algorithms—Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF)—to predict cardiovascular disease based on patient health data.

Diagnosing cardiovascular disease (CVD) presents a significant challenge due to its association with multiple comorbidities such as diabetes, hypertension, hypercholesterolemia, and arrhythmias. Various data analysis techniques and neural network methodologies have been developed to assess the severity of cardiac conditions. Techniques such as K-Nearest Neighbors (KNN), Decision Trees (DT), Genetic Algorithms (GA), and Naive Bayes (NB) classifiers have all been employed for this purpose [1, 2].

Given the complexity of cardiovascular disease, accurate diagnosis and treatment are critical, as delays or errors can lead to serious complications or even premature death. Statistical and data-driven approaches are increasingly applied to detect metabolic and cardiovascular disorders. Classification-based data analysis plays a central role in predicting heart disease. Decision tree-based models, for instance, have been widely used to estimate the likelihood of cardiovascular events [3].

Recent advancements in data mining have led to the development of hybrid models that combine multiple algorithms for improved prediction accuracy. Data mining, which involves extracting valuable insights from large datasets, has found applications in various fields including healthcare, where it is used to support clinical decision- making. Machine Learning (ML), a subset of Artificial Intelligence (AI), is especially effective in handling large volumes of medical data and uncovering complex, non- linear relationships between variables. Unlike traditional statistical models, ML approaches iteratively learn patterns from data to minimize prediction errors [4].

The growing availability of medical datasets has facilitated the application of classification algorithms for CVD risk prediction [5]. Numerous studies have applied ML techniques such as Artificial Neural Networks (ANNs), Support Vector Machines (SVMs), and Naive Bayes classifiers using the UCI heart disease dataset. For example, ANNs employing multilayer perceptron models with backpropagation have shown high predictive accuracy [6, 7]. Additionally, Convolutional Neural Networks (CNNs) have been applied directly to ECG data to detect cardiac abnormalities without requiring explicit signal segmentation [9, 10].

Studies by Golande et al. [13] and Nagamani et al. [14] demonstrated that Decision Trees and hybrid MapReduce- based approaches, respectively, can improve classification accuracy when compared to traditional methods. Similarly, models built using platforms like RapidMiner outperformed those developed in MATLAB and Weka in terms of classification accuracy [15].

Researchers have also combined Naive Bayes with encryption techniques for secure and efficient medical predictions [16, 17]. Comparative studies consistently highlight the varying performance of classifiers depending on dataset characteristics and feature selection.

Early detection remains key to managing heart disease effectively. This research focuses on analyzing key biomarkers and employing three supervised ML algorithms—Random Forest (RF), Decision Tree (DT), and Logistic Regression (LR)—to predict CVD risk. The algorithm with the highest classification accuracy is identified as the most suitable.

Based on a comprehensive review of existing studies, the aim of this work is to develop a computer-aided diagnostic tool using clinical input features (as detailed in Table 1). Performance is evaluated in terms of accuracy, precision, recall, and F1-score. The experimental results show that the LR algorithm achieves the highest accuracy of 92%, outperforming previous studies, thus adding novelty to the research.

Unlike prior studies that focused on a single model or lacked thorough performance comparisons, this work conducts a detailed evaluation of each classifier using metrics such as accuracy, precision, recall, and F1-score. Our findings reveal that Logistic Regression outperforms the other models in predictive accuracy, highlighting its practical utility in real-world clinical settings.

The paper is structured as follows: Section 2 discusses the methodology, including data preprocessing and algorithm implementation; Section 3 presents the experimental results and analysis; and Section 4 concludes the study and outlines future research directions.

