Journal of Population Therapeutics & Clinical Pharmacology

RESEARCH ARTICLE
DOI: 10.53555/jptcp.v30i18.3206

A HYBRID ENSEMBLE FRAMEWORK FOR CARDIAC DISEASE RISK STRATIFICATION WITH MACHINE LEARNING

Muhammad Imran[1], Sadaqat Ali[2], Hadi Abdullah[3], Abdul Majid Soomro[4], Muhammad Ahsan Raza[5*], Tahir Abbas[6]

[1,2,6]Department of Computer Science, Times Institute Multan, Pakistan.
Email: m.imran@bzu.edu.pk
[3]Department of Computer Science, Lahore Garrison University, Pakistan.
[4]Department of Computer Science, National College of Business Administration & Economics, Multan, Pakistan.
[5]*Department of Information Sciences, University of Education, Lahore, Multan Campus, Pakistan.
***Corresponding Author:** Muhammad Ahsan Raza
*Department of Information Sciences, University of Education, Lahore, Multan Campus, Pakistan.

Abstract:
Cardiovascular disease is one of the top health concerns to humanity and is gradually increasing daily. Predicting it timely and taking the necessary steps for its intervention is crucial. Precisely predicting cardiac disease is a challenging job that a human or application can do. The complexity of the cardiovascular system compels the use of Artificial Intelligence (AI) to find the solution. Machine learning techniques (sub-set of artificial intelligence) have done tremendous work in medical sciences by providing vast answers to their queries. Computer scientists have used different machine-learning methods for the identification of cardiac disease. This study aims to enhance the accuracy of the prophecy of cardiac disease to reduce the risk factors. It proposes a hybrid ensemble framework to analyze the cardiac data based on essential features for optimum prediction results. This ensemble framework uses multiple machine-learning classification methods to approach the optimal solution. This study uses the Cleveland open access dataset to discuss the working performance of famous classification techniques like Decision Tree, Naive Bayes, SVM, KNN, logistic regression, RF, Gradient Boosting, and XGB Classifier. It proposes a Hybrid Ensemble Framework based on this analysis to enhance the results. The proposed method shows incredible results using the Adaptive Boosting Ensemble technique. AdaBoost is used with hyperparameters on the results retrieved from the applied ML methods and gets more accuracy. The accuracy of this proposed method is evaluated using an open-access Cleveland dataset, which has various cardiac modalities, clinical records, and physiological measurements. Our proposed Hybrid Ensemble Framework achieved an accuracy of 91.80%, precision= 0.94, f1-score=0.92, macro avg= 0.92, and recall = 0.93. The results obtained by the other machine-learning algorithms are less than our model. The comparison of previously completed results is also examined to reflect the improvement in the proposed technique. Moreover, this technique opens new doors for real-world clinical solutions, and it advances the cardiac disease risk stratification field by introducing an innovative and applicable approach by merging ML and ensemble methods. The HEF enhances prediction accuracy and provides valuable insights into the key factors influencing cardiac disease risk, ultimately facilitating more informed clinical decision-making. Our findings underscore the potential of this hybrid ensemble framework as a valuable tool for improving the detection and management of

cardiac diseases, ultimately reducing the burden of CVD (cardiovascular disease) on healthcare systems and society.

Keywords: Heart disease prediction; Machine learning; Ensemble classifier; Hybrid Technique; Decision Tree; Naive Bayes; SVM; KNN; logistic regression; RF; Gradient Boosting; XGB

I-Introduction:

Cardiac disease is the main reason of death not only in developing countries but also in modern nations, presenting a question to world healthcare strategies. In recent years, heart disease and cancer have been the primary causes of death, making up around 43.5% of all deaths during this period. A recent report from the American Heart Association reveals that cardiovascular disease claims more lives annually than all types of cancer and chronic lower respiratory diseases combined [1]. The prediction and consequences of cardiovascular patients have notably improved over the past few years, thanks to innovations in technology and techniques. Notably, machine learning, particularly within the neural network, is rapidly evolving and structuring the processes of cardiology [2,3]. This advancement opens new doors to enhance various aspects of cardiovascular procedures. The rapid adoption of machine learning is driven by the exponential growth of data, particularly in the cardiac field. The most common symptoms of cardiac disease are usually focused on the chest, indicating pain, tightness, pressure, and discomfort. Heart attack is also noticed in specific patients' conditions regarding coldness in arms or legs, numbness, shortness of breath, and weakness if blood circulation is blocked or lessened in those body portions [4].

When discussing risk factors, we mean things that increase the chance of getting sick. If we can fix or remove these factors, we can lower the risk of getting sick. This study focuses on the main risk factors for heart disease, such as age, sex, lifestyle factors (smoking, physical activity, alcohol, and stress), and metabolic syndrome factors (insulin resistance, dyslipidemia, abdominal obesity, high blood pressure, and dietary factors). If people are exposed to these factors, their chances of getting heart disease increase; while removing or improving the elements, we can decrease the risk of heart disease.

The dataset consists of multiple features, and working with all parts may not be helpful and could lead to inaccurate results. Therefore, this study aims to increase the precision of cardiac disease diagnosis using a combination of classification and feature selection. It chooses the most essential features in the dataset, and for data preparation, we divide the dataset into testing and training sets. We then classify the parts using eight machine learning techniques: K-nearest neighbor (KNN), Random Forest, Naïve Bayes, AdaBoost, Decision Tree, Logistic Regression, and Support Vector Machine. By combining these methods, we have improved the accuracy of heart disease diagnosis in different ways. The novelty of this study is that working with these eight machine-learning techniques has never been done before. This proposed method has two main benefits: it reduces the number of features used in datasets and increases the accuracy of diagnosis.

Cardiac disease is responsible for a growing number of deaths worldwide every year. Computer scientists are using machine learning and data mining methods to resolve healthcare industry issues to assist medical professionals in diagnosing and identifying diseases. Machine learning techniques can identify patterns, relationships, and knowledge that may not be possible for ordinary statistical methods [5]. Machine learning is now widely used for analyzing organic compounds, healthcare, and weather forecasting. It is becoming more critical in the healthcare sector and crucial for predictive analysis [6-7]. Machine learning and data mining techniques predict and analyze stroke, diabetes, cancer, and heart disease. These methods are marvelous in diagnosing and predicting cardiovascular issues [8]. Some famous machine learning techniques have shown extraordinary performance, such as KNN (K-Nearest-Neighbor), commonly used in data mining for pattern recognition and classification. The research performed by Paris et al. has shown enhanced results by the voting technique used in the ensemble than the different classifiers [9]. The study also includes the KNN in diagnosing heart disease on a benchmark dataset. This facilitates comparisons with other data mining techniques employed on the same dataset. Additionally, the effectiveness of KNN can be improved by incorporating voting. We aim to provide a framework for predicting heart disease by utilizing diverse machine-learning techniques catering to medical experts and researchers, emphasizing cardiac imaging and modalities. This research also evaluates accuracy loss for each fold to measure the framework's efficiency on benchmark datasets.

The machine-learning algorithms are used to classify the different data types and mainly to diagnose the probability of disease chances [10]. They provide a way to predict the new class of samples through the training dataset [11]. This type of classification is known as supervised classification, which trains itself through labeled data [12-13]. This proposed model uses a classification technique to diagnose heart disease using an objective clinical dataset of cardiac patients.

In this study, we have used eight different linear and non-linear machine learning techniques to predict heart disease and found the best method. The machine learning classifiers that we have used are support vector machine (SVM), logistic regression (LR), k-nearest neighbors (KNN), decision tree (DT), random forest (RF), gradient boosting classifier (GB), naïve Bayesian (NB) and extreme gradient boosting (XGB). The main objectives of this study include:

- To build a system for the prediction of cardiac disease
- To identify the prime attributes of the dataset for the enhancement of results
- To propose a state-of-the-art framework using Ensemble machine-learning techniques

The organization of this research is as follows: section 2 investigates the work done by previous researchers in this area of interest. Section 3 discloses the research methodology in detail, briefly describing how to normalize the data and its preprocessing and analysis. In section 4, results are provided with comparison tables of each used machine learning technique, and in section 5, an intelligent conclusion is given along the way to future work.

II-Literature Review

Computer scientists have been using statistical analysis on the data collected by healthcare professionals. They use different data mining techniques to investigate the issues raised by the clinicians. Medical experts provide the data to scientists for classification and prediction of diseases. Data scientists have done numerous works for the prediction of heart disease. They have identified the risk factors related to heart disease. The most common factors that cause cardio attacks are diabetes, smoking, blood pressure, cholesterol, hypertension, age, and family history of heart disease [14].

