Technical Journal, University of Engineering and Technology (UET) Taxila, Pakistan *Vol. 28 No. 2-2023*
ISSN:1813-1786 (Print) 2313-7770 (Online)

Performance Analysis of Machine Learning Algorithms for Early Prognosis of Cardiac Vascular Disease

M. Hussain[1], A. Shahzad[2], F. Liaquat[3], M. A. Arshed[4], S. Mansoor[5], Z. Akram[6]

[1,3,4,5]*School of Systems and Technology, University of Management & Technology, Lahore, Pakistan*
[2]*Department of Computer Science, University of Engineering & Technology, Lahore, Pakistan*
[6]*Department of Computer Science, Lahore Leads University, Lahore, Pakistan*

[2] 2017mscs10@student.uet.edu.pk

Abstract- Cardiovascular disease, also known as heart disease, is on the rise. It is imperative to anticipate possible illnesses in advance, which is a difficult task that demands precision and efficiency. The main objective of this research paper is to identify patients who are at a higher risk of developing heart disease based on specific medical characteristics. To accomplish this, a heart disease prediction model was created that utilizes a patient's medical history to estimate the probability of a heart condition diagnosis. In this research, heart disease will be predicted using the dataset at hand, which includes 14 key attributes used for analysis. This study evaluates 18 machine learning models on a binary classification task using various performance metrics. The KNeighborsClassifier demonstrated the best performance across all metrics, achieving the highest train and test accuracy, precision, recall, F1-score, and AUC among all models. The ExtraTreesClassifier and GaussianProcessClassifier also performed well, while Gaussian Naïve Bayes, LinearSVC, NuSVC, and LogisticRegressionCV performed the worst. These findings suggest that the KNeighborsClassifier is the most suitable model for the binary classification task in question. This prediction can help clinicians analyze illness risk factors and assess patient scenarios. By focusing more on the condition's risk factors, it can be improved even further.

Keywords- Cardiovascular Disease, Machine Learning, Decision Tree, Linear Models, K-Nearest Neighbors, Ensemble Models

I. INTRODUCTION

Heart disease encompasses a range of conditions that can affect the heart. According to the World Health Organization, cardiovascular diseases are the leading cause of death worldwide, resulting in an estimated 17.9 million deaths annually. Multiple risk factors contribute to the development of heart disease, including obesity, high cholesterol, hypertension, and elevated triglyceride levels. The American Heart Association has identified additional symptoms, such as irregular heartbeat, sleep difficulties, rapid weight gain, and swollen legs, which could indicate the presence of heart disease. Neglecting these signs can lead to severe consequences. Notably, males are more susceptible to heart disease than females. Preventive measures, such as maintaining a healthy weight, quitting smoking, adopting a nutritious diet, and engaging in regular exercise, are effective in averting heart disease. Medical interventions, such as medications and surgeries, are also employed. As part of preventive strategies, computer technologies, particularly Machine Learning algorithms, are commonly employed to predict the occurrence of this disease prior to infection.

Healthcare organizations, including clinics and hospitals, face a significant challenge in providing high-quality services at affordable prices. Ensuring accurate patient diagnosis and efficient treatment delivery are essential for delivering such care. To this end, the heart disease database contains both numerical and categorical data, which undergoes thorough cleaning and filtering to remove any extraneous information before further processing. The proposed method enables the extraction of precise hidden information from a historical heart disease database, revealing patterns and correlations related to heart disease. Moreover, it offers solutions to complex diagnostic queries, assisting medical experts in making well-informed clinical decisions.

1. Machine Learning is a powerful technology utilized for testing that relies on both training and testing phases. Unlike traditional programming, machine learning empowers machines to learn

from data and enhance their performance over time. The amalgamation of AI and machine learning is sometimes referred to as Machine Intelligence, as it involves training machines to process and utilize data effectively. In this research, the testing data incorporates biological parameters such as cholesterol levels, blood pressure, sex, and age. Eighteen different algorithms, including ensemble models, linear models, Naïve Bayes models, k-nearest neighbor, support vector machines (SVM) models, and decision tree classifiers, are employed to evaluate the accuracy of the data. The objective of the study is to determine the algorithm that exhibits the highest accuracy. The research paper calculates the accuracy of each approach and uses these calculations to draw conclusions about which algorithm is most effective. This paper primarily offers the following contributions:

- Analyze the correlation and interdependence among various features in the dataset.
- An analysis was conducted to compare the performance of various machine-learning algorithms.
- Discovery of the optimal algorithm based on performance for predicting heart disease.

II. LITERATURE REVIEW

Cardiovascular disease diagnosis has been the focus of numerous research studies. Various machine learning procedures have been employed for the diagnosis, resulting in varying probabilities of different approaches by Garg, [1]. KNN and Random Forest machine learning algorithms were used to predict cardiac disorders. After collecting and analyzing the data, the balance was examined, and a correlation between different features and their impact on the desired value was discovered. The UCI dataset from Kaggle was used in this study, and it was separated into 80/20 for training and testing. An observation was made that the target attribute displayed a positive correlation with both Chest Pain and Maximum Heart Rate achieved. Using KNN and Random Forest, this model had an accuracy of 86.885% and 81.967%, respectively.

Kasabe and Narang [2] proposed a prediction model based on a web application that was trained on the UCI dataset, which was split 75-25 for testing and training purposes. The most accurate predictive models were based on Logistic Regression with an accuracy rate of 82.89%, followed by SVM at 81.57%, and Decision Tree and Naive Bayes at 80.43% each. This online application can be used by end-users as an initial test to evaluate their cardiac health and seek medical assistance if needed.

Shah, [3] provided a set of models using the WEKA tool and supervised learning methods. In order to estimate the likelihood of an individual having cardiovascular disease, four classification methods were employed, namely KNN, NB, DT, and RF. Prior to integration and reduction, the dataset underwent cleaning, smoothing, normalization, and aggregation processes. The KNN algorithm yielded the most accurate results.