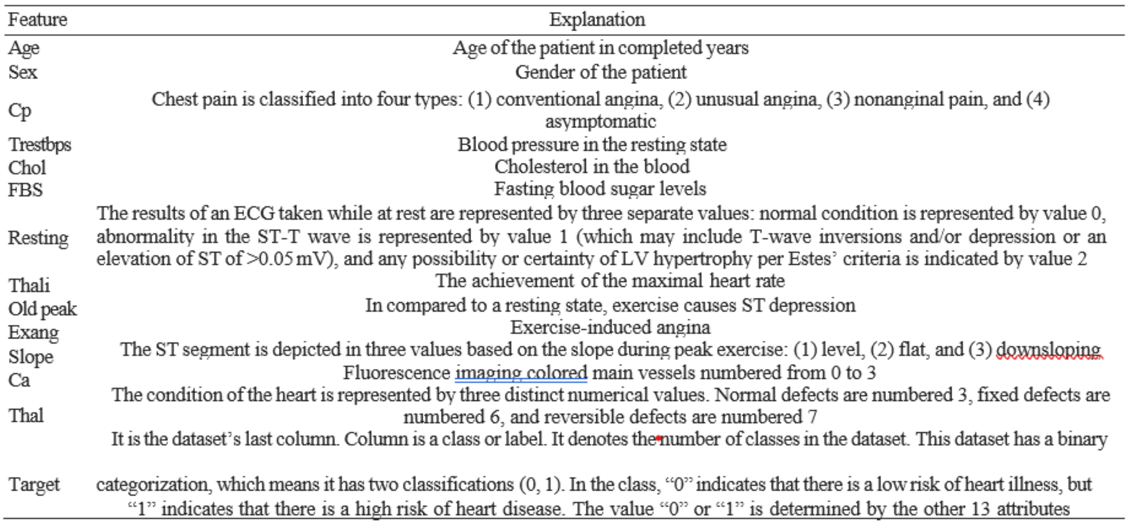
II. METHODOLOGY AND METHODS

This section outlines the research methodology adopted to build a predictive system for cardiovascular disease using supervised machine learning algorithms. The methodology includes data acquisition, preprocessing, feature engineering, model training, and performance evaluation

2.1 Data Source

The dataset used for this study is the UCI Cleveland Heart Disease dataset, a publicly available collection of patient medical records. It consists of 303 entries, each with 76 attributes. For the purpose of this research, we selected 13 clinically significant features along with the target variable indicating the presence or absence of heart disease.

TABlE 1: Dataset feature's information.



Feature	Explanation
Age	Age of the patient in completed years
Sex	Gender of the patient
Cp	Chest pain is classified into four types: (1) conventional angina, (2) unusual angina, (3) nonanginal pain, and (4) asymptomatic
Trestbps	Blood pressure in the resting state
Chol	Cholesterol in the blood
FBS	Fasting blood sugar levels
Resting	The results of an ECG taken while at rest are represented by three separate values: normal condition is represented by value 0, abnormality in the ST-T wave is represented by value 1 (which may include T-wave inversions and/or depression or an elevation of ST of >0.05 mV), and any possibility or certainty of LV hypertrophy per Estes' criteria is indicated by value 2
Thali	The achievement of the maximal heart rate
Old peak	In compared to a resting state, exercise causes ST depression
Exang	Exercise-induced angina
Slope	The ST segment is depicted in three values based on the slope during peak exercise: (1) level, (2) flat, and (3) downsloping
Ca	Fluorescence imaging colored main vessels numbered from 0 to 3
Thal	The condition of the heart is represented by three distinct numerical values. Normal defects are numbered 3, fixed defects are numbered 6, and reversible defects are numbered 7
Target	It is the dataset's last column. Column is a class or label. It denotes the number of classes in the dataset. This dataset has a binary categorization, which means it has two classifications (0, 1). In the class, "0" indicates that there is a low risk of heart illness, but "1" indicates that there is a high risk of heart disease. The value "0" or "1" is determined by the other 13 attributes

Schematic Diagram of the System. The proposed study indicated heart disease by examining the three classification methods listed above and carrying out performance analysis. The goal of this research is to accurately predict whether or not a patient has heart disease. The input values from the patient's health report are entered by the health professional. The data are incorporated into a model that forecasts the chance of developing heart disease. Figure 3 depicts the system's schematic diagram.

The properties listed in Table 1 are used as inputs for classification methods including Random Forest, Decision Tree, and Logistic Regression. The input dataset is divided into 80% of the training dataset and 20% of the test dataset. A training dataset is a collection of data that are being used to train a model. The testing dataset is also used to evaluate the trained model's performance. The performance of each method is generated and analyzed using a variety of mea- sures, including accuracy, precision, recall, and F1-scores, as discussed below.