Several machine-learning techniques are used to predict cardiovascular disease. Data analysts have applied different machine learning methods to identify heart disease. They have used other datasets for the diagnosis of cardiac disease. The prediction results of the various researchers cannot be equated because they have used different datasets and techniques. However, with time, standard datasets have been focused on and investigated with other machine-learning techniques. Computer scientists have used different modalities and features in available datasets to improve results. They have modified the preprocessing techniques and ways to refine the datasets [15]. As stated in [16], they have chosen thirteen features and three strategies for predicting heart disease. In this study, they have used artificial neural network (ANN), Multivariate Adaptive Regression (MAR), and Logistic Regression (LR) methods to build the hybrid framework for the prediction of heart disease. The result is improved by using Pearson correlation coefficients for the missing values in the Cleveland dataset [17]. Genetic algorithm is also used for heart disease prediction along with other machine learning algorithms and obtained an accuracy of 89% on the dataset of fifty patients [18]. Data analysis techniques work in different ways. Most of them focus on the neighbor’s field points and make the classification based on similarity [19-20]. This study [21] explores the linear regression technique and obtained 87.1% accuracy. R. Perumal et al. [22] used the linear regression and SVM machine learning models on Cleveland datasets and secured 87% accuracy. Some authors have used feature selection models to enhance the accuracy of results [23]. Orange and Weka data analysis tools are also used to find the prediction results [24]. Mohan et al. [25] combined linear regression and random forest to improve the results. In research [26], scientists used decision trees, gaussian NB, and linear regression models to get an accuracy of 82.75%. In another study, particle swarm optimization adopted selected features [27].

Table 1: *ML Techniques and Results on Cleveland Dataset*

SR#	RESEARCH YEAR/REFERENCE	ML-TECHNIQUES	RESULT ACCURACY
1	2007 [37]	Fuzzy-AIRS, KNN	87.00
2	2009 [38]	Bagging Algo	81.41
3	2009 [39]	Neural Network Ensembles	89.01
4	2011 [40]	Decision Tree	84.10
5	2019 [30]	PART, NB Net, C4.5, MLP, Bagging, Boosting, stacking	85.48
6	2019 [34]	HRLFM	88.4
7	2019 [33]	NB, RF, MLP, BN	85.48
8	2019 [35]	AES, NB	89.77
9	2020 [29]	KNN, LR, SVM	87.00
10	2020 [31]	Gaussian Naïve Bayes, DT, LR	82.75
11	2023 [32]	SVM, DT, KNN, RF, LR, XGB	87.91

III- Methodology

A – Dataset and Preprocessing

This study has used the Cleveland standard dataset for heart disease prediction of the Cleveland Clinic Foundation [28]. This benchmark dataset has 76 attributes that store patient information of different types. However, most analysts have used only 13 out of 76 features. The reason behind selecting specific points is that these are the key attributes through which the existence of heart disease can be predicted. It has 303 records of cardiac patients represented by the Target column, containing 0 for absence of heart disease and 1 for presence. The age field includes information on a patient’s age; gender is represented by sex attributes, representing 1 for males and 0 for females. Chest pain attribute CP has four values: two values for fasting blood sugar FBS, resting electrocardiogram RESTECG has 3 classes, and exercise angina EXANG has two values. ST represents a slope that has 3 values. The other four fields have numerical input those are resting bp TRESTBPS, cholesterol CHOL, OLDPEAK, and AGE as shown in figure 1.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	37	1	3.0	130	250	0.0	0	187	0	3.5	3	0	3	0
1	41	0	2.0	130	204	0.0	2	172	0	1.4	1	0	3	0
2	56	1	2.0	120	236	0.0	0	178	0	0.8	1	0	3	0
3	57	0	4.0	120	354	0.0	0	163	1	0.6	1	0	3	0
4	57	1	4.0	140	192	0.0	0	148	0	0.4	2	0	6	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
298	77	1	4.0	125	304	0.0	2	162	1	0.0	1	3	3	1
299	64	1	4.0	145	212	0.0	2	132	0	2.0	2	2	6	1
300	38	1	1.0	120	231	0.0	0	182	1	3.8	2	0	7	1
301	61	1	4.0	138	166	0.0	2	125	1	3.6	2	1	3	1
302	58	1	4.0	114	318	0.0	1	140	0	4.4	3	3	6	1

303 rows × 14 columns

Figure 1: *Selected Columns of Dataset*

After uploading the CSV file of the Cleveland dataset to Google Colab, we normalized it by removing missing and duplicate values. Preprocessing of the dataset is essential for better performance of machine-learning techniques, so we separated the categorical and numerical values. Categorical data is converted to numerical values as shown in figure 2.

```
['sex', 'cp', 'fbs', 'restecg', 'exang', 'slope', 'ca', 'thal', 'target']
[3.  2.  4.  1.  1.1]
   age  sex  trestbps  chol  thalach  oldpeak  target  cp_1.1  cp_2.0  cp_3.0  \
0   37    1       130   250      187      3.5       0       0       0       1
1   41    0       130   204      172      1.4       0       0       1       0
2   56    1       120   236      178      0.8       0       0       1       0
3   57    0       120   354      163      0.6       0       0       0       0
4   57    1       140   192      148      0.4       0       0       0       0

   ...  slope_2  slope_3  ca_0.1  ca_1  ca_2  ca_3  ca_?  thal_6  thal_7  \
0  ...        0        1       0     0     0     0     0       0       0
1  ...        0        0       0     0     0     0     0       0       0
2  ...        0        0       0     0     0     0     0       0       0
3  ...        0        0       0     0     0     0     0       0       0
4  ...        1        0       0     0     0     0     0       1       0

   thal_?
0       0
1       0
2       0
3       0
4       0

[5 rows x 26 columns]
```

Figure 2 *Encoding of Categorical Data*

The assessment of the classification method is reflected through different performance indices, which consist of precision, sensitivity, accuracy, and F1 score. Accuracy is the reflection of the overall performance of the machine-learning model. The model's finding capability of the target class is measured by the Recall metric whether a model can search all objects in the target class or not. Precision measures the predicting capability of the machine-learning model. F-Score is the harmonic mean between sensitivity and precision while Sensitivity describes the ratio of the number of accurate positive predictions to the total actual positive instances. Mathematically, these attributes are represented by the following expressions:

$$Classification\ Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (1)$$

$$Precision = \frac{TP}{TP+FP} \qquad (2)$$

$$Sensitivity = \frac{TP}{TP+FN} \qquad (3)$$

$$F1 = \frac{2*Precision*Sensitivity}{Precision+Sensitivity} = \frac{2TP}{2TP+FP+FN} \qquad (4)$$

Where FN, TN, FP, and TP represent the false negative, true negative, false positive, and true positive respectively.

After the conversion of categorical data, feature scaling is performed on the numerical data using the standard scale to balance the features between certain ranges. The next step taken is to split the dataset into training and testing. The division of the dataset is 80% for training and 20% for testing purposes. The chosen eight classifiers were applied to the dataset and found the optimum accuracy after taking certain hyperparameters as shown in Figure 3.

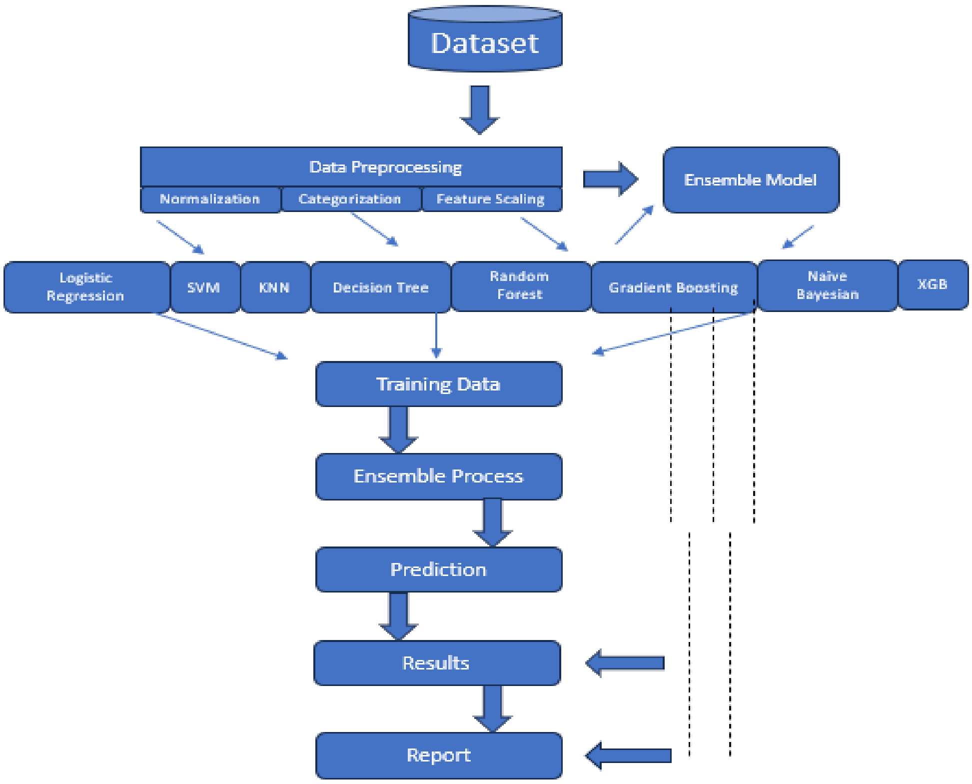

***Figure 3:** The Proposed HEF Model for Predicting Heart Disease*

B- Machine Learning Techniques

In this study, we have used 8 machine-learning techniques consisting of Decision Tree, Naive Bayes, SVM, KNN, logistic regression, RF, Gradient Boosting, and XGB. We observed the working of each machine-learning classification method and by updating their hyperparameters found the best results. By observing their results, a relationship among them is built to obtain more accurate and precise predictions of heart disease. As shown in the result table, it became possible to propose a framework that has given more accurate and enhanced results. We have applied the AdaBoost Ensemble classifier to enhance the accuracy of the prediction model. The working and results of each used machine-learning classifier are discussed below.