The research paper states that the author Jindal, [4] developed a system named EHDPS, which stands for effective heart disease prediction system, that achieved an accuracy rate of 87.5%. The system employed three classification algorithms, namely KNN, RF, and LR. Among these three algorithms, KNN and Logistic Regression outperformed RF, with KNN having the highest accuracy rate of 88.52%. A cardiac disease prediction web application was developed by Pandita, [5], which takes in medical data of patients and predicts if they are suffering from any cardiac illness. The proposed model employs five ML algorithms, and the one with the highest accuracy is applied. The web app is developed using a framework that is based on Flask, and the design is implemented through the use of HTML/CSS. KNN algorithm produced the highest accuracy of 89.06%, whereas the lowest accuracy of 84.38% was contributed by Logistic Regression, according to the source.

Five algorithms were implemented by Alqahtani, [6], namely, RF, KNN, DT, and XGB. Out of these, the four algorithms that yielded the highest accuracy were RF (88.65%), KNN (86.45%), DT (86.35%), and XGB (88.19%). According to Arumugam, [7], the DT model consistently outperforms the NB and SVM models, and both DT and SVM models attain 90% accuracy. A sensitive score of 0.85 was achieved in the study [8]. Siuly and Zhang [9] reported that RF had the best performance with an accuracy rate of 88% in predicting heart disease, among the variety of techniques used including LR, SVM, GNB, Light, RF, and XGB. Our proposed approach attains the best accuracy when compared to other approaches that also use the UCI repository dataset. Ten ML classifiers' accuracy, precision (specificity), recall (sensitivity), and F-Measure were assessed in this paper.

In study the authors Arshed and Riaz [10], applied fifteen machine learning model to compare the performance for cardiovascular patient identification. The authors of this research paper utilized a machine learning technique along with the RapidMiner tool to classify heart disease. Their findings revealed that the Logistic Regression model demonstrated the most effective performance with an accuracy of 85.22% when compared to other machine learning models that

were evaluated. this study suggests that machine learning can be a useful tool for diagnosing heart disease, and Logistic Regression may be particularly effective for this purpose.

This study includes a comprehensive statistical analysis of the input datasets to comprehend how data range affects CVD (cardiovascular diseases) forecasts and eighteen machine learning consider for experiments. It includes a correlation investigation of the patients' continuous and categorical characteristics. To comprehend the correlation's relevance between significant features, scatter plots and data visualization for pairs of significant features were acquired. The results section discusses and examines these. In this research paper, three models were implemented, namely, random forest, logistic regression, and decision tree yielded the best accuracy. Furthermore, the correlation of the target variable was checked.

In their study, Angraal, [11] employed Machine Learning techniques to make predictions about the likelihood of hospitalization and mortality in individuals with heart failure. The researchers utilized five different methods, including SVM, Gradient Descent Boosting, Random Forest, LASSO regularization and LR with forwarding selection variable. To validate their findings, a three-year follow-up was conducted using 5-fold cross-validation. The results showed that Random Forest was the most effective method, with a mean C-statistic value of 0.76 for predicting hospitalization and 0.72 for predicting mortality. Furthermore, the researchers suggested that the inclusion of time-to-event analysis could potentially improve the outcomes of the proposed model.

Ashok Dwivedi [12] utilized various Machine Learning techniques, including SVM, KNN, ANN, Naive Bayes, Classification Tree, and Logistic Regression, to predict the likelihood of heart disorder occurrence. In addition, the performance of these methods was evaluated using the ROC curve. The results showed that Logistic Regression had the highest accuracy of 85%, with 89% sensitivity and 81% specificity. However, it is important to note that the model's reliability needs to be confirmed by testing it on large datasets.

Aljaaf, [13] proposed the method of multi-level risk assessment, which incorporates three risk factors (obesity, lack of physical activity, and smoking) along with existing attributes to assess the risk of heart failure. The researchers utilized the Decision Tree method, which yielded a prediction accuracy of 86.53%. However, the performance of radical feature selection approaches could potentially enhance the performance of this model. In their study, Mohan, [14] proposed a hybrid approach that combines two distinct approaches to improve accuracy. The hybrid approach resulted in an accuracy of 88.4%, which was higher than all other methods tested.

Ozcan and Peker [15] utilized the Classification and Regression Tree (CART) algorithm, a supervised machine learning technique, to predict heart disease and establish decision rules that clarify the relationship between input and output variables. The results of the study ranked the features that influence heart disease based on their importance, and the model's accuracy of 87% validates its reliability. Additionally, the decision rules extracted in the study can simplify the use of clinical purposes without requiring additional knowledge. The proposed algorithm can assist healthcare professionals and patients who face cost and time constraints in the diagnosis and treatment processes of heart disease.

Ogundepo and Yahya [16] presents a predictive analysis of heart disease risk factors using two independent datasets. The Cleveland data was analyzed using the Chi-square test of independence, and ten classification models were trained for class prediction. The support vector machine was found to have the best predictive performance with 85% accuracy, 82% sensitivity, 88% specificity, 87% precision, 91% area under the ROC curve, and 38% log loss value. The study highlights the importance of bio-clinical categorical variables in predicting heart disease and the potential of machine learning models in supporting healthcare professionals in the diagnosis and treatment of heart disease.

YILMAZ and YAĞIN [17] evaluated three different machine learning models, Random Forest (RF), Logistic Regression (LR), and Support Vector Machine (SVM), for the classification of coronary heart disease. The models were optimized using a 3-repeats 10-fold repeated cross-validation method and evaluated based on various performance metrics, including Accuracy, F1 Score, Specificity, Sensitivity, Positive Predictive Value, Negative Predictive Value, and Confusion Matrix. Results showed that RF had the highest accuracy of 0.929, followed by SVM with 0.897 and LR with 0.861. RF also had the highest sensitivity value, making it clinically important in minimizing overlooked heart patients. Overall, the study concludes that the RF model outperformed the other models in accurately classifying coronary heart disease.

Patel, [18] present three machine learning algorithms (J48, LMT, and Random Forest) were compared for their accuracy in predicting heart disease. The J48 decision tree algorithm was found to be the best classifier with the highest accuracy of 56.76% and the least time to build the model (0.04 seconds). The study also found that applying reduced error pruning to J48 improved its performance. In contrast, LMT had the

lowest accuracy of 55.77% and took the longest time to build the model (0.39 seconds).