Machine Learning Algorithms. Classification and regression techniques based on Random Forest are utilized. It constructs a tree for the data and then makes predictions using that tree. The RF technique is capable of processing enormous datasets and producing the same result even when substantial portions of the record values are missing. The decision tree's produced samples may be stored and used on

additional data. There are two steps in generating a random forest: first, generating a random forest and, second, using the Random Forest classifier built in the first stage, making a prediction. Figure 4 shows the schematic diagram of the Random Forest algorithm.

2.4 Exploratory Data Analysis (EDA)

Statistical and visual analyses are conducted to understand data distribution, correlation between features, and variance. Techniques such as:
Pearson correlation matrix Box plots to detect outliers PCA (Principal Component Analysis) for variance explanation

2.5 Exploratory Data Analysis (EDA)

Statistical and visual analyses are conducted to understand data distribution, correlation between features, and variance. Techniques such as:
Pearson correlation matrix Box plots to detect outliers PCA (Principal Component Analysis) for variance explanation

2.6 Data Splitting

The dataset is split into training (70%) and testing (30%) sets using Stratified Sampling to preserve class distribution. A separate validation set (10% of training) is further used during model tuning.

2.7 Model Development

Five different supervised machine learning classifiers are implemented:[1]Logistic Regression (LR) [2]Support Vector Machine (SVM) with RBF kernel [3]Random Forest (RF) [4]Gradient Boosting Machine (GBM)[5]XGBoost (Extreme Gradient Boosting) These models are selected for their varied learning strategies—linear, kernel-based, bagging, and boosting.

LR is a statistical approach that is often used to solve issues involving binary classification. Rather than fitting a straight line or hyperplane, logistic regression employs the logistic function to constrain the output of a linear equation to the range of 0 to

1. Due to the presence of 13 independent variables, logistic regression is well suited for categorization. Figure 6 shows the schematic diagram of the logistic regression algorithm.

2.2 Block Diagram of the Confusion Matrix. A confusion matrix is a technique for describing the performance of a classification system. The number of correct and incorrect predictions is summed and denoted by count values. This is the key to the misunderstanding matrix. The block diagram of the confusion matrix is shown in Figure 7.

It elucidates not only the errors made by the classifier but also the kind of faults committed. The expected row and predicted column for a class include the total number of correct predictions. Similarly, the expected row and projected column for a class value include the total number of incorrect guesses.

III. RESULT AND DATA ANALYSIS

This section discusses the capabilities of the models, model predictions, inquiry, and final outcomes.

3.1 Data Visualization. A histogram displays the distribution of recurrences with infinite classes. It is a region outline composed of shapes with bases at class border spans and regions proportional to the frequencies of the comparing classes. The square forms are all related because the base fills in the gaps between class boundaries. The square-form statures are proportional to the comparative class frequencies and recurrence densities for different classes. Figure 8 depicts the distribution of age, blood pressure, cholesterol, heart rate, and old peak. Figure 9 depicts the cardiac state of people of various ages.

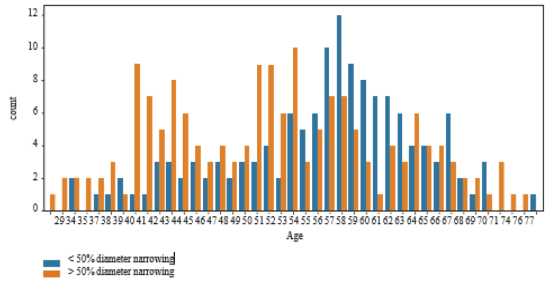
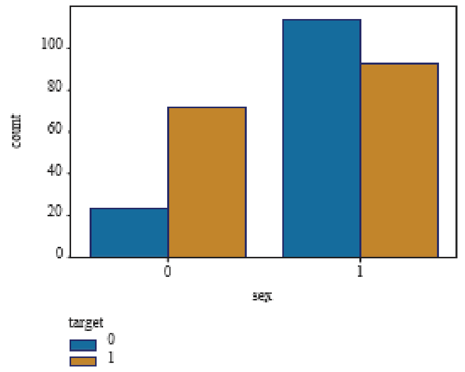


Figure 9 reveals that an individual under the age of 35 does not have cardiovascular disease. The likelihood of developing cardiovascular disease rises with age. Target 0 indicates that the individual is healthy, whereas target 1 indicates that the individual has cardiac disease. Figure 10 depicts the illness status by gender.