1. Logistic Regression Classifier

This classifier predicts the class based on observations (scores) whether the given input belongs to a certain type of class or not. The Logistic Regression classifier converts the probabilities into binary (0,1) for the prediction of classes [41]. We have used the Skit Learn library to take the LR object. We got an accuracy of 88.52% after training its object and then finding predictions on test data as shown in the overall result table. The working performance of the model is represented by multiple metrics like ROC (Receiver Operating Characteristic), PRC (Precision-Recall Curve), and FIG (Feature Importance Graph), CM (Confusion Metric) as shown in Figure 3, and its classification report is reflected in Table 2.

***Table 2:** Classification Report of Logistic Regression*

ATTRIBUTE	PRECISION	RECALL	F1-SCORE	SUPPORT
CLASS 0	0.93	0.84	0.89	32
CLASS 1	0.84	0.93	0.89	29
ACCURACY			0.89	61
MACRO AVG	0.89	0.89	0.89	61
WEIGHTED AVG	0.89	0.89	0.89	61

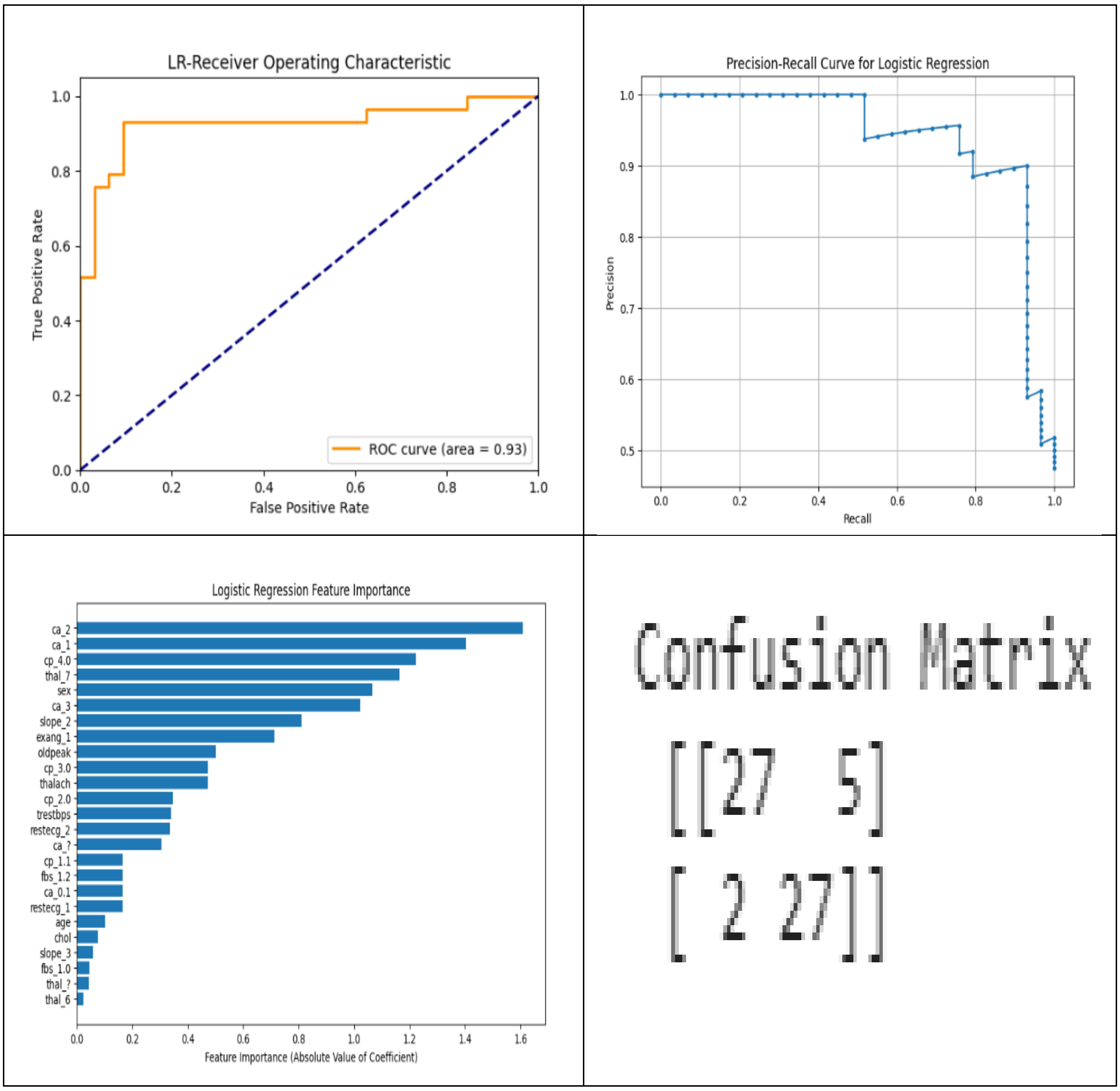

Figure 4: *LR ROC, PR, FI-Graph, Confusion Matrix*

2. Support Vector Machine Classifier

SVM classifier is used to perform both regression and classification tasks. After converting each data item to numerical format, it performs and relates them to certain coordinates. The classification task is then applied by suitable hyper-plane parameters [42]. We have used the Skit Learn library to take the SVM object. We got an accuracy of 81.96% after training its object and then finding predictions on test data, as shown in the overall result table. The working performance of the model is represented by multiple metrics like ROC (Receiver Operating Characteristic), PRC (Precision-Recall Curve), and FIG (Feature Importance Graph), CM (Confusion Metric), as shown in Figure 4-B, and its classification report is reflected in Table 3.

Table 3: *Classification Report of Support Vector Machine*

ATTRIBUTE	PRECISION	RECALL	F1-SCORE	SUPPORT
CLASS 0	0.86	0.78	0.82	32
CLASS 1	0.78	0.86	0.82	29
ACCURACY			0.82	61
MACRO AVG	0.82	0.82	0.82	61
WEIGHTED AVG	0.82	0.82	0.82	61

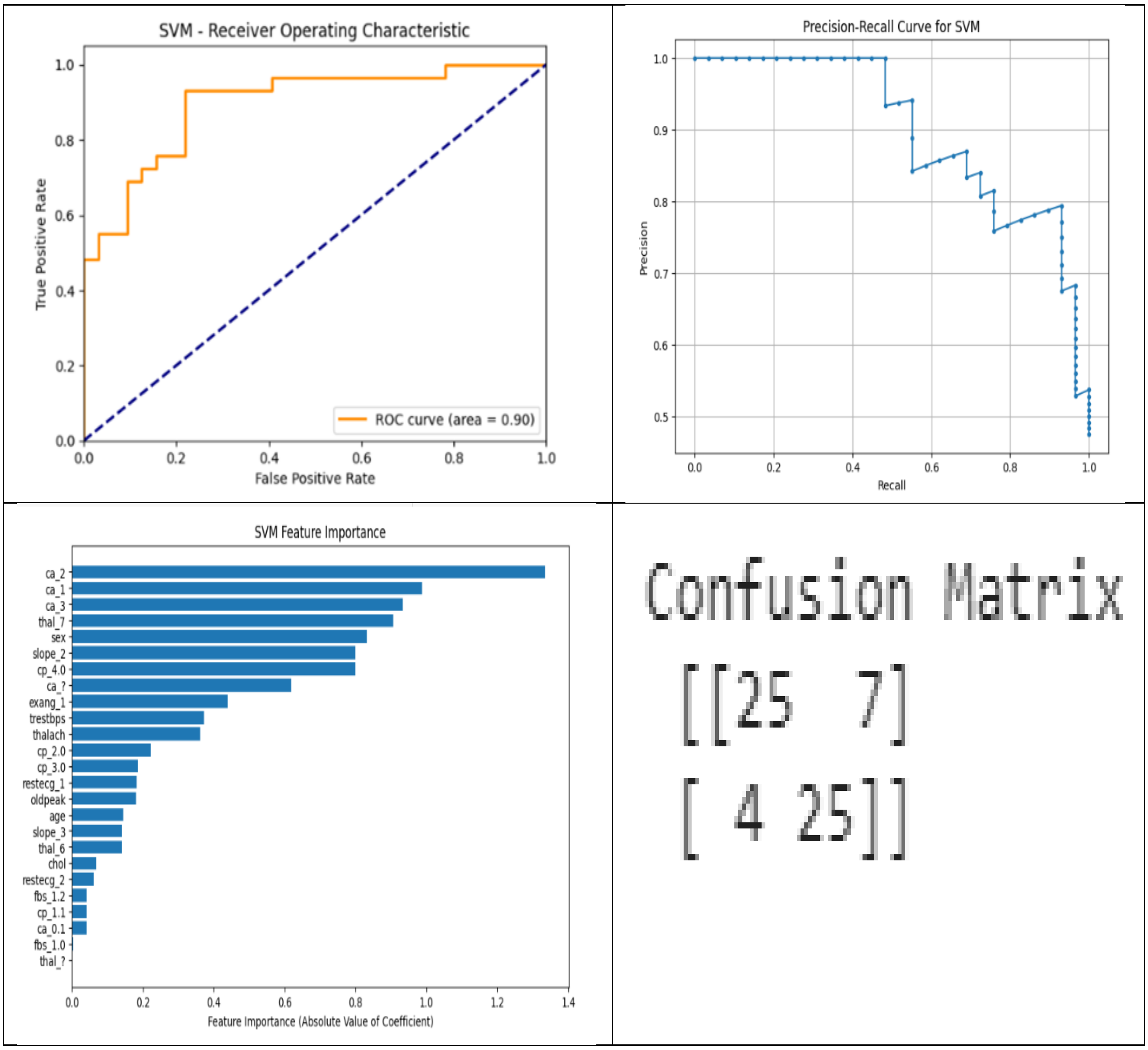

Figure 4-B: *SVM ROC, PR, FI-Graph, and Confusion Matrix*

3. K-Nearest Neighbour Classifier

KNN is one of the most used machine learning techniques for data prediction. It is a regression and classification method used for the analysis of different types of data. This model measures the resemblance technique, and new points are discovered. In our study, KNN has shown an accuracy of 83.60% against the parameter k=9. The working performance of the model is represented by multiple metrics like ROC (Receiver Operating Characteristic), PRC (Precision-Recall Curve), and FIG (Feature Importance Graph), CM (Confusion Metric), as shown in Figure 5, and its classification report is reflected in Table 4.