Khemphila and Boonjing [19] presents a classification approach for diagnosing heart disease using a Multi-Layer Perceptron (MLP) with Backpropagation learning algorithm and a feature selection algorithm. Information Gain is used to select the most important attributes, reducing the original 13 to 8. The MLP algorithm is used to classify the diagnosis of patients, with an accuracy difference of 1.1% between 13 features and 8 features in the training dataset, and 0.82% in the validation dataset. This approach has the potential to reduce the number of tests needed for accurate diagnosis, improving healthcare efficiency, see Table 1 for literature summary.

Table 1. Cardiac Vascular Disease Literature Summary

Year	Author	Model	Accuracy
2021	Garg, [1]	KNN Random Forest,	86.885% 81.967%
2020	Kasabe and Narang [2]	Logistic Regression SVM Naive Bayes Decision Tree	82.89% 81.57% 80.43% 80.43%
2021	Harshit Jindal, [4]	KNN Logistic Regression KNN & LR based model	88.52% 88.5% 87.5%
2021	Pandita, [5]	Logistic Regression KNN SVM Naive Bayes Random Forest	84.38% 89.06% 87.50% 85.94% 87.50%
2022	Alqahtani, [6]	RF KNN DT XGB	88.65%, 86.45%, 86.35% 88.19%
2022	Doppala,. [8]	RF into fetal echocardiography	Sensitivity 0.85, Specificity 0.88,
2021	Arshed and Riaz [10]	15 Machine Learning Models	Accuracy 85.22 with Logistic Regression
2023	Ozcan and Peker [15]	Classification and Regression Tree (CART) algorithm	87%
2023	Ogundepo and Yahya [16]	10 classification models	85% with Support Vector Machine

III. Methodology of the System

The objective was to determine if a patient would experience CVD based on a specific set of clinical data. The confusion matrix of each technique indicated that 243 (80%) out of 303 occurrences in the dataset were used to train the two models. After the acquisition of the training data, 60 cases were provided to the trained models to determine their class before testing. This work aims to predict CVD risk using a computerized prediction method that can be helpful for medical practitioners. The methodology involves a series of procedures that convert provided data into patterns that can be easily understood by consumers. Figure 1 presented in the paper outlines a methodology that consists of three stages, namely data collection, feature extraction, and data exploration. In the first stage, data is collected from various sources and made ready for processing. The collected data is then preprocessed in the second stage using procedures such as handling missing values, data cleansing, and standardization, depending on the nature of the data. In the third stage, a classifier is applied to the preprocessed data, which is then used to build the model. The proposed model employs three classifiers, namely Random Forest, Logistic Regression, and Decision Classifier, to classify the preprocessed data. Finally, the model is implemented and evaluated for accuracy and performance using various indicators, which helps in assessing the effectiveness of the proposed methodology. Overall, the Figure 1 provides a useful

overview of the methodology and its various stages, highlighting the importance of data preprocessing and the use of multiple classifiers in improving the accuracy of the model. The following steps are taken to predict CVD:

- Step 1: Load Data.
- Step 2: Data cleaning through data wrangling, which involves removing null values and outliers.
- Step 3: Significant features extraction
- Step 4: Data Splitting (75% and 25%)
- Step 5: The data is trained and tested.
- Step 6: Applied machine learning models
- Step 7: Evaluation

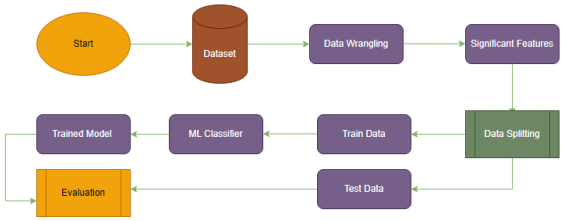


Figure 1. Flowchart of Methodology

Data Collection

The dataset used to develop the detection models for CVD was obtained from Kaggle and was converted into a comma-separated file (.csv) [20]. This dataset includes 303 samples and 76 attributes. The study only considered 13 important test attributes, including age, sex, pain, chest, cholesterol, maximum heart rate, fasting blood sugar, resting blood pressure, ST depression (old peak), the slope of peak ST segment, exercise-induced angina, thallium stress result, number of major vessels, and a target output (1 = patient with CVD, 0 = patient without CVD). Table 2

shows examples of these attributes. The desired value is whether a person has CVD (close to 1) or not (close to 0). The dataset is unbalanced as there were 165 patients with CVD and 138 without CVD as shown in Table 2.

Table 2: CVD Dataset Attributes Description

Attribute	Description	Range/Values
age	Age expressed in years.	25-90
sex	Sex (0 = female, 1 = male)	0, 1
cp	The type of pain experienced in the chest.	1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic
trestbps	Resting blood pressure, measured in mm Hg.	94-200
chol	Cholesterol level in blood serum, measured in mg/dl.	126-564
fbs	Fasting blood sugar over 120 mg/dl.	0 = false, 1 = true
restecg	Resting electrocardiographic results	0 = normal, 1 = ST-T wave abnormality, 2 = left ventricular hypertrophy
thalach	Maximum heart rate during exercise	71-202
exang	Exercise induced angina	0 = no, 1 = yes,
oldpeak	Exercise-induced ST depression relative to rest.	0-6.2
slope	The ST segment slope during peak exercise.	1 = upsloping, 2 = flat, 3 = down sloping
ca	The number of major vessels visualized by fluoroscopy.	0, 1, 2, 3
thal	Thalassemia	3 = normal, 6 = fixed defect, 7 = reversible defect
target	Diagnosis of heart disease (1 = present; 0 = absent)	1, 0

Data Preprocessing of CVD

In our research paper, we employed various data wrangling techniques, including the handling of null values and outliers. Data wrangling is a crucial step in preparing raw data for analysis, and addressing missing values and outliers is vital to ensure the integrity and reliability of our findings.