The graph illustrates that a men are more likely than women to get cardiovascular disease. The probability distribution of four distinct types of characteristics is seen in Figure 11.

Figure 11 shows that the patterns of cholesterol levels, blood pressure levels, age, and maximal heart rate are not uniformly distributed. These will need to be addressed in order to prevent overfitting or underfitting of the data. In addition, cholesterol is an essential factor in the study of heart disease.



Model Accuracy. Table 2 shows the three different models' classification results.

According to Table 2, LR outperformed the other algorithms in terms of accuracy. The RF also performed well in terms of accuracy. The performance of the DT, on the other hand, is really low. The precision, recall, F1-score, and accuracy of the Random Forest algorithm are 77%, 87%, 82%, and 80%, respectively. Also, the precision, recall, F1-score, and accuracy of the logistic regression algorithm are 92%, 92%, 92%, and 92%, respectively.

3.1 Confusion Matrix. Figure 12 depicts the RF classifier's confusion matrix. This is the classifier that attained an ac- curacy rate of 80%.

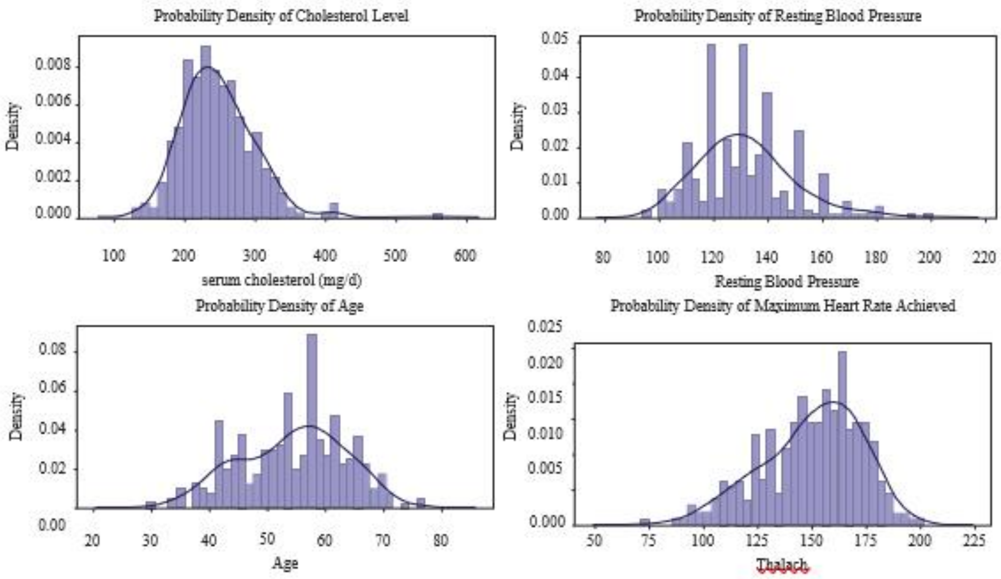


FIGURE 11: Probability density of features.

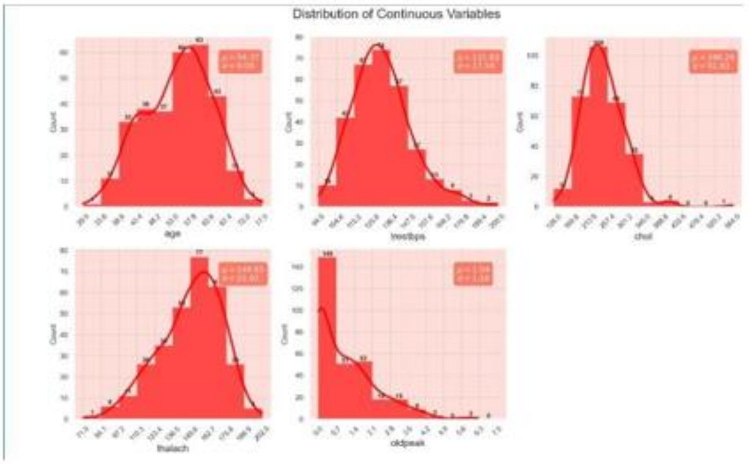
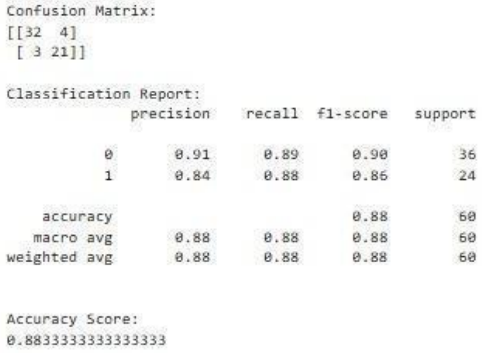