Table 4: *Classification Report of K-Nearest Neighbors*

ATTRIBUTE	PRECISION	RECALL	F1-SCORE	SUPPORT
CLASS 0	0.86	0.78	0.82	32
CLASS 1	0.78	0.86	0.82	29
ACCURACY			0.82	61
MACRO AVG	0.82	0.82	0.82	61
WEIGHTED AVG	0.82	0.82	0.82	61

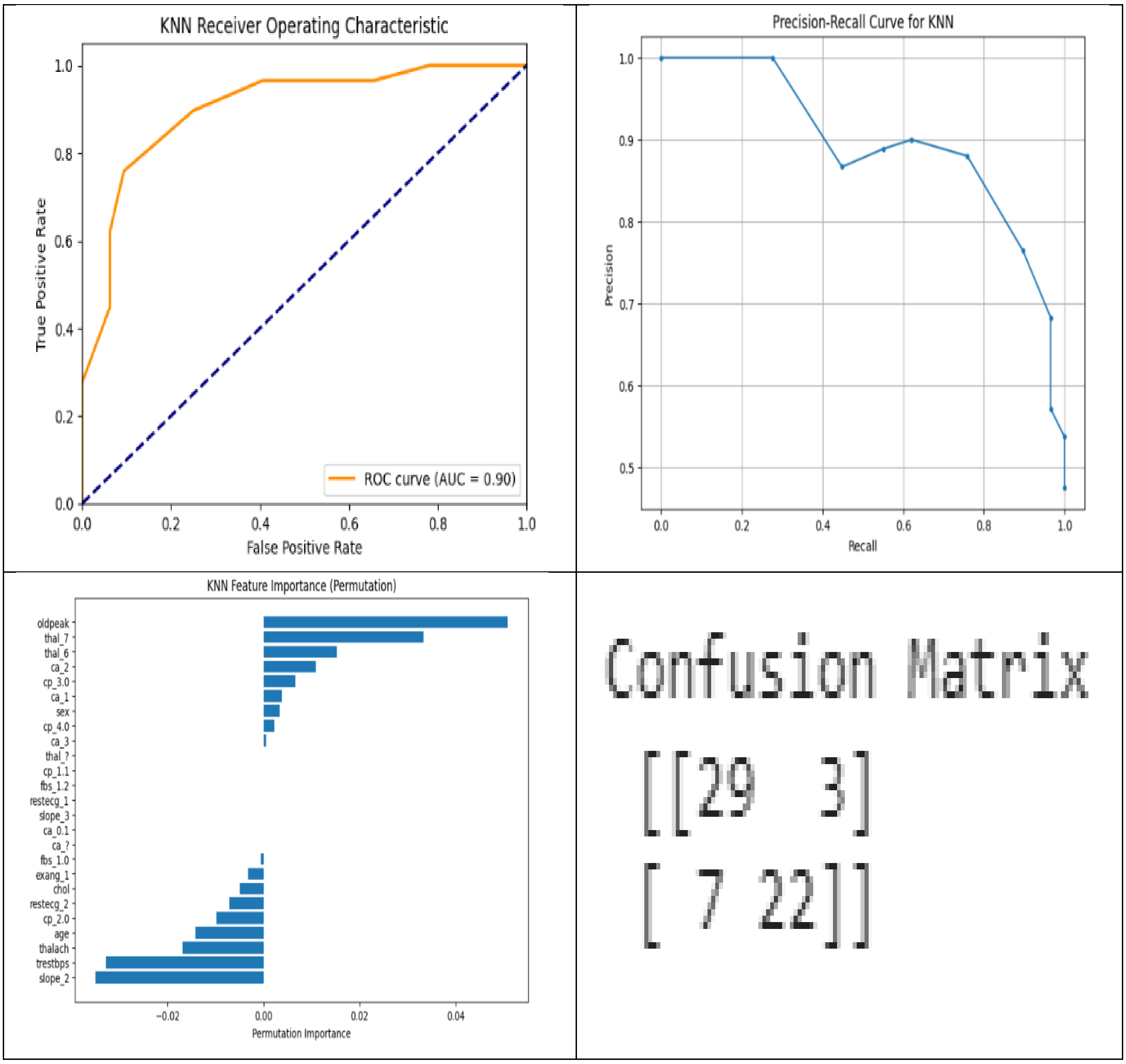

Figure 5: *KNN ROC, PR, FI-Graph, and Confusion Matrix*

4. Decision Tree Classifier

This classifier has a Tree shape-like diagram representing the attributes as internal nodes. It has shown 81.96% accuracy using the Grid Search cross-validation technique with hyperparameters cv=10 and scoring='accuracy'. The working performance of the model is represented by multiple metrics like ROC (Receiver Operating Characteristic), PRC (Precision-Recall Curve), and FIG (Feature Importance Graph), CM (Confusion Metric), as shown in Figure 6, and its classification report is reflected in Table 5.

Table 5: *Classification Report of Decision Tree*

ATTRIBUTE	PRECISION	RECALL	F1-SCORE	SUPPORT
CLASS 0	0.82	0.84	0.83	32
CLASS 1	0.82	0.79	0.81	29
ACCURACY			0.82	61
MACRO AVG	0.82	0.82	0.82	61
WEIGHTED AVG	0.82	0.82	0.82	61