To address missing values, we systematically identified instances where data was incomplete or unavailable. We carefully considered the nature of the missing data and applied appropriate strategies, such as removing rows or columns with missing values or imputing them using estimation techniques. This meticulous approach allowed us to minimize the impact of missing data on our analysis and ensure the accuracy and completeness of our dataset.

In addition, we recognized the significance of outliers in our research. Outliers are extreme values that can significantly affect statistical analyses and lead to misleading results. Therefore, we implemented robust outlier detection methods, taking into account statistical techniques and domain knowledge. We carefully examined and appropriately handled outliers by either removing them, transforming them, or treating them separately in our analysis. By doing so, we ensured that the presence of outliers did not compromise the validity and reliability of our findings. By effectively applying these data wrangling techniques, we were able to establish a solid foundation for our research. We obtained a clean and refined dataset that was accurate, complete, and representative, enabling us to conduct rigorous and trustworthy analyses. This, in turn, enhanced the validity and reliability of our research outcomes, providing a strong basis for drawing meaningful insights and conclusions. For machine learning algorithms to work, it is necessary to preprocess the raw data by converting it into a clean and usable format. This involves transforming the data into a numeric representation, which requires converting any categorical labels into binary-valued column vectors. An unpleasant issue in the data set is missing values or Nan (not a number). The empty rows must either be dropped or filled in with a mean or interpolated data.

Machine Learning Algorithms

In this research work, several machine learning algorithms were applied to analyze and model the data. Ensemble methods, including AdaBoost, GradientBoostingClassifier and bagging classifier were utilized to combine the predictions of multiple models and improve accuracy. Gaussian Processes, GaussianProcessClassifier specifically were also used for probabilistic classification. Linear models, such as LogisticRegressionCV, SGDClassifier PassiveAggressiveClassifier, Perceptron and RidgeClassifierCV were employed for their simplicity and interpretability. Naïve Bayes algorithms, including BernoulliNB and GaussianNB, were applied for classification based on probability theory. The Nearest Neighbors algorithm (KNN) was used for pattern recognition and classification. Support Vector Machine models, including SVC, Linear SVC, and NuSVC, were applied for classification and regression analysis. Finally, RandomForestClassifier, extra trees classifiers, and Decision Trees were employed as simple yet powerful tools for decision-making and classification. Total eighteen machine learning models consider in this study for experiments.

IV. Experimental Analysis

Data Visualization of CVD

In this section, the dataset details like outliers, EDA and variable correlation presented visually.

Detecting and Removing Outliers

An outlier is defined as an observation that significantly deviates from the other values in a random sample taken from a population. However,

determining what constitutes abnormal behavior is ultimately subject to the judgement of the analyst or a consensus process. Figure 2 show the outlier in the dataset of CVD. There are many ways to detect the outlier like z-Score and IQR Score. We use IQR Score to detect the outlier. The interquartile range (IQR) is a statistical measure of dispersion that represents the range of values within which the middle 50% of the data falls. It is also known as the H-spread and is calculated as the difference between the upper and lower quartiles, specifically the 75th and 25th percentiles.

In other words, the IQR is defined as Q3 minus Q1. So, there is too much outlier in the dataset we remove the outlier from the dataset.

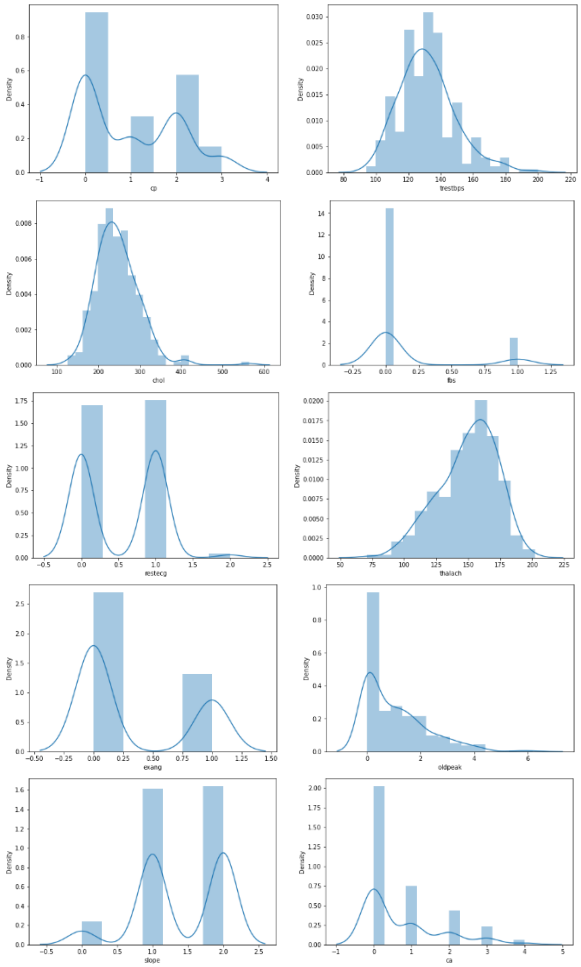


Figure 2. Outlier of CVD Dataset

Exploratory Data Analysis (EDA)

According to the graph shown in Figure 3, there appears to be a correlation between increased blood pressure as a function of age and a higher maximum heart rate, which is commonly associated with heart disease.

Figure 3. EDA of CVD Dataset

Evaluation Metrics

To assess the effectiveness of machine learning algorithms, a range of performance metrics are used, including a confusion matrix that incorporates true positive (TP), false positive (FP), false negative (FN), and true negative (TN) values. This enables the evaluation of both current and future data. A matrix table, known as perplexity, is often used to evaluate the performance of classifiers. Five criteria are used to assess each classifier's performance: classification accuracy, F1-score, specificity, sensitivity and receiver operating characteristics (ROC). The classification accuracy is calculated using four factors: TN, FP, TP, and FN. It assesses the proportion of cases that are correctly classified. The sensitivity represents the percentage of patients with CVD who are accurately diagnosed, while specificity shows the proportion of patients who are free of CVD and were correctly classified.