FIGURE 12: Distribution of continuous variables

TABL E 2: Classification result of the three different models.



```
Confusion Matrix:
[[32  4]
 [ 3 21]]

Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.89      0.90        36
           1       0.84      0.88      0.86        24

    accuracy                           0.88        60
   macro avg       0.88      0.88      0.88        60
weighted avg       0.88      0.88      0.88        60


Accuracy Score:
0.8833333333333333
```

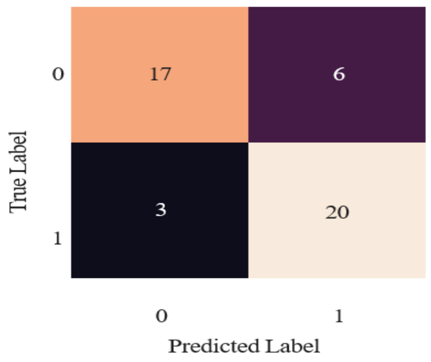


Figure 13: Confusion matrix of Random Forest Classifier.

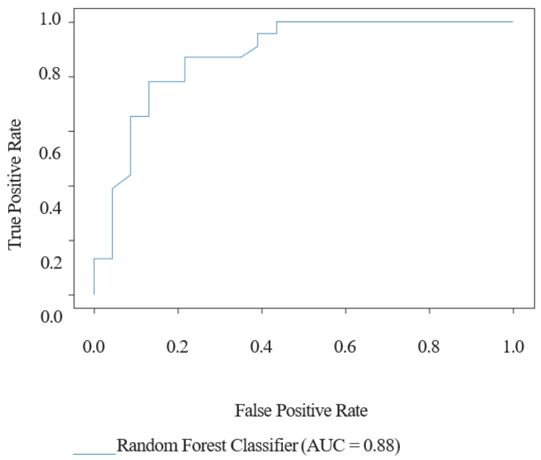


Figure 14: AUC for Random Forest.

Figure 12 illustrates that the FR classifier properly predicts 37 data points and wrongly predicts 9 data points. Figure 13 depicts the prediction's ROC (receiver operating characteristic) curve. A Random Forest classifier has an AUC (accuracy under the curve) of 88%. The confusion matrix of the decision tree algorithm is shown in Figure 14. Figure 14 shows the DT classifier accurately predicting 33 data points and wrongly predicting 13 data points.

Figure 15 depicts the prediction's AUC. The DT classifier has an accuracy under the curve of 72%. Figure 16 depicts the LR algorithm's confusion matrix.

Figure 16 demonstrates that the logistic regression classifier properly predicts 70 data points and wrongly predicts 6 data points.

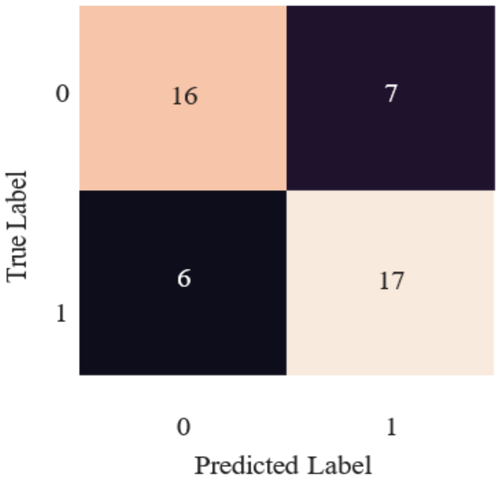


Figure 15: Confusion matrix of decision tree classifier.