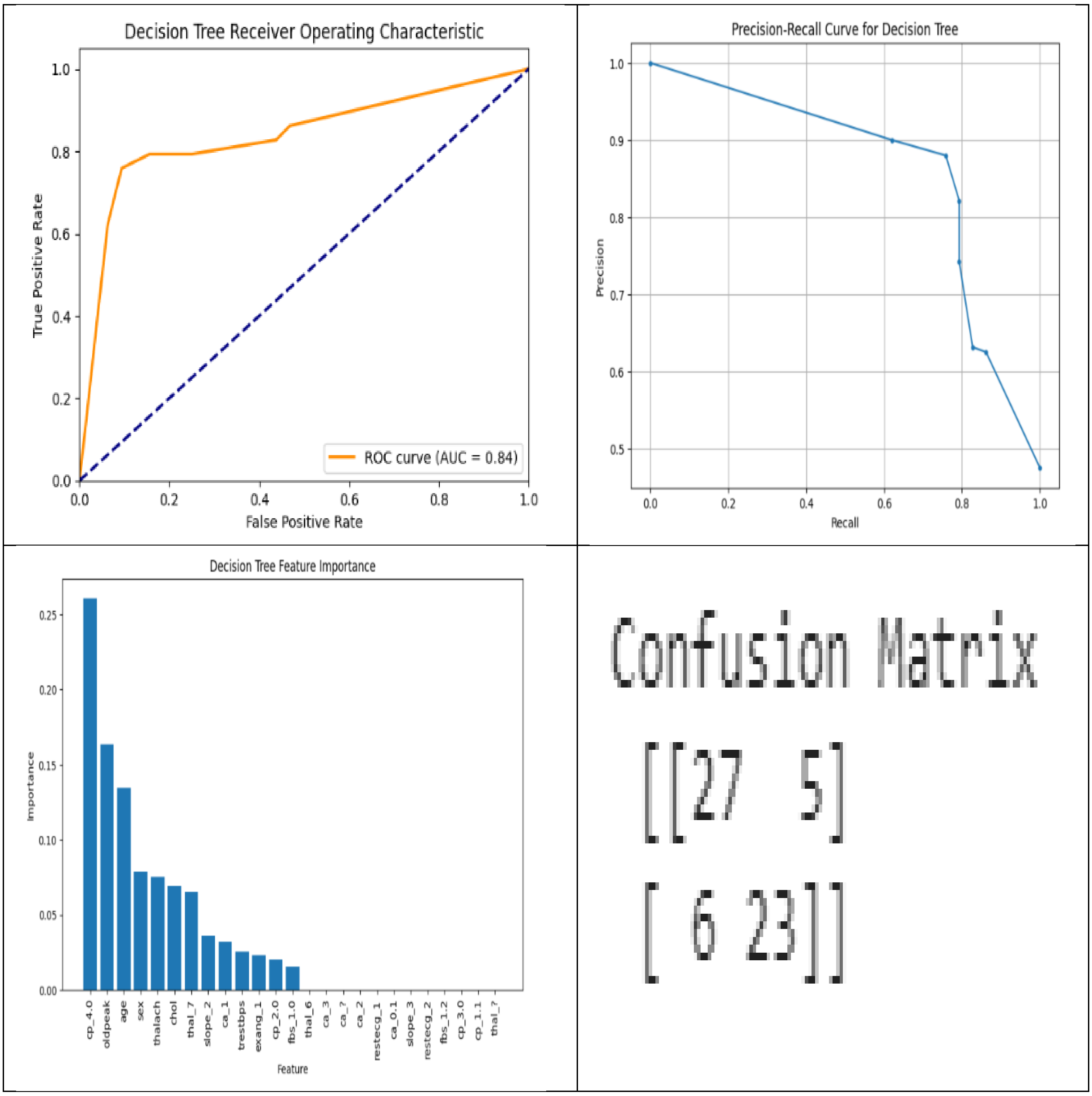

Figure 6: *Decision Tree ROC, PR, FI-Graph, and Confusion Matrix*

5. Random Forest Classifier

The accuracy result measured by using the RF classifier is 80.32%. It is one of the most robust supervised classification techniques mainly used for prediction. It builds a logical forest of trees using a given dataset instead of a single tree. The RF classifier ensembles all trees' output to produce more accurate results. The working performance of the model is represented by multiple metrics like ROC (Receiver Operating Characteristic), PRC (Precision-Recall Curve), and FIG (Feature Importance Graph), CM (Confusion Metric), as shown in Figure 7, and its classification report is reflected in Table 6.

Table 6: *Classification Report of Random Forest*

ATTRIBUTE	PRECISION	RECALL	F1-SCORE	SUPPORT
CLASS 0	0.79	0.84	0.82	32
CLASS 1	0.81	0.76	0.79	29
ACCURACY			0.80	61
MACRO AVG	0.80	0.80	0.80	61
WEIGHTED AVG	0.80	0.80	0.80	61

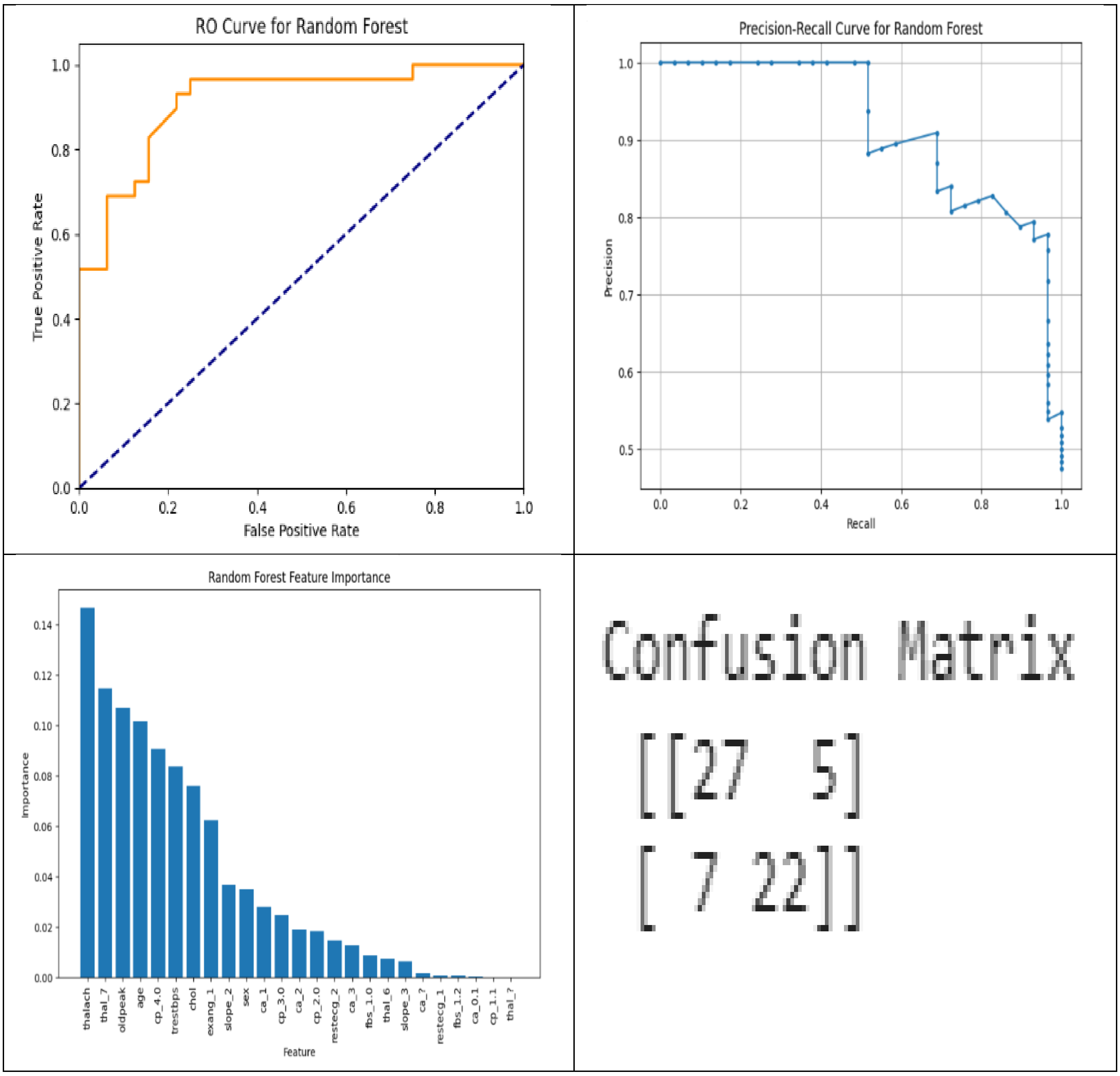

Figure 7: *Random Forests ROC, PR, FI-Graph, and Confusion Matrix*

6. Gradient Boosting Classifier

Like other ML methods, this machine learning technique is also used for regression and classification problems. The GB classifier constructs a stage-level decision tree structure to generate a prediction model. Decision Trees are usually preferred for the boosting purpose of gradient classifiers. In our model, the accuracy of GB was measured at 78.68. The working performance of the model is represented by multiple metrics like ROC (Receiver Operating Characteristic), PRC (Precision-Recall Curve), and CM (Confusion Metric), as shown in Figure 8, and its classification report is reflected in Table 7.

Table 7: *Classification Report of Gradient Booster*

ATTRIBUTE	PRECISION	RECALL	F1-SCORE	SUPPORT
CLASS 0	0.79	0.81	0.80	32
CLASS 1	0.79	0.76	0.77	29
ACCURACY			0.79	61
MACRO AVG	0.79	0.79	0.79	61
WEIGHTED AVG	0.79	0.79	0.79	61

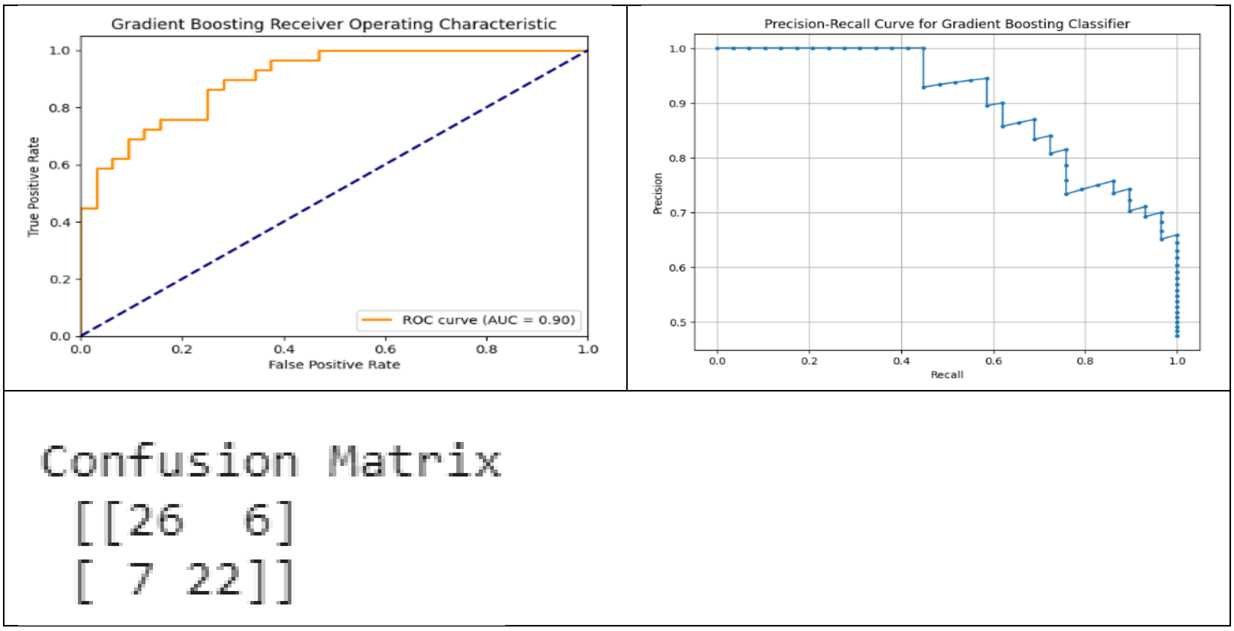

Figure 8: *Gradient Boosting ROC, PR, and Confusion Matrix*

7. Naïve Bayes

NB classifier is another machine learning algorithm used to predict a target class. It is a Bayesian theorem that uses probabilities for calculation. It calculates the probabilities for each attribute in a dataset and adopts the BOW (bag of words) technique to predict the class. We found the accuracy for this model is 85.24. The working performance of the model is represented by multiple metrics like ROC (Receiver Operating Characteristic), PRC (Precision-Recall Curve), and CM (Confusion Metric), as shown in Figure 9, and its classification report is reflected in Table 8.