Accuracy:

The metric of accuracy is used to measure performance and indicates the percentage of accurate predictions made on the test data. To calculate accuracy, the total number of correct predictions is divided by the total number of predictions made, as illustrated in equation (1).

$$A = \frac{correct\ predictions}{all\ predictions} \quad (1)$$

Precision:

Precision is a metric that measures the proportion of true positive predictions among all the examples that were predicted to belong to a particular class, and it can be calculated using an equation (2).

$$P = \frac{TP}{TP+FP} \quad (2)$$

Recall:

Recall is a metric that indicates the proportion of correctly predicted examples for a particular class out of all the examples that truly belong to that class, as shown in equation (3).

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

F1- Score:

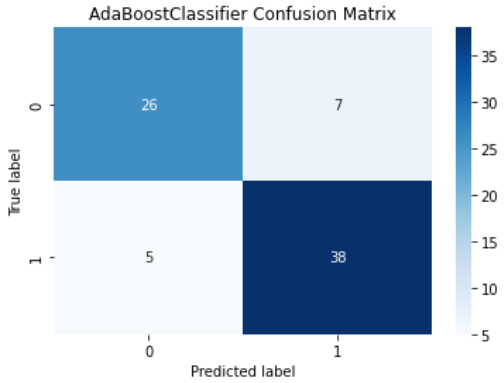
The F1-Score is a performance metric that incorporates a classifier's recall and precision using their harmonic mean, resulting in a single score. It is commonly used to compare the effectiveness of two classifiers. Formula (4) is used to calculate the F1-Score.

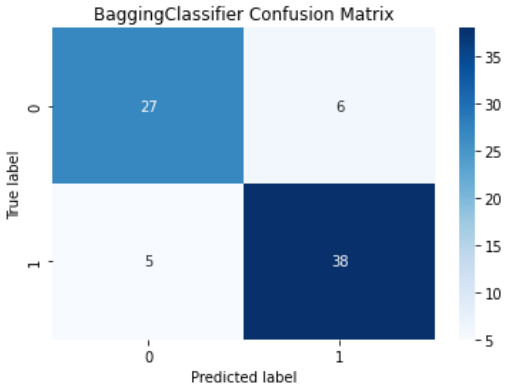
$$F1 = \frac{2*(P*R)}{P+R} \quad (4)$$

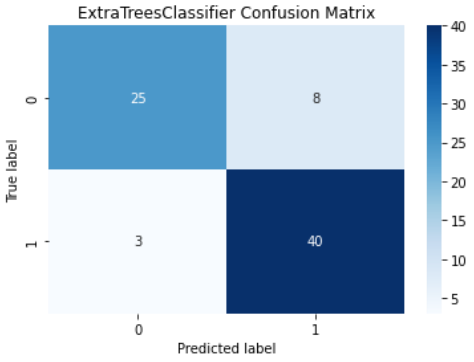
Confusion Metrics

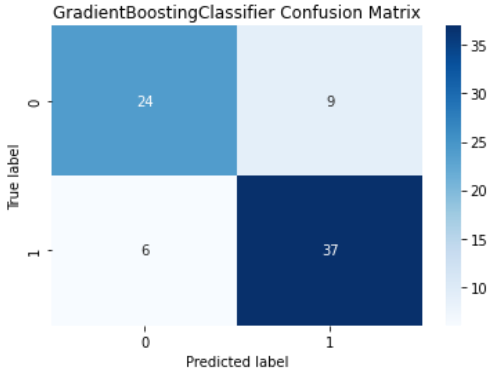
A confusion matrix compares the predicted and actual labels of a dataset, displaying the number of true negatives, false negatives, true positives, and false positives for each class. It helps evaluate a classification model's performance and reveals the types of errors made, such as false positives or false negatives, providing insights for improvement, as shown in Table 3.

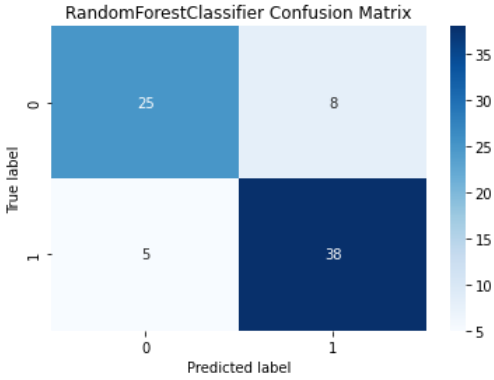
Table 3: Confusion Matrix of 18 different machine learning models