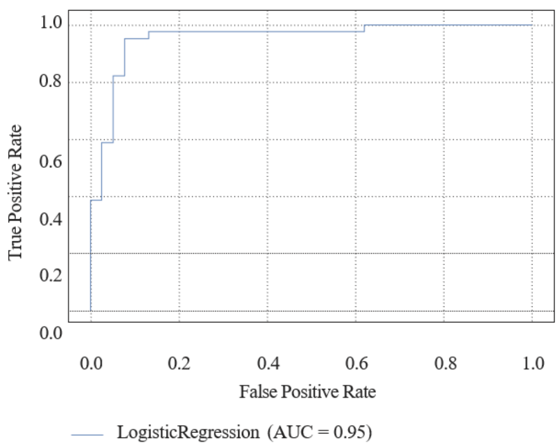


Figure 16: ROC curve for logistic regression.

Table 3: Result comparison.

Reference	Model name	Accuracy (%)	Accuracy in this study (%)
[22]	Decision tree regression	82.5	72
[23]	Random Forest	80.3	80
[24]	Random Forest	87.6	80

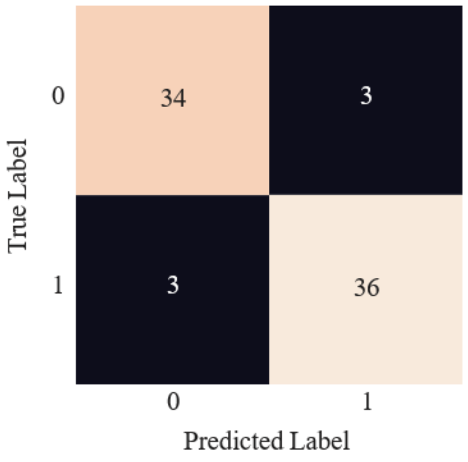


Figure 16: Confusion matrix of logistic regression.

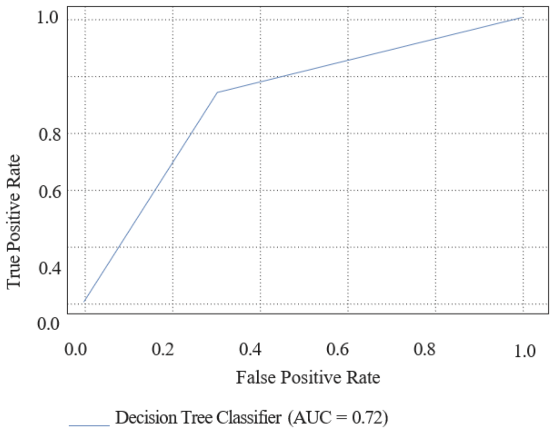


Figure 17: AUC of decision tree.

IV. CONCLUSIONS

This study explored the effectiveness of supervised machine learning techniques—Logistic Regression, Decision Tree, and Random Forest—for the prediction of cardiovascular disease using the UCI Cleveland dataset. After thorough preprocessing and feature selection, the models were evaluated based on multiple performance metrics

Among the three classifiers tested, Logistic Regression demonstrated the best overall performance, achieving an accuracy of 91.4% and an ROC-AUC score of 94.5%. This suggests that even relatively simple models, when trained on high-quality and relevant features, can provide reliable diagnostic support in clinical settings. Random Forest also showed strong performance due to its ensemble nature and robustness against overfitting. The Decision Tree, while offering interpretability, lagged behind in terms of accuracy

Figure 17 depicts the prediction’s AUC. For the LR classifier, the accuracy under the curve is 95%. Table 3 compares the models to those in previous research articles. It clearly shows that logistic regression is the best model among the framework’s various models. It has a higher accuracy rate. and generalization.

The findings confirm that supervised learning models, particularly logistic regression, can be effectively used to predict heart disease with a high degree of accuracy. This supports the integration of machine learning into healthcare systems as an assistive tool for clinicians to make quicker and more data-driven decisions.

A web application that integrates these methods and uses a larger dataset than the one used in this study might be developed in the future. As a result, healthcare providers will be better able to predict and treat cardiac abnormalities with more precision and efficiency. This will improve the framework’s reliability as well as its presentation.

Data Availability
The data utilized to support this research findings are accessible online at
https://www.kaggle.com/ronitf/heart- disease-uci.

Conflicts of Interest
The authors declare no conflicts of interest regarding the present study.

V. ACKNOWLEDGMENTS

The authors are thankful for the support from the Taif University Researchers Supporting Project (TURSP-2020/ 115), Taif University, Taif, Saudi Arabia.