Table 8: *Classification Report of Naive Bayes*

ATTRIBUTE	PRECISION	RECALL	F1-SCORE	SUPPORT
CLASS 0	0.90	0.81	0.85	32
CLASS 1	0.81	0.90	0.85	29
ACCURACY			0.85	61
MACRO AVG	0.85	0.85	0.85	61
WEIGHTED AVG	0.86	0.85	0.85	61

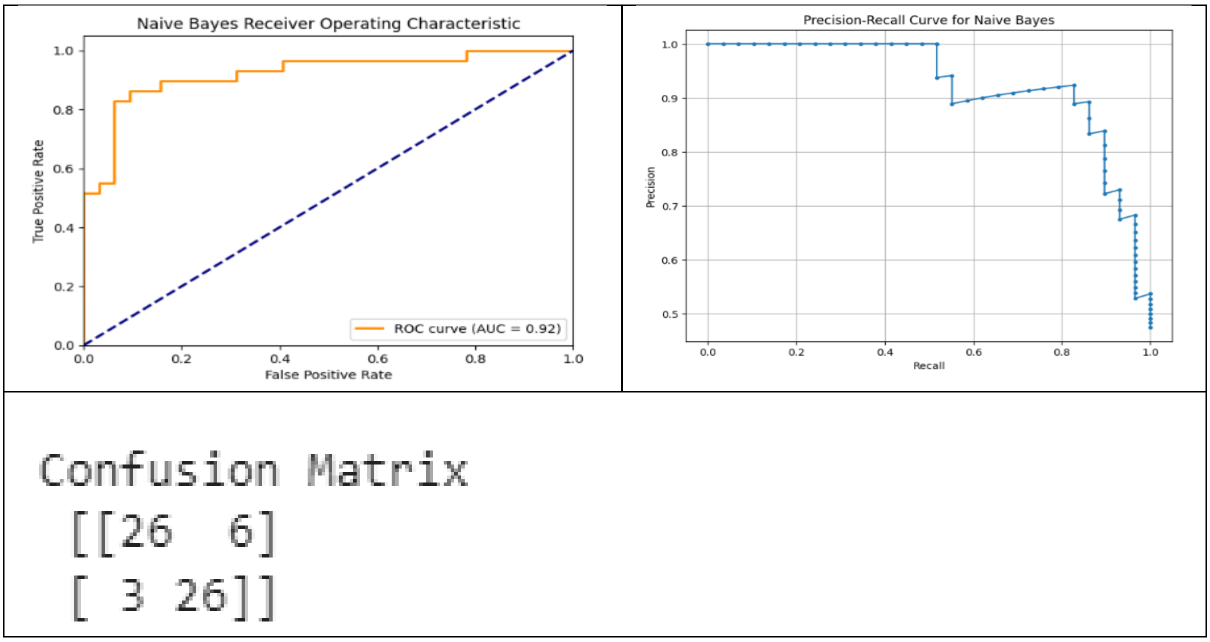

Figure 9: *Naïve Bayes ROC, PR, and Confusion Matrix*

8. Extreme Boosting Classifier

This technique belongs to the ensemble classifier family used to improve weak classifiers' working. XGB (Extreme Gradient Boosting) and GB (Gradient Boosting) both use the gradient descent algorithm to enhance the working of poor learners. In our model, XGB has shown 81.96% accuracy by adopting a hyperparameter grid of Randomized Search with cv=9 as a parameter. The working performance of the model is represented by multiple metrics like ROC (Receiver Operating Characteristic), PRC (Precision-Recall Curve), and CM (Confusion Metric), as shown in Figure 10, and its classification report is reflected in Table 9.

***Table 9:** Classification Report of XG Boost*

ATTRIBUTE	PRECISION	RECALL	F1-SCORE	SUPPORT
CLASS 0	0.76	0.88	0.81	32
CLASS 1	0.83	0.69	0.75	29
ACCURACY			0.79	61
MACRO AVG	0.80	0.78	0.78	61
WEIGHTED AVG	0.79	0.79	0.78	61

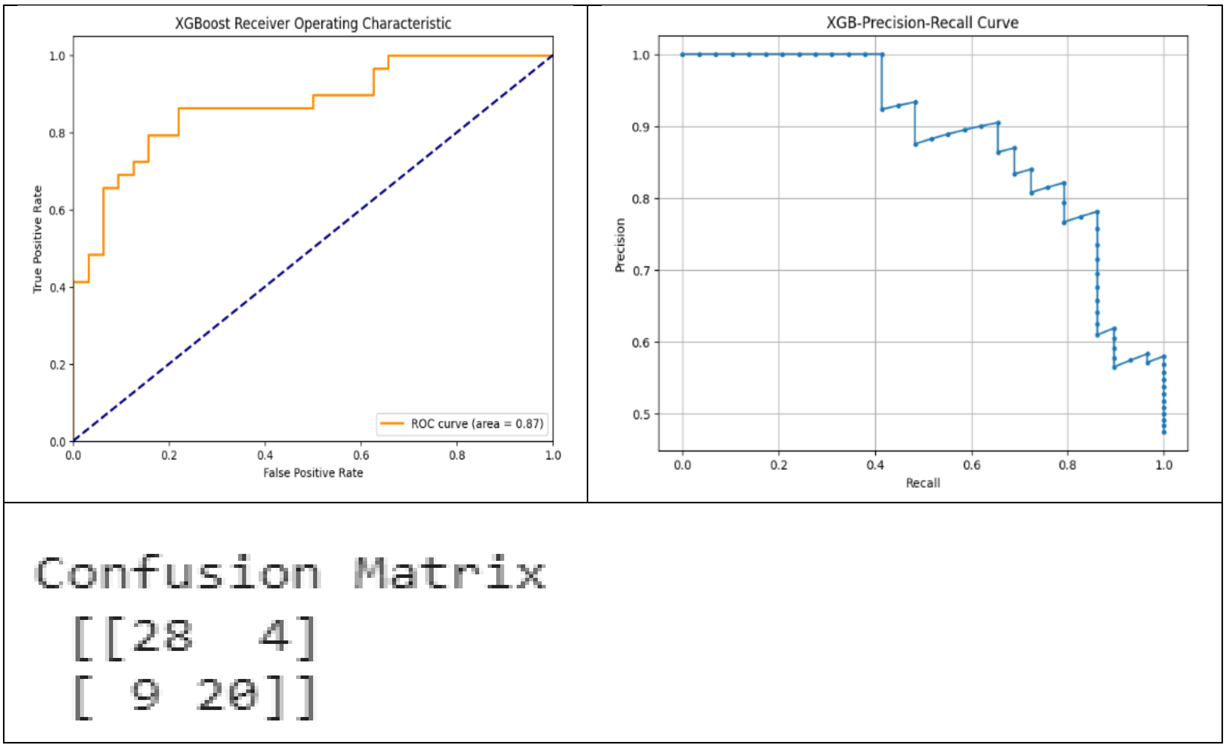

***Figure 10:** XG Boost ROC, PR, and Confusion Matrix*

IV. Results

This study aims to predict cardiovascular disease using machine learning techniques. In this research, we have proposed a Hybrid Ensemble Framework (HEF) that has shown the best results among other machine-learning techniques, as shown in Table 10. We have used 8 machine-learning classifiers on the open-access dataset of Cleveland, and our HEF method has demonstrated its superiority over others. We have used an Adaptive Booting ensemble classifier in conjunction with Logistic Regression to improve the accuracy. Our framework has achieved an accuracy of 91.80%, surpassing all other ML models. We have used the scikit-learn library to import the classifiers to predict heart disease. Logistic Regression approached the maximum accuracy of 88.52% among eight models. Naïve Bayes followed the LR and reached 85.24%, while the Support Vector Machine, K-Neighbor Network, Decision Tree, Random Forest, Gradient Booster, and Extreme Gradient Booster achieved accuracy between 78.68% and 83.60%. We used Logistic Regression as a base classifier in the AdaBoost Ensemble classifier with specific hyperparameters to attain maximum occurs n_estimators =158 and learning_rate =1.0. The efficiency of our HEF framework is reflected in Figure 11, which shows the coefficients of Logistic Regression classifiers in AdaBoost. The notable margin in accuracy between HEF and other ML techniques makes it more worthy for real-world applications, especially for predicting cardiac disease.