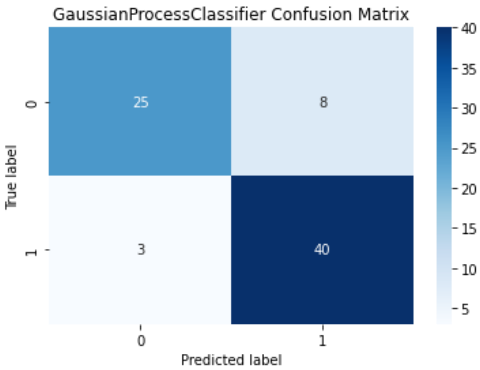


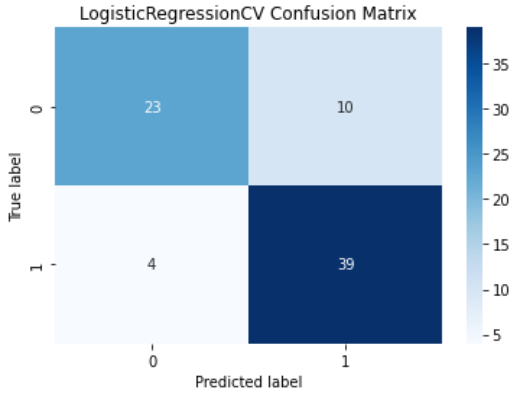


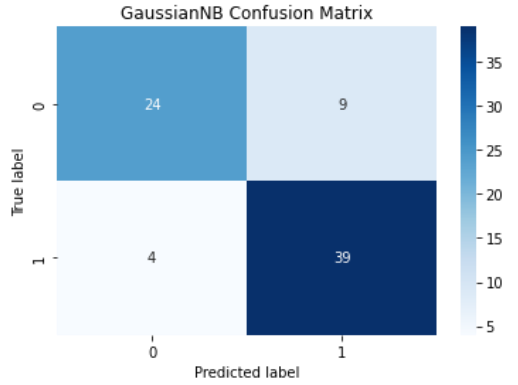


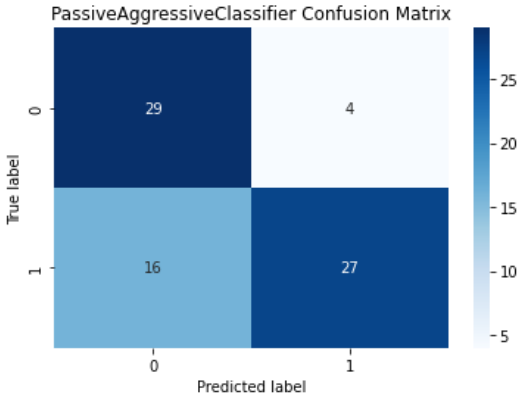


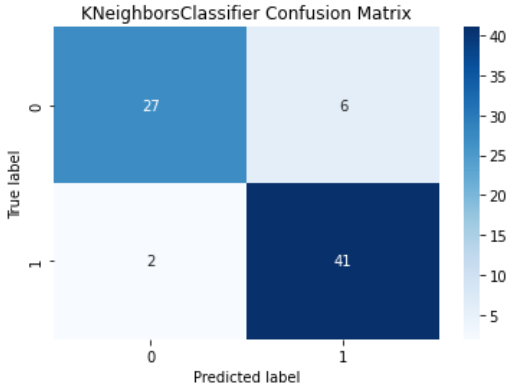


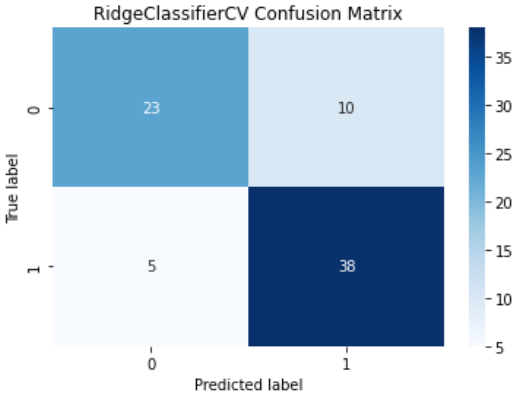


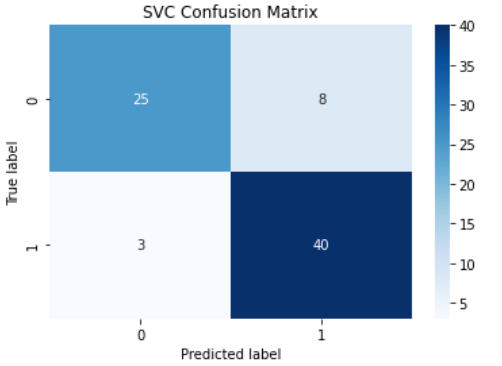

LogisticRegressionCV Confusion Matrix
23
10
4
39
True label
Predicted label

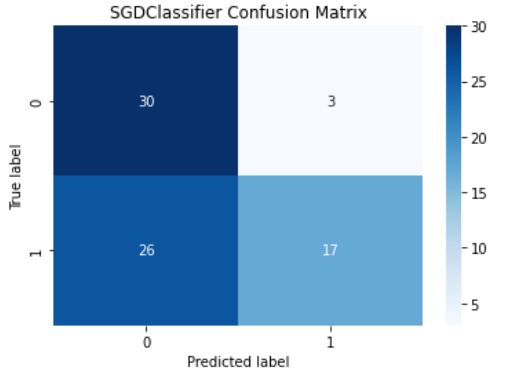

GaussianNB Confusion Matrix
24
9
4
39
True label
Predicted label

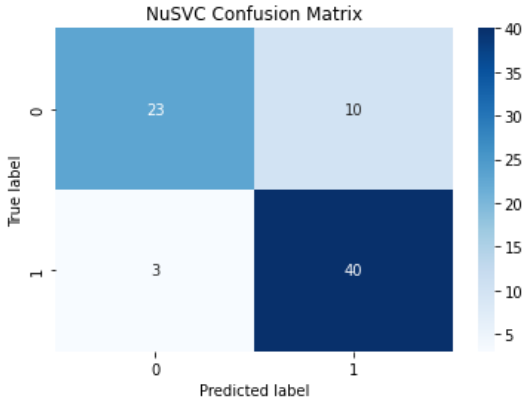

PassiveAggressiveClassifier Confusion Matrix
29
4
16
27
True label
Predicted label