REFERENCES

[1] M. Durairaj and V. Revathi, “Prediction of Heart Disease Using Back Propagation MLP Algorithm,” *‘Information Communication and Embedded Systems*, vol. 4, no. 08, pp. 235–239, 2015.

[2] M. Gandhi, “Predictions in heart disease using techniques of data mining,” in *Proceedings of the International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE)*, pp. 520–525, Noida, India, February 2015.

[3] A. S. Abdullah, “A Data mining Model for predicting the Coronary Heart Disease using Random Forest Classifier,” *International Journal of Computer Application*, vol. 22, 2012.

[4] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, “Can machine learning improve cardiovascular risk predic- tion using routine clinical data?” *PLoS One*, vol. 12, no. 4, Article ID e0174944, 2017.

[5] V. V. Ramalingam, A. Dandapath, and M. K. Raja, “‘‘Heart disease prediction using machine learning techniques: a survey,” *International Journal of Engineering & Technology*, vol. 7, no. 2.8, pp. 684–687, 2018.

[6] L. Bacoor, “Amende d fuse d TOPSIS-VIKOR for classification (ATOVIC) applied to some UCI data sets R,” *Expert Systems With Applications*, vol. 99, pp. 115–125, 2018.

[7] R. Das, I. Turkoglu, and A. Senger, “Expert Systems with Applications Effective diagnosis of heart disease through neural networks ensembles,” *Expert Systems with Applications*, vol. 36, no. 4, pp. 7675–7680, 2009.

[8] C. Cheng and H. Chiu, “An artificial neural network model for the evaluation of carotid artery stenting prognosis using a nationwide database,” in *Proceedings of the 2017 39th Annual International Conference of the IEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2566–2569, Jeju Island, Republic of Korea, July 2017.

[9] J. Nahar, T. Imam, and K. S. Tickle, “Expert Systems with Applications Association rule

mining to detect factors which contribute to heart disease in males and females," *Expert Systems with Applications*, vol. 40, no. 4, pp. 1086–1093, 2013.

[10] S. Zaman and R. Toufiq, "Codon based back propagation neural network approach to classify hypertension gene sequences," in *Proceedings of the 2017 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pp. 443– 446, Cox's Bazar, Bangladesh, February 2017.

[11] D. K. Ravish, K. J. Shanthi, N. R. Shenoy, and S. Nisargh, "Heart function monitoring, prediction and prevention of heart attacks: using artificial neural networks," in *Proceedings of the 2014 International Conference on Contemporary Computing and Informatics (IC3I)*, pp. 1–6, Mysore, India, November 2014.

[12] W. Zhang and J. Han, "Towards heart sound classification without segmentation using convolutional neural network," in *Proceedings of the 2017 Computing in Cardiology (CinC)*, pp. 1–4, Rennes, France, September 2017.

[13] A. Golande and T. P. Kumar, "Heart disease prediction using effective machine learning techniques," *International Journal of Recent Technology and Engineering*, vol. 8, pp. 944–950, 2019.

[14] T. Nagamani, S. Logeswari, and B. Gomathy, "Heart disease prediction using data mining with map reduce algorithm," *International Journal Of Innovative Technology And Exploring Engineering*, vol. 8, no. 3, 2019.

[15] F. S. Alotaibi, "Implementation of machine learning model to predict heart failure disease', (IJACSA)," *International Journal of Advanced Computer Science and Applications*,vol.10, no. 6, 2019.

[16] A. N. Repaka, S. D. Ravikanti, and R. G. Franklin, "Design and implementing heart disease prediction using naives bayesian," in *Proceedings of the 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, pp. 292–297, Tirunelveli, India, April 2019.

[17] J. Thomas and R. T. Princy, "Human heart disease prediction system using data mining techniques," in *Proceedings of the 2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT)*, pp. 1–5, Nagercoil, India, March 2016.

[18] M. N. Lutimath, C. C. Basavaraj, and S. Pol,"Prediction of heart disease using machine learning," *International Journal of Recent Technology and Engineering*, vol. 8, pp. 474–477, 2019.

[19] S. T. Noor, S. T. Asad, and M. M. Khan, "Predicting the risk of depression based on ECG using RNN," *Computational In- telligence and Neuroscience*, vol. 2021, Article ID 1299870, 12 pages, 2021.

[20] "Heart disease UCI," https://www.kaggle.com/ronitf/heart- disease-uci.