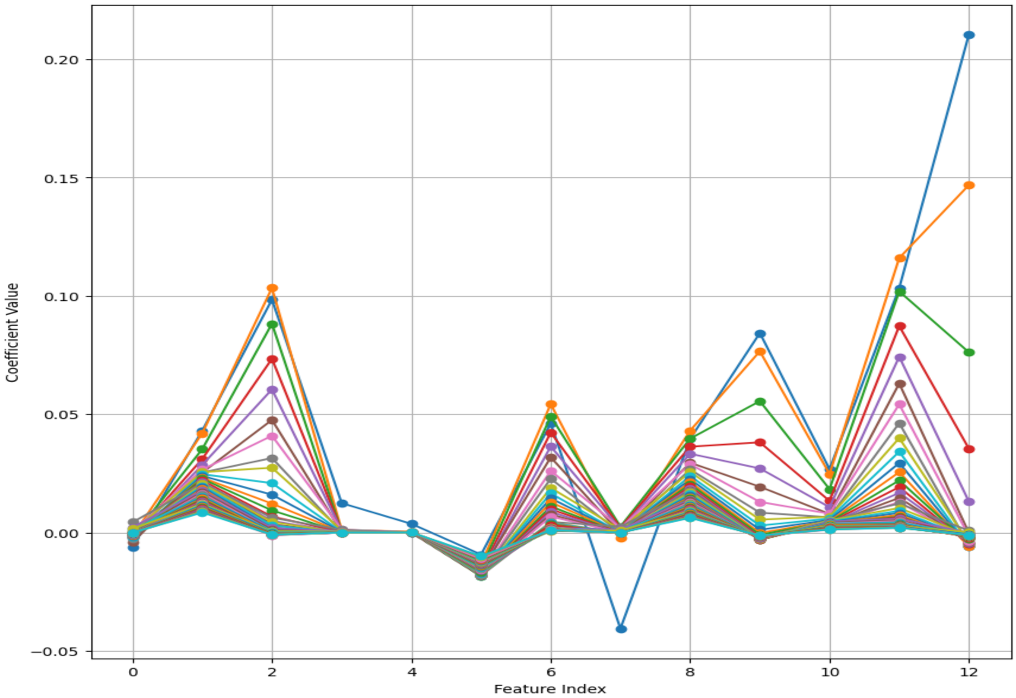

Figure 11: *Coefficients of Logistic Regression Classifiers in AdaBoost*

The accuracy of each used model is shown below in Table 10 to reflect the noticeable difference between the proposed and existing models.

Table 10: *Enhanced Accuracy Results of ML Techniques*

SR#	MACHINE-LEARNING CLASSIFIER	ACCURACY
1	Logistic Regression	88.52
2	Support Vector Machine	81.96
3	K-Neighbor Network	83.60
4	Decision Tree	81.96
5	Random Forest	80.32
6	Gradient Booster	78.68
7	Naïve Bayes	85.24
8	Extreme Gradient Booster	78.68
9	Proposed Ensemble Framework	91.80

The working performance of the model is represented by multiple metrics like ROC (Receiver Operating Characteristic), PRC (Precision-Recall Curve), and CM (Confusion Metric), as shown in Figure 11, and its classification report is reflected in Table 11.

Table 11: *Classification Report of AdaBoost*

ATTRIBUTE	PRECISION	RECALL	F1-SCORE	SUPPORT
CLASS 0	0.94	0.91	0.92	32
CLASS 1	0.90	0.93	0.92	29
ACCURACY			0.92	61
MACRO AVG	0.92	0.92	0.92	61
WEIGHTED AVG	0.92	0.92	0.92	61

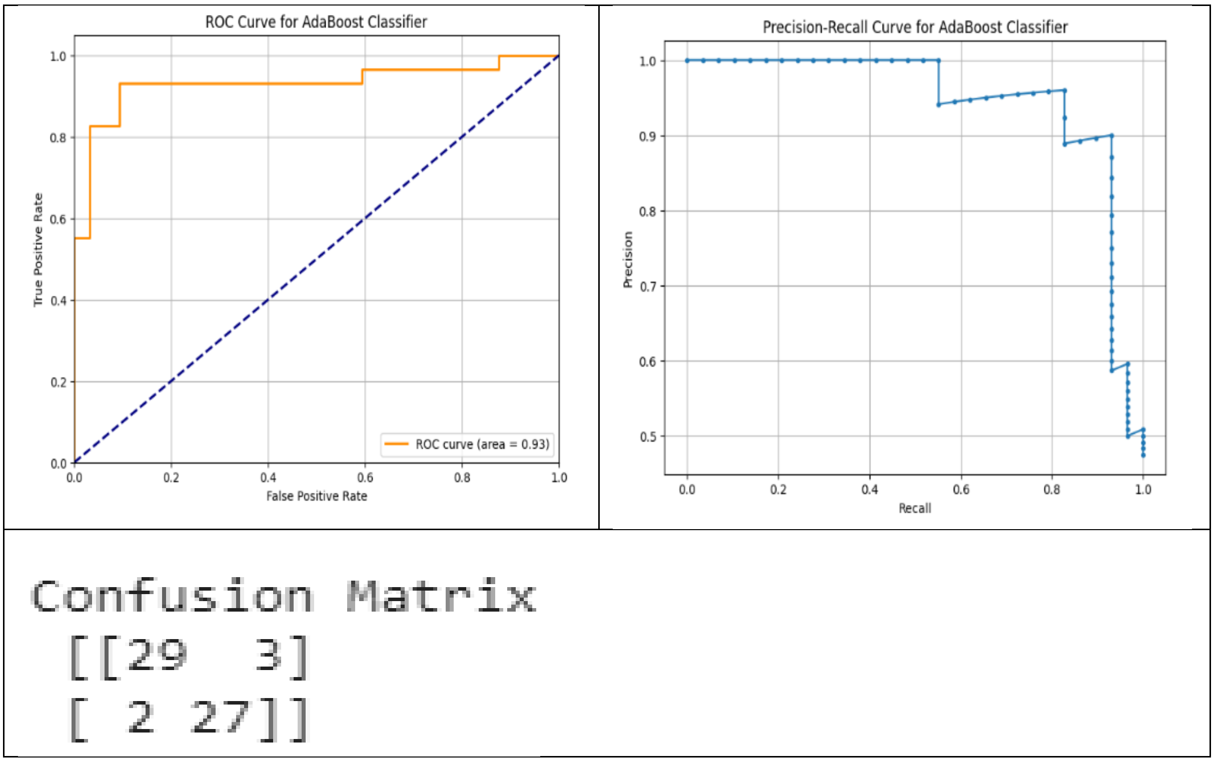

Figure 11: *Ada Boost ROC, PR, and Confusion Matrix*

The comparison of results accuracy is graphically represented by a Bar graph in Figure 12.

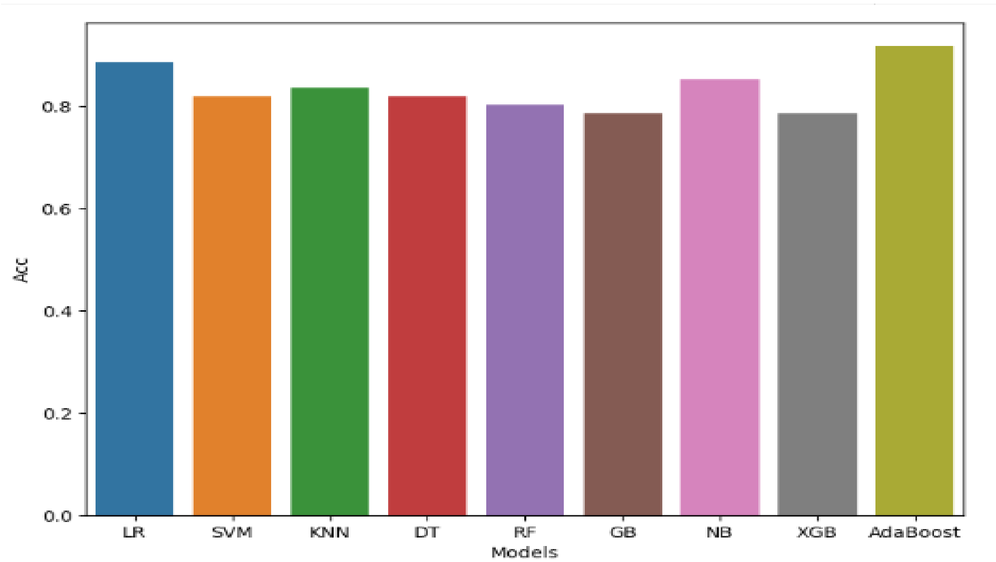

Figure 12: *Result Accuracy Bar Graph of Machine-Learning models*

V. Conclusion

This study has proved that enhanced accuracy can be achieved by using the HEF model, and heart disease can be predicted easily with fewer chances of errors. The comparative analysis has shown the importance of machine-learning techniques in healthcare decision-making. Our innovative Hybrid Ensemble Framework, which combines AdaBoost with Logistic Regression, has proved its supremacy over other machine-learning models with an accuracy of 91.80%. It proved its working and efficiency outstanding in risk stratification. The findings of our research work have shown the precision and effectiveness of ensemble methodologies and the promise of AdaBoost in enhancing the results of base classifiers. In the future, the accuracy of results may be improved by using deep learning techniques.