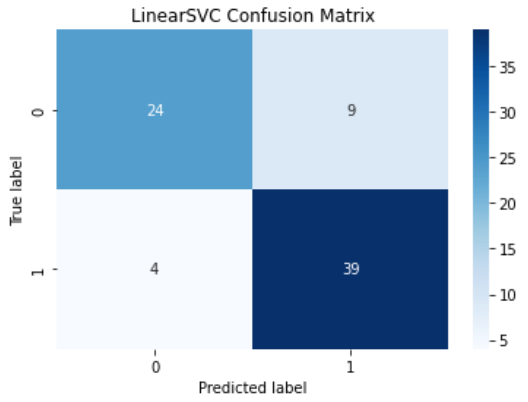

KNeighborsClassifier Confusion Matrix
27
6
2
41
True label
Predicted label

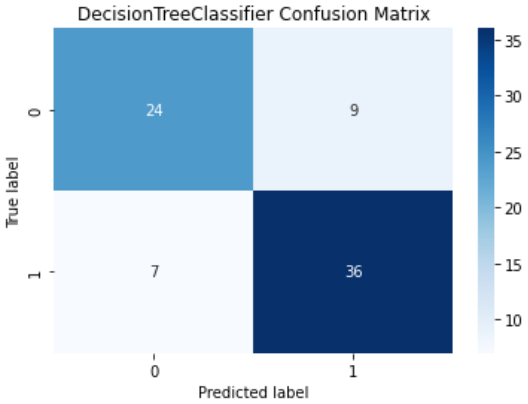

RidgeClassifierCV Confusion Matrix
23
10
5
38
True label
Predicted label


SVC Confusion Matrix
25
8
3
40
True label
Predicted label


SGDClassifier Confusion Matrix
30
3
26
17
True label
Predicted label


NuSVC Confusion Matrix
23
10
3
40
True label
Predicted label





V. RESULTS COMPARISONS

The Table 4 provides the results of 18 different machine learning models on a binary classification task. The models were calculated based on their train accuracy, test accuracy, recall, precision, F1-score, and AUC. The KNeighborsClassifier achieved the highest train accuracy (0.876) and test accuracy (0.8947), while also having the highest precision (0.872) and recall (0.953), resulting in the highest F1-score (0.912) and AUC (0.885) among all models. The Extra Trees Classifier and Gaussian Process Classifier achieved the second-highest test accuracy (0.8553) and AUC (0.843). The Gaussian Naïve Bayes model had the lowest train and test accuracy (0.854 and 0.828, respectively) among all models, with similar performance to Linear SVC, Logistic Regression CV and Nu SVC.

Table 4: Evaluation Scores of ML Models

	ML Models	Train Accuracy	Test Accuracy	Precision	Recall	F1-score	AUC
1	KNeighborsClassifier	0.8767	**0.8947**	0.8723	0.9534	0.9111	0.8858
2	ExtraTreesClassifier	1.0000	0.8553	0.8333	0.9302	0.8791	0.8439
3	GaussianProcessClassifier	0.9163	0.8553	0.8333	0.9302	0.8791	0.8439
4	Bagging Classifier	0.9912	0.8553	0.8636	0.8837	0.8735	0.8509
5	SVC	0.8855	0.8553	0.8333	0.9302	0.8791	0.8439
6	AdaBoostClassifier	0.9383	0.8421	0.8444	0.8837	0.8636	0.8358
7	RandomForestClassifier	1.0000	0.8289	0.8260	0.8837	0.8539	0.8206
8	LinearSVC	0.8458	0.8289	0.8125	0.9069	0.8571	0.8171
9	NuSVC	0.8634	0.8289	0.8000	0.9302	0.8602	0.8136
10	GaussianNB	0.8546	0.8289	0.8125	0.9069	0.8571	0.8171
11	LogisticRegressionCV	0.8414	0.8158	0.7959	0.9069	0.8478	0.8019
12	GradientBoostingClassifier	1.0000	0.8026	0.8043	0.8604	0.8314	0.7938
13	RidgeClassifierCV	0.8458	0.8026	0.7916	0.8837	0.8351	0.7903
14	Decision Tree Classifier	1.0000	0.7895	0.8000	0.8372	0.8181	0.7822
15	KNeighborsClassifier	0.8767	0.8947	0.8723	0.9534	0.9111	0.8858
16	ExtraTreesClassifier	1.0000	0.8553	0.8333	0.9302	0.8791	0.8439
17	GaussianProcessClassifier	0.9163	0.8553	0.8333	0.9302	0.8791	0.8439
18	Bagging Classifier	0.9912	0.8553	0.8636	0.8837	0.8735	0.8509

Figure 4 and Figure 5 are the train and test accuracy comparison graph of eighteen different machine learning algorithms. From the graph, it is observable that the highest test accuracy achieved with KNN.

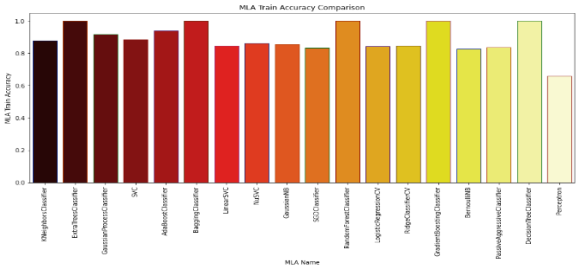


Figure 4: Machine Learning Algorithms Train Accuracy Comparison (Graph)

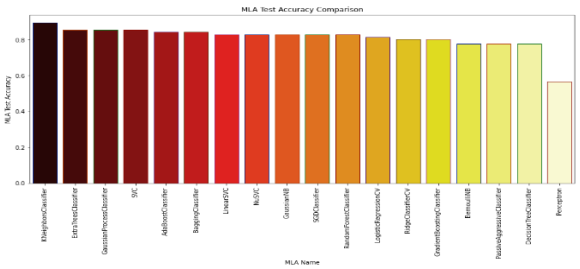


Figure 5: Machine Learning Algorithms Test Accuracy Comparison (Graph)

ROC Curves

The results show the AUC values for ROC curves of various machine learning classifiers. The AdaBoostClassifier and Gaussian Process Classifier have the highest AUC value of 0.84, followed by KNeighborsClassifier with an AUC value of 0.89. The RandomForestClassifier, Bagging Classifier and

ExtraTreesClassifier have AUC values between 0.77 and 0.82. The Gradient Boosting Classifier, Logistic Regression CV, Passive Aggressive Classifier, SGDClassifier, GaussianNB, SVC, and LinearSVC have AUC values between 0.79 and 0.84. The RidgeClassifierCV and NUSVC have AUC values of 0.79 and 0.81 respectively. Finally, the Perceptron and BernoulliNB have the lowest AUC values of 0.61 and 0.77 respectively, as shown in Figure 6.