References:

1. S.S. Virani, et al., Heart disease and stroke statistics, Update. Circulation, 2021 143 (8) (2021) e254–e743
2. Chu, M., Wu, P., Li, G., Yang, W., Gutiérrez-Chico, J. L., & Tu, S. (2023). Advances in Diagnosis, Therapy, and Prognosis of Coronary Artery Disease Powered by Machine Learning Algorithms. JACC: Asia, 3(1), 1-14.
3. Al'Aref, S. J., Singh, G., Choi, J. W., Xu, Z., Maliakal, G., van Rosendael, A. R., ... & Min, J. K. (2020). A boosted ensemble algorithm for determination of plaque stability in high-risk patients on coronary CTA. Cardiovascular Imaging, 13(10), 2162-2173.
4. Sowmiya C, Sumitra P. Analytical study of heart disease diagnosis using classification techniques. In: IEEE international conference on intelligent techniques in control, optimization and signal processing (INCOS), March 2017; 2017. p. 23–5. https://doi.org/10.1109/ITCOSP.2017.8303115.
5. J. Han and M. Kamber, "Data Mining Concepts and Techniques," Morgan Kaufmann Publishers, 2006.
6. I.-N. Lee, S.-C. Liao, and M. Embrechts, "Data mining techniques applied to medical information," Med. inform, 2000.
7. M. K. Obenshain, "Application of Data Mining Techniques to Healthcare Data," Infection Control and Hospital Epidemiology, 2004.
8. S. C. Liao and I. N. Lee, "Appropriate medical data categorization for data mining classification techniques," MED. INFORM, vol. 27, no. 1, pp. 59–67, 2002.
9. I. H. M. Paris, L. S. Affendey, and N. Mustapha, "Improving Academic Performance Prediction using Voting Technique in Data Mining," World Academy of Science, Engineering and Technology, 2010.
10. M. Diwakar, A. Tripathi, K. Joshi, M. Memoria, P. Singh, and N. Kumar, "Latest trends on heart disease prediction using machine learning and image fusion," Mater. Today Proc., vol. 37, no. Part 2, pp. 3213–3218, 2020, doi: 10.1016/j.matpr.2020.09.078.
11. C. J. Harrison and C. J. Sidey-Gibbons, "Machine learning in medicine: a practical introduction to natural language processing," BMC Med. Res. Methodol., vol. 21, no. 1, pp. 1–18, 2021, doi: 10.1186/s12874-021-01347-1
12. M. Pérez-Ortiz, S. Jiménez-Fernández, P. A. Gutiérrez, E. Alexandre, C. Hervás-Martínez, and S. Salcedo-Sanz, "A review of classification problems and algorithms in renewable energy applications," Energies, vol. 9, no. 8, pp. 1–27, 2016, doi: 10.3390/en9080607.
13. S. Pandey, M. Supriya, and A. Shrivastava, "Data Classification Using Machine Learning Approach," no. June, 2018, doi: 10.1007/978-3-319-68385-0.
14. L. Shahwan-Akl, "Cardiovascular Disease Risk Factors among Adult Australian-Lebanese in Melbourne," International Journal of Research in Nursing, 2010.
15. I. H. M. Paris, L. S. Affendey, and N. Mustapha, "Improving Academic Performance Prediction using Voting Technique in Data Mining," World Academy of Science, Engineering and Technology, 2010.
16. Shao YE, Hou CD, Chiu CC. Hybrid intelligent modeling schemes for heart disease classification. Appl Soft Comput 2014; 14:47–52.
17. Yekkala I, Dixit S, Jabbar MA. August. Prediction of heart disease using ensemble learning and Particle Swarm Optimization. In: 2017 international conference on smart technologies for smart nation (Smart Tech Con). IEEE; 2017. p. 691–8.
18. Amin SU, Agarwal K, Beg R. April. Genetic neural network-based data mining in the prediction of heart disease using risk factors. In: 2013 IEEE conference on information & communication technologies. IEEE; 2013. p. 1227–31.
19. Tan PN, Chawla S, Ho CK, Bailey J, editors. Advances in knowledge discovery and data mining, Part II: 16th Pacific-Asia conference, PAKDD 2012, Kuala Lumpur, Malaysia, may 29-June 1, 2012, Proceedings, Part II, vol. 7302. Springer; 2012.

20. Chandel K, Kunwar V, Sabitha S, Choudhury T, Mukherjee S. A comparative study on thyroid disease detection using K-nearest neighbor and Naive Bayes classification techniques. CSI Trans. ICT 2016;4(2–4):313–9.
21. Ambrish, G.; Ganesh, B.; Ganesh, A.; Srinivas, C.; Mensinkal, K. Logistic Regression Technique for Prediction of Cardiovascular Disease. Glob. Transit. Proc. 2022, 3, 127–130.
22. Perumal, R. Early Prediction of Coronary Heart Disease from Cleveland Dataset Using Machine Learning Techniques. Int. J. Adv. Sci. Technol. 2020, 29, 4225–4234.
23. Shah, S.M.S.; Batool, S.; Khan, I.; Ashraf, M.U.; Abbas, S.H.; Hussain, S.A. Feature action through Parallel Probabilistic Principal Component Analysis for Heart Disease Diagnosis. Phys. A Stat. Mech. Its Appl. 2017, 482, 796–807.
24. Kodati, S.; Vivekanandam, R. Analysis of Heart Disease Using in Data Mining Tools Orange and Weka Sri Satya Sai University Analysis of Heart Disease Using in Data Mining Tools Orange and Weka. Glob. J. Comput. Sci. Technol. C 2018, 18, 17–21.
25. Mohan, S.; Thirumalai, C.; Srivastava, G. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. IEEE Access 2019, 7, 81542–81554.
26. Ananey-Obiri, D.; Sarku, E. Predicting the Presence of Heart Diseases Using Comparative Data Mining and Machine Learning Algorithms. Int. J. Comput. Appl. 2020, 176, 17–21.
27. Vijayashree, J.; Sultana, H.P. A Machine Learning Framework for Feature Selection in Heart Disease Classification Using Improved Particle Swarm Optimization with Support Vector Machine Classifier. Program. Comput. Softw. 2018, 44, 388–397.
28. Janosi, Andras, Steinbrunn, William, Pfisterer, Matthias, and Detrano, Robert. (1988). Heart Disease. UCI Machine Learning Repository. https://doi.org/10.24432/C52P4X.
29. Reddy, K.V.V.; Elamvazuthi, I.; Aziz, A.A.; Paramasivam, S.; Chua, H.N.; Pranavanand, S. Heart Disease Risk Prediction Using Machine Learning Classifiers with Attribute Evaluators. Appl. Sci. 2021, 11, 8352.
30. Pavithra, V.; Jayalakshmi, V. Hybrid Feature Selection Technique for Prediction of Cardiovascular Diseases. Mater. Today Proc. 2021; in press.
31. Ananey-Obiri, D.; Sarku, E. Predicting the Presence of Heart Diseases Using Comparative Data Mining and Machine Learning Algorithms. Int. J. Comput. Appl. 2020, 176, 17–21.
32. Ahamad, G.N.; Fatima, H.; Zakariya, S.M.; Abbas, M. Influence of Optimal Hyperparameters on the Performance of Machine Learning Algorithms for Predicting Heart Disease. Processes 2023, 11, 734.
33. Latha CBC, Jeeva SC. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Informatics in Medicine Unlocked Jan. 2019; 16:100203. https://doi.org/10.1016/j.imu.2019.100203.
34. Mohan S, Thirumalai C, Srivastava G. Effective heart disease prediction using hybrid machine learning techniques. IEEE Access 2019; 7:81542–54. https://doi.org/10.1109/ACCESS.2019.2923707.
35. Repaka AN, Ravikanti SD, Franklin RG. Design and implementing heart disease prediction using naives bayesian. In: 2019 3rd international Conference on Trends in Electronics and informatics (ICOEI); Apr. 2019. p. 292–7. https://doi.org/10.1109/ICOEI.2019.8862604.
36. N. Cheung, "Machine learning techniques for medical analysis," School of Information Technology and Electrical Engineering, B.Sc. Thesis, University of Queenland, 2001.
37. K. Polat, S. Sahan, and S. Gunes, "Automatic detection of heart disease using an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and k-nn (nearest neighbour) based weighting preprocessing," Expert Systems with Applications, pp.625–663,.2007.
38. M. C. Tu, D. Shin, and D. Shin, "Effective Diagnosis of Heart Disease through Bagging Approach," Biomedical Engineering and Informatics, IEEE, 2009.Expert Systems with Applications, Elsevier, pp. 7675–7680, 2009.

39. Shouman, Mai & Turner, Timothy & Stocker, Rob. (2012). Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients. International Journal of Information and Education Technology. 2. 220-223. 10.7763/IJIET. 2012.V2.114.
40. Sharma, A.; Mishra, P.K. Performance Analysis of Machine Learning Based Optimized Feature Selection Approaches for Breast Cancer Diagnosis. Int. J. Inf. Technol. 2022, 14, 1949–1960.
41. M. Diwakar, A. Tripathi, K. Joshi, M. Memoria, P. Singh, and N. Kumar, “Latest trends on heart disease prediction using machine learning and image fusion,” Mater. Today Proc., vol. 37, no. Part 2, pp. 3213–3218, 2020, doi: 10.1016/j.matpr.2020.09.078.