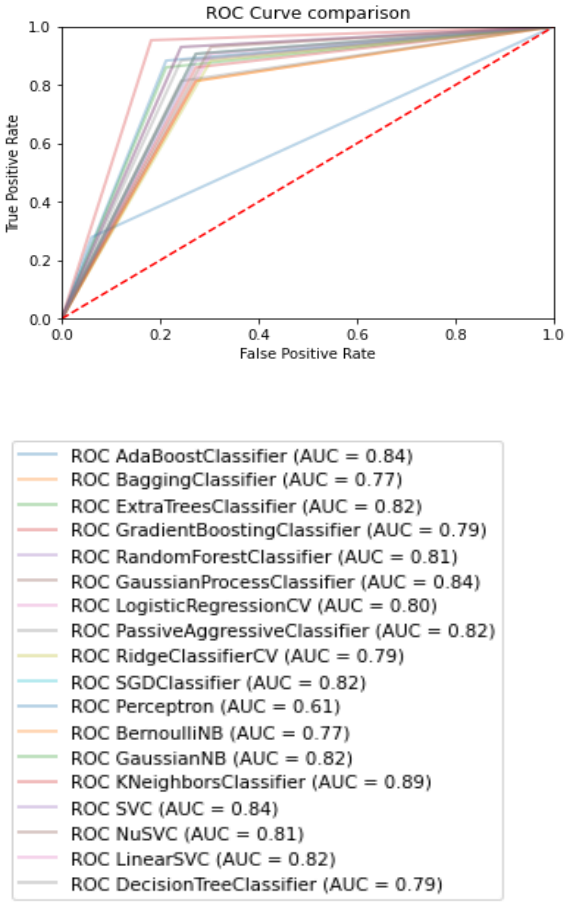


Figure 6: AUC Scores of ML Models

VI. Conclusion

Heart problems are becoming increasingly common, particularly in our country. Therefore, identifying the illness before it becomes severe can significantly lower the risk of death. Extensive research has been conducted in the field of cardiac disease prediction. Our study focuses on the identification and prognosis of cardiac disease using machine learning algorithms. this study evaluated the performance of 18 different machine learning models on a binary classification task using various performance metrics. The results showed that the KNeighborsClassifier demonstrated the best performance across all metrics, achieving the highest train and test accuracy, precision, recall, F1-score, and AUC among all models. The ExtraTreesClassifier and GaussianProcessClassifier also performed well, while the Gaussian Naïve Bayes model and several other models performed poorly. Based on these findings, the KNeighborsClassifier appears to be the most suitable model for the binary classification task. Our future work involves developing the best application for patients and doctors to detect heart diseases with data oversampling of existing dataset.

References

[1] A. Garg, B. Sharma, and R. Khan, "Heart disease prediction using machine learning techniques," in *IOP Conference Series: Materials Science and Engineering*, 2021, p. 012046.

[2] R. Kasabe and G. Narang, "Heart Disease Prediction using Machine Learning," *International Journal of Engineering Research & Technology (IJERT)*, vol. 9, pp. 353-355, 2020.

[3] D. Shah, S. Patel, and S. Bharti, "Heart disease prediction using machine learning techniques. Sn Comput Sci 1: 345," ed, 2020.

[4] H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, "Heart disease prediction using machine learning algorithms," in *IOP conference series: materials science and engineering*, 2021, p. 012072.

[5] A. Pandita, S. Vashisht, A. Tyagi, and S. Yadav, "Prediction of Heart Disease using Machine Learning Algorithms," 2021.

[6] A. Alqahtani, S. Alsubai, M. Sha, L. Vilcekova, and T. Javed, "Cardiovascular Disease Detection using Ensemble Learning," *Computational Intelligence and Neuroscience*, vol. 2022, 2022.

[7] K. Arumugam, M. Naved, P. P. Shinde, O. Leiva-Chauca, A. Huaman-Osorio, and T. Gonzales-Yanac, "Multiple disease prediction using Machine learning algorithms," *Materials Today: Proceedings*, 2021.

[8] B. P. Doppala, D. Bhattacharyya, M. Janarthanan, and N. Baik, "A reliable machine intelligence model for accurate identification of cardiovascular diseases using ensemble techniques," *Journal of Healthcare Engineering*, vol. 2022, 2022.

[9] S. Siuly and Y. Zhang, "Medical Big Data: Neurological Diseases Diagnosis Through Medical Data Analysis," *Data Science and Engineering*, vol. 1, 06/01 2016.

[10] M. A. Arshed and F. Riaz, "Machine Learning for High Risk Cardiovascular Patient Identification," *Journal of Distributed Computing and Systems (JDCS)*, vol. 4, pp. 34-39, 2021.

[11] S. Angraal, B. J. Mortazavi, A. Gupta, R. Khera, T. Ahmad, N. R. Desai, *et al.*, "Machine learning prediction of mortality and hospitalization in heart failure with preserved ejection fraction," *JACC: Heart Failure*, vol. 8, pp. 12-21, 2020.

[12] A. K. Dwivedi, "Performance evaluation of different machine learning techniques for prediction of heart disease," *Neural Computing and Applications*, vol. 29, pp. 685-693, 2018.

[13] A. J. Aljaaf, D. Al-Jumeily, A. J. Hussain, T. Dawson, P. Fergus, and M. Al-Jumaily, "Predicting the likelihood of heart failure with a multi level risk assessment using decision tree," in *2015 Third International Conference on Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE)*, 2015, pp. 101-106.

[14] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE access*, vol. 7, pp. 81542-81554, 2019.

[15] M. Ozcan and S. Peker, "A classification and regression tree algorithm for heart disease modeling and prediction," *Healthcare Analytics*, vol. 3, p. 100130, 2023.

[16] E. A. Ogundepo and W. B. Yahya, "Performance analysis of supervised classification models on heart disease prediction," *Innovations in Systems and Software Engineering*, pp. 1-16, 2023.

[17] R. YILMAZ and F. H. YAĞIN, "Early detection of coronary heart disease based on machine learning methods," *Medical Records*, vol. 4, pp. 1-6, 2022.

[18] J. Patel, D. TejalUpadhyay, and S. Patel, "Heart disease prediction using machine learning and data mining technique," *Heart Disease*, vol. 7, pp. 129-137, 2015.

[19] A. Khemphila and V. Boonjing, "Heart disease classification using neural network and feature selection," in *2011 21st International Conference on Systems Engineering*, 2011, pp. 406-409.

[20] fedesoriano. (September 2021). *Heart Failure Prediction Dataset*. Available: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction