International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Exploring the Potential of Supervised Learning Methods for the Investigation and Identification of Cardiovascular Diseases Using Machine Learning

B. Blessed Sam[1]*, M. Prakash[2]

Submitted:10/03/2024 **Revised**: 25/04/2024 **Accepted**: 02/05/2024

Abstract: Cardiovascular-disease continues to be a worldwide concern, especially impacting countries that are economically disadvantaged. The vast data sets that are generated by the healthcare business may now be processed by machine learning algorithms. This research uses Grid-Search CV to tune parameters and compare six classification techniques for selecting the most accurate prediction model. It is innovative in applying these algorithms to assess patients' risk factors based on medical information. Study results identified major risk factors of cardiovascular disease. The machine proceeds to the next classification step, if an individual's F1-Score exceeds 91.3%. Within the scope of the study, a number of different machine learning approaches were evaluated for classification tasks. The accuracy of logistic regression was 0.90%, the accuracy of support vector classifier was 0.91%, the accuracy of k-nearest was 0.92%, the accuracy of Gaussian Naive Baye's was 0.91%, and the accuracy of decision tree classifier was 0.89%. Finally, the Random Forest Classifier had the highest accuracy of all the approaches that were investigated, coming in at 0.96%. These findings demonstrate that varied approaches are used by machine learning systems when it comes to classification problems.

Keywords: *Cardiovascular Health Evaluation, Classifiers, Predictive Modeling, Diagnostic Evaluation metrics, Risk Prediction.*

1. Introduction

The use of digitization and digital transformation has helped to drive tremendous growth in the medical profession all around the globe over the course of the last ten years. This growth may be attributed to the globalization of the healthcare industry. However, in today's fast-paced environment, the vast majority of people choose to avoid medical checkups unless they have severe health-related concerns.

This is especially true among the younger generations. In a similar vein, the vast majority of people do not go in for regular cardiac exams since the traditional procedures are time-consuming and difficult to implement in day-to-day life. If a person is uninformed of their present cardiac status, consequences may lead to serious health difficulties, and in the worst circumstances, untimely deaths.[1].

In order to monitor individual's heart health in a way that is both efficient and convenient, a sophisticated system is required. The Internet of Things (IoT) has arisen in recent times as a significant contributor to the field of healthcare as a result of its core characteristics, which include connection, sensing, reliability, consistency, and cognitive capabilities. This development came about as a result of the fact that IoT has these characteristics for heart disease.

Utilizing sensing devices to monitor and track important health signs such as Blood Pressure, Pulse Rate and Electrocardiogram (ECG), it is a method for revolutionizing modern healthcare by providing personalized and proactive treatment. This care is provided via the provision of a means to revolutionize current healthcare. Live data may be gathered from the human body by using a number of different wearable sensors that are designed for health monitoring.

Moreover, the use of cardiac recordings and cardiac CT scans, which are essential for identifying coronary heart disease, are frequently exorbitant and unattainable for numerous low- and middle-income nations. Henceforth, prompt identification of cardiovascular ailment is paramount to mitigate its physiological and monetary toll on persons and establishments.

According to projections, the cumulative count of fatalities resulting from cardiovascular diseases (CVDs), predominantly attributable to cardiac ailments and cerebrovascular accidents. Henceforth, it is of the utmost significance to employ data mining and machine learning methodologies to estimate the probability of contracting cardiovascular ailments so as to preserve lives and curtail the financial strain on the community.

These data may be examined and converted into health records, which then have the potential to be used in the prevention of cardiovascular disease, treatment of cardiovascular disease, and recovery from cardiovascular disease [2]. Because there is no one symptom that defines

[1] *Department of Data Science and Business Systems, School of Computing, SRM Institute of Science and Technology, Kattankulathur, Chennai-603203, Tamil Nadu, India.*
[2] *Department of Data Science and Business Systems, School of Computing, SRM Institute of Science and Technology, Kattankulathur, Chennai-603203, Tamil Nadu, India.*
** Corresponding Author Email: bb4439@srmist.edu.in*

cardiovascular illness, improving the promptness and accuracy of diagnosis is a problem. There is not one particular sign or symptom that represents cardiovascular diseases, the importance of a certain characteristic in the diagnosis of a disease is brought into focus here.

In most cases, diagnoses are arrived at by a physician after doing an assessment of the patient's present state of health, going through the patient's medical history, and considering any other pertinent criteria [3]. In addition, there is the possibility that machine learning algorithms may be able to accelerate the process of diagnosing cardiovascular disorders and improve the degree of accuracy that can be reached in one's projections.

This is an exciting potential where several research initiatives have made use of a wide variety of machine learning algorithms to facilitate the clinical diagnosis of cardiovascular disorders [4]. This has been done with the goal of providing an aid to clinician. In the end, it is essential important to evaluate the functionality, efficacy, and user satisfaction of the program in relation to its stated aims to determine whether it was successful or not in medical-aspects.

The purpose of this study is to make a prediction about the possibility of a diagnosis of cardiovascular heart disease. Through the use of this data, we are able to ascertain whether or not the patient is at danger for developing heart disease. By classifying individuals according to twelve medical factors that suggest the possibility that they will acquire heart illness or not. In order to find out how well various machine learning algorithms predict the start of heart illness.

Here in this study six distinct prediction-modeling approaches are employed to accomplish this goal. Regardless, most of these studies have relied on very little samples, thus the results may not extrapolate to larger groups.

Using a larger and more diverse dataset, this work aims to avoid this limitation. The most important objective of this study is to provide an accurate prediction about the likelihood of heart disease occurring in the human body.

2.Related Work

Several investigations have been carried out regarding the prognostication of heart disease employing both neural networks and conventional machine learning methodologies. As per the research conducted by Dinesh et al. [5] the usage of machine learning algorithms for predicting cardiovascular disease was investigated. The study examined with multiple heart disease datasets from the database of UCI and suggested the usage of logistic regression. Ali et al. [6] proposed in a scholarly study the use of ontology-centered ideas for patients based on their clinical documentation and data collected at the same time. Again, a small number of studies have suggested that you might be able to tell if someone has a heart problem just by looking at their electrocardiogram. Nashif et al. [7] conducted a study on the detection of heart disease using ML algorithms. They proposed a real-time cardiovascular health monitoring system, which includes an application UI for both doctors and patients to collect information as user input, in addition to wearable sensors. Additionally, the paper had identified some sensors that may have a lower impact on the disease prediction model, potentially limiting the mobility of the wearable device. Mohan. S et al. [8] introduced a novel approach for employing machine learning to discover essential traits in order to improve the predictability of cardiovascular disease. The prediction model is shown with numerous different feature combinations and well-known classification techniques. Using a hybrid random forest approach with a linear model to predict heart disease, they improved performance with a prediction accuracy of 88.7%. Ram et al. [9] identified the use of machine learning in edge computing has been explored as a means of detecting anomalies and enhancing data accuracy. Also, mobile health monitoring has the potential to predict outcomes. On top of that, the multi-modal sensor information underwent pre-processing to clean the data and classify the events in the dataset. Umar. U et al. [10] has proposed a Cardiac Healthcare System that makes use of IoT technology to provide ubiquitous healthcare. The Smart Cardiac Care System provides continuous monitoring in real time, ensures patient confidentiality, and limits the need for physical examinations by healthcare professionals in cardiac units. The model stood out due to the hybridization of variables and electrocardiographic (ECG) data. The system might alert you about unusual statistics. Cloud servers enable healthcare providers to access information regarding patients from anywhere. A fog-assisted IoT-based medical cyber framework was introduced by Karthick et al. [11] for patients with cardiovascular issues. A suggested healthcare support system for cardiac patients involves monitoring and reporting to hospitals. They developed a cardiac monitoring application with Fog-assisted IoT technology to boost the decision support system. The software was compared to traditional analogue healthcare applications to showcase its superior performance. The study exclusively utilized real-time HBR measurements acquired from clinical ECG tests. The intent of Jameel Ahmed et al. [12] is to interpret current data on cardiovascular disorders to predict and prevent early onset of heart disease. The dataset of patients with heart disease was collected and stored in a system that uses the cloud. The data is pre-processed and examined using machine learning techniques to predict heart problems. Khan et al. [13] an Internet of Things framework that makes use of a Modified Deep Convolutional Neural Network has been cleverly developed to predict heart disease that allowed for real-time tracking of heart rate and other vital signs. An innovative illness detection system has been designed to skillfully categorize sensor data into two unique categories: normal and abnormal. This system has been built to identify

diseases. The system will swiftly and precisely identify any irregularities that are discovered in the event that they are detected. In the following table, a demonstration of the significant amount of effort that was done by a number of authors in order to forecast risk by using categorization strategies and the study of the methodology by employing various algorithms throughout the course of time and the accuracy they obtained from various learning classification (supervised) algorithms.

3.Proposed System

This research aims to use a range of features as potential predictors of future cardiovascular disease (CVD) risk. Various classification algorithms have been examined to assess their effectiveness in predicting CVD, with most of them depending on information obtained from the repository.

Based on the findings, accuracy ranges from around 80% to 96%. To begin, investigations that make use of private datasets are hindered by inconsistencies in database size and structure, for which bigger public datasets are not readily accessible. However, based on the datasets that are available to the public, experiments could not achieve the same level of performance if they are moved from the benchmark domain to the patient care domain.

The below fig 1. Shows the working methodology of predictive modeling architecture. Following the completion from the selection procedure, the following records were chosen:

270 records were taken from the Stat-log data set, 303 documents were taken from the Cleavand database, 294 records were taken from the Hungarian file, 123 records were taken from the Swiss file, 200 records were taken from the Long Beach VA database, and 303 records were taken from the Cleveland database. The datasets mentioned above were combined to produce a total of 1190 data points.

All of these datasets were derived using identical parameters to the others. an effect on the diagnosis and treatment of coronary artery disease. It is to everyone's advantage to compile CAD data from a variety of countries and to develop associated networks. The bulk of the datasets that were analyzed have just a limited number of feature types. It is essential to take into consideration that the performance of ML approaches may be impacted by the quantity of specimens and characteristics, which can have a considerable effect on the findings that are ultimately obtained.

Table 1. Existing Methods

Author	Dataset	Algorithm	Accuracy	Techniques
Mert Ozcan,2023[14]	Medical records	Decision Tree Classifier	87	Classification
Chintan M. Bhatt,2023[15]	UCI-Repository (70,000 patients)	Random Forest	95	Multilayer Perceptron with Cross Validation
Surendra Reddy Vinta,2023[16]	UCI-Repository (303 patients	Random Forest	78.2	Classification
Umarani Nagavelli,2022 [17]	Cleavand and Stat log (600 patients)	XgBoost	95.9	Classification
ElSeddawy, 2022 [18]	UCI-Repository (303 patients)	Random Forest	89.01	Classification
Neha Nandal, 2022 [19]	UCI-Repository (294)	XgBoost	94	Classification
Srichand Doki,2021 [20]	UCI-Repository (303 patients)	XgBoost	85.96	Classification
Ali. A Barhoom,2021 [21]	Kaggle (303 patients)	Gaussian-NB	74	Classification
Khan &Mondal, 2020 [22]	Kaggle (462 patients)	Logistic Regression	72.72	Regression
Yuvraj Nikhate, 2020 [23]	UCI (303 patients)	Support Vector Machine	92.22	Optimized stacked SVM based expert system

J. Jeyaganesan,2020 [24]	Kaggle (303 patients)	Random Forest	93	Classification
Devansh Shah ,2020 [25]	UCI-Repository (303 patients)	K-Nearest Neighbors	90.8	Classification
Ahmed F. Otoom, 2015 [26]	UCI-Repository (303 patients)	Support Vector Machine	84.5	Classification
K. Vembandasamy,2015 [27]	UCI-Repository (303 patients)	Naïve Bayes	86.4	Classification
D.R. Patil, 2014 [28]	UCI-Repository (303 patients)	Random Forest	85.55	Learning Vector Quantization
M.A. Jabbar,2013 [29]	UCI-Repository (303 patients)	Random Forest and Chi square	83.7	10-fold cross validation

This work implies to generate CAD records that have enhanced distinguishing qualities. The major objective of this inquiry is to determine which machine learning algorithm has been shown to be the most accurate in terms of providing a prognosis for cardiovascular disease. In the first step of the process, the data from all of the different sources are blended together into one large CSV file.

Following this, the method of data preprocessing is applied, which entails the elimination of outliers, a normalization of the data, the determination of the most advantageous feature sets from the standardized data, which is then followed by hyperparameter tuning, and finally, the validation of the data for the purpose of processing it into the machine learning section. Performing this action is done with the purpose of determining the most effective performance measurements, risk status, and the incidence of heart disease.

3.1 Dataset

This research used 1190 records from a dataset that included data from numerous similar sources. Data was organized into columns and stored as a.csv file. It had no null values since all variables were category data types.

This data collection has several peculiarities. This technique clips the top and lower tenths of the percentile range for continuous variables with significant standard deviations.

There were also unusual events like systolic blood pressure below diastolic. These samples of fake data were given before being deleted so the model could make predictions based on real data. Finally, the number components were ungrouped.

To ensure uniformity between measurements, they were standardized between 0 and 1. To make real-data forecasts, the component of interest that was dangerously near to equilibrium was eliminated. Finally, numeric variables, which were not categorical, were normalized between 0 and 1 to ensure consistency across all dataset measures.

Table 2. Dataset Description

Attributes	**Domain**	**Data Type**
Age	Years [29-77]	Real
Sex	Male-1, Female-0	Binary
Chest Pain Type	1=Typical angina 2=Atypical angina 3=non-anginal pain 4=no pain	Nominal
Resting Blood Pressure	94-200 mm/Hg	Real

Cholesterol	126-564 mg/dl	Real
Fasting Blood Sugar	>120 mg/dl, 1=Yes, 0=No	Binary
Resting Electrocardiogram	0=normal Nominal 1=having ST-T wave abnormality 2=showing probable or definite left ventricular hypertrophy	Nominal
Maximum Heart Rate	71-202 bpm	Real
Exercise Angina	1=Yes, 0=No	Binary
Old Peak	0-6.2 (Measured in Depression)	Real
ST_Slope	1=upsloping 2= fat 3=down sloping	Ordered
Heart Disease	output class [1- heart disease, 0 - Normal]	Binary

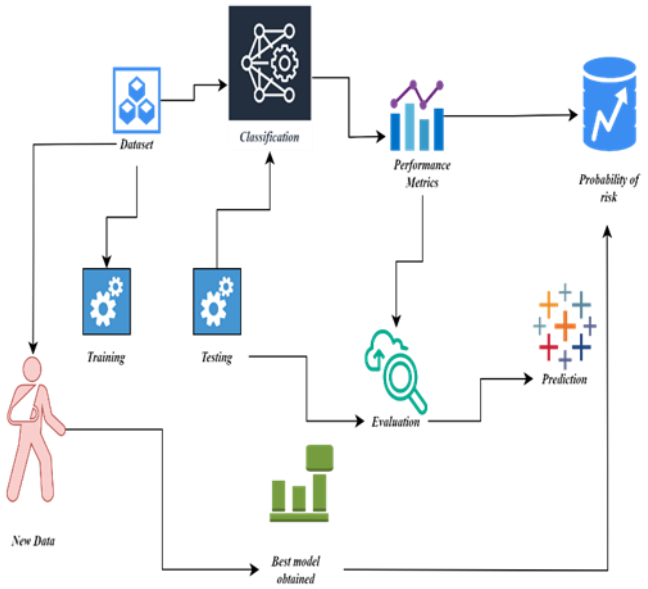


Fig.2 Outline Representation

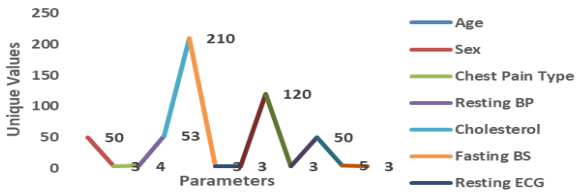


Fig 3. Number of unique values in the dataset.

3.2 Data Preprocessing

A methodology employed during the collection, cleansing, and examination of unprocessed data with the intention of training and developing machine learning models. Information that has frequently been compiled from a variety of sources constitutes raw data, which

should not be utilized for rapid analysis. Data preprocessing, which should not be regarded as unprocessed data intended for rapid analysis, can be conceptualized as data preparation. Consequently, it is imperative to scrutinize the data collection process to detect and subsequently eliminate any outliers or cluttered information. In this work, the variety of cleansing techniques are utilized during the course of investigation, including locating any missing numbers, identifying and removing any anomalies (outliers), and carrying out these procedures in sequential order. Here the system utilized Python programming within the Anaconda framework to examine the dataset for missing values. This action was taken to guarantee the dataset's completeness. The procedure of collecting supplementary data, conducting an analysis of it, and eliminating it from the model's consideration set [30]. There are many beneficial methods that can be used to achieve positive outcomes. Data preprocessing methods include the management of missing values, the conversion of feature types, and various other processes. Data preprocessing encompasses a variety of techniques utilized to improve the level of data. Feature Scaling involves the normalization of data through the use of Min-Max and Z-Score Normalization techniques. The range of values for independent variables, also known as features of the data, may be made more comparable by using a technique known as feature scaling. A method that is often used for the purpose of standardizing the range of independent variables or features included within a dataset is known as feature scaling [31]. This procedure is usually referred to as data normalization within the realm of data processing, and it is typically carried out as a component of the data pretreatment step of the workflow. The primary goal of this study is to determine which clinical characteristics may most effectively enhance the accuracy with which heart disease is predicted. Feature selection is the process of narrowing down a vast number of candidates features to a more manageable set of characteristics that will have the greatest impact on the final result.

A) Min-Max Normalization: The conversion of each numerical feature value into a new value that is based on the lowest and maximum values of the feature is a common and widespread use of this approach in classification algorithms. This technique is an example of how classification algorithms are used. This strategy makes a contribution to the enhancement of the algorithms' correctness as well as the efficiency with which they perform their functions.

$$\mathbf{x'= \frac{X-Min(x)}{Max(x)-Min(x)}} \quad \{1\}$$

From the above formula, Min indicates the minimum value in the selected feature, Max reflects the maximum value in the selected feature, x' represents a new selected value after applying normalization, and x represents a selected value from a numerical feature. The method described here effectively scales the values of a feature to a positive range between 0 and 1. This can be achieved by simply subtracting the minimum value of the feature from each value and then breaking down by the range of the feature's values.

B) Z-Score Standardization: The use of z-scores, which have the ability to quantify the number of standard deviations that each value deviates from the mean of the feature, is one way that may be utilized for the purpose of standardizing data. For the goal of providing an overview of the benefits and drawbacks connected with the use of z-scores for the purpose of standardizing data in predictive modelling, a subsequent information will be presented.

$$\mathbf{z = \frac{(x-\mu)}{\sigma}} \quad \{2\}$$

The suggested strategy involves transforming each numerical feature value based on the standard deviation and mean of the feature. This is achieved by applying standardization, where z represents the new value obtained after the transformation, and x represents the original value picked from the numerical feature and μ represents the standard deviation.

C) Feature Scaling: A technique that is used for the goal of standardizing the range of independent variables or characteristics of data is known as scaling of features. A variety of methods are used in order to accomplish this. When it comes to the area of data processing, this technique is sometimes referred to as data normalization, and it is something that is typically carried out at the stage of data preparation [32]. Table.3 includes a variety of independent variables, such as age, maximum heart rate, and cholesterol. Each of these variables has a range that falls somewhere between 36 and 60 years, 98 to 170 beats per minute, and 164 to 301 milligrams per deciliter, respectively. The ranges of these variables are shown in the table. The use of feature scaling would be of tremendous aid in bringing all of these variables into the same range. The strategy that was used in this particular instance concentrated on either the number 0 or anything that fell within the range of 0 to 1, depending on the scaling technique that showed up to be employed. The system approach rescaled the input data into a new fixed range of between -1 and 1 or 0 to 1, depending on the preference. Min-Max normalization is a linear transformation that may be used to normalize the raw original dataset. This technique maintains the relationship between the value that is input and the value that is scaled, where X is the value that was originally entered, X' is the value that was normalized, and max and min are the maximum and minimum values of the feature, respectively. The adjusted number for the data rescales the new value of

a features, which is indicated by X', while the prior value of the feature is represented by X. X' is the new value of the feature.

TABLE 3. Min-Max and Z-score (Single Feature Scaling)

Age	Min-Max (X)	Z-score (z)	Feature scaling
49	0.298507	-0.29951	0.298507
37	0.46932	0.716489	0.46932
48	0.354892	0.035867	0.354892
54	0.323383	-0.15155	0.323383
39	0.562189	1.268878	0.562189
45	0.393035	0.262741	0.393035
54	0.344942	-0.02332	0.344942
37	0.343284	-0.03318	0.343284
48	0.470978	0.726353	0.470978
37	0.349917	0.006275	0.349917
58	0.271973	-0.45734	0.271973
39	0.338308	-0.06277	0.338308
49	0.38806	0.233149	0.38806
42	0.349917	0.006275	0.349917
54	0.452736	0.617848	0.452736
38	0.325041	-0.14169	0.325041
43	0.333333	-0.09237	0.333333
60	0.411277	0.371246	0.411277
36	0.442786	0.558664	0.442786

Outlier Detection:

Outliers, or uncommon data items, may impair statistical tests and contradict their procedures. Analysts will encounter outliers and must decide how to manage them. Due to disruption, deleting them from records may be preferable [33]. Outliers may only be removed in certain scenarios.

The Interquartile range is used to estimate the distance between data points and the mean. It is split into quarters. IQR is a good approach for discovering outliers in a normally distributed dataset. If there is a larger difference between two values, the data points will be further apart, lowering the mean. Identification of significant deviations is done via IQR. From below Fig.4, Cholesterol has a lower limit of 103.34 and an upper limit of 400.75, while Q1 is 203.4 and Q3 is 304.5.

Data points that are less than the lowest and more than the maximum are outliers. In contrast, the left subplot shows five data points over the top limit. No cholesterol data falls below the lower limit. In the right subplot, these harsh cases are ignored. The heart rate parameter has a 60.9-point lower limit and 190.5-point upper bound. So set Q1 at 120.9 and Q3 at 150.6. Outlier data points are those below and above the minimum and maximum.

The left subplot reveals no heart rate data points below the lower limit, whereas the right subplot shows nine data points over the upper limitation. These extreme cases have been ignored. The previous peak feature ranges from -2.3 to 3.4, whereas Q1 and Q3 are zero and 1.9. Outliers are points below the lower limit and over the upper bound.

The left subplot shows that no data point goes below the lower limit, yet five data points exceed the higher limit for this characteristic. In the right subplot, these harsh cases are ignored. Lower limit of 90.5, highest limit of 194.5 for Blood Pressure. Data points outside the minimum and maximum are outliers. In the left subplot, two data points are below the lower limit for this property, while none surpass it. These extreme cases were ignored, as seen in the right subplot.

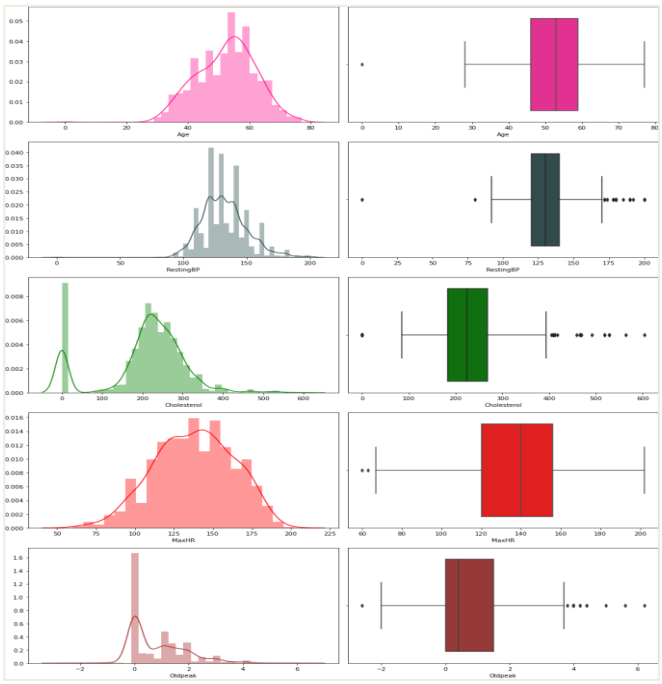


Fig 4. Outlier Identification (Graphical Representation)

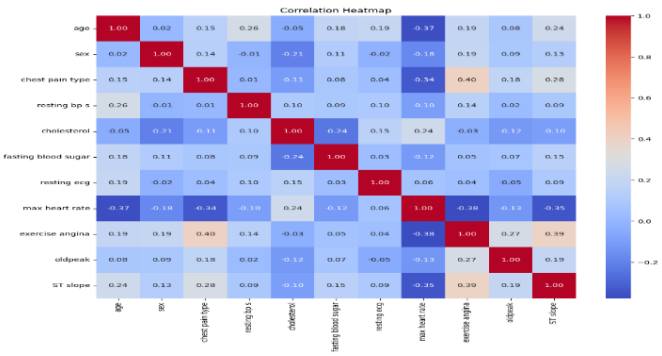


Fig 5. Heat Map

Once ensuring that there is no imbalance in the data, the Seaborn library is used to determine the relationship between the data and present it as a heat map. The above fig.5 aims to investigate the relationship and association between variables. The heatmap colors reflect how well each attribute matches the output target class, while other attributes represent additional qualities. A brighter hue usually indicates a stronger bond. Target attribute was most highly connected with chest pain type, specific types of chest pain, exercise-induced angina, and maximal heart rate.

Feature Selection

Predictive modelling uses several characteristics. However, other characteristics may be more important. Thus, to improve the model's predicted accuracy, each feature's value in relation to others must be examined. Predicting heart disease requires weighing demographic features.

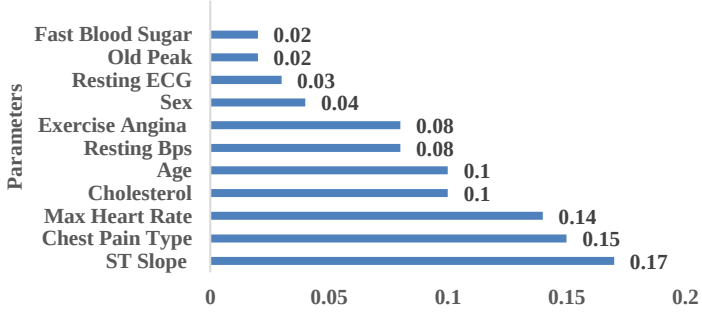


Fig 8. Feature Importance

Inferences are drawn in fig.8 on the relative significance of a number of model characteristics to the prediction made by the model. It simply determines whether or not a certain variable should be included in a model and the predictions that follow from that model. Prioritizing factors that have a significance of more than 0.6 include the kind of chest discomfort, the blood sugar level after fasting, the resting ECG, and the gender of the patient.

Each model in the study provides customizable controls for assessing performance [34]. Using these settings, it may improve or degrade model performance. Before using a machine learning model, improve its performance to assure correctness. Experts must modify hyperparameters to improve the model's learning. Fitting a model to a training dataset and fine-tuning its hyperparameters is a common way to increase its performance.

3.5 Hyperparameter Tuning

TABLE 4: Best Parameters are 'max_features':2, n -estimators :130 with a score of 0.93.

max features	**n_ estimators**	**Accuracy**
1	10	0.91250
1	20	0.93250
1	30	0.90750
1	40	0.83125
1	50	0.85000

3.6 Tuning with GridSearchCV

Param grid is a parameter that enables to transmit the grid of parameters that user is looking for. The grid should be structured as a dictionary, where the keys are the parameter labels of the estimator, and the data are arrays of values to be used for each variable [35].

Table 5: Tuning Parameters

Grid Search CV Runs (0-31)	0	1	2	3
mean_fit_time	0.006242	0.006247	0	0
std_fit_time	0.007644	0.007651	0	0
mean_score_time	0	0.003125	0.003125	0.003124
std_score_time	0	0.00625	0.006249	0.006248
param_C	0.001	0.001	0.001	0.001
param_gamma	0.001	0.01	0.1	1
split0()_test_score	0.0347826	0.0347826	0.0347826	0.0347826
split1()_test_score	0.0347826	0.0347826	0.0347826	0.0347826
split2()_test_score	0.363636	0.363636	0.363636	0.363636
split3()_test_score	0.363636	0.363636	0.363636	0.363636
split4()_test_score	0.409091	0.409091	0.409091	0.409091
mean_test_score	0.366403	0.366403	0.366403	0.366403
std_test_score	0.022485	0.022485	0.022485	0.022485
rank_test_score	27	27	27	27

Runs (32-35) score	32	33	34	35
mean_fit_time	0	0	0.002699	0
std_fit_time	0	0	0.005398	0
mean_score_time	0.003132	0	0	0.003132
std_score_time	0.006263	0	0	0.006264
param_C	100	100	100	100
param_gamma	0.1	1	10	100
split0()_test_score	1	0.956522	0.869565	0.521749
split1()_test_score	0.956522	0.956522	0.913043	0.521749

split2()_test_score	1	1	1	0.590909
split3()_test_score	0.863636	0.863636	0.818182	0.590909
split4()_test_score	0.954545	0.954545	0.954545	0.691919
mean_test_score	0.954941	0.946245	0.911067	0.581423
std_test_score	0.049799	0.044708	0.063488	0.058964
rank_test_score	11	13	18	24

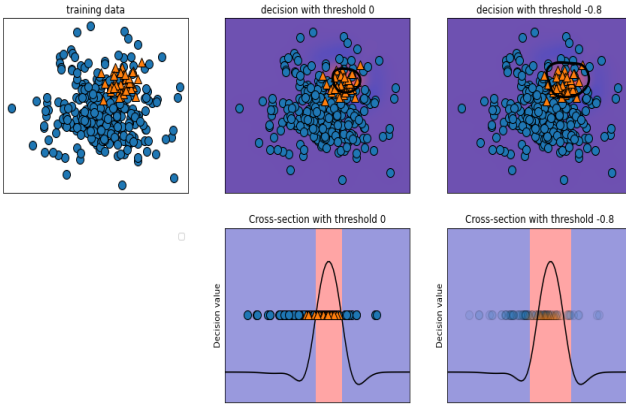


Fig 6. Decision Threshold

From table 5. it is inferred that the best parameter obtained the best parameters are {'C': 100, 'gamma': 0.01, 'kernel': 'rbf'} with the best cross validation score of 0.97. Sklearn does not provide a direct way to set the decision threshold. However, from the above fig.6, it does grant them access to the decision scores (output of the decision function) that are implemented in making predictions. The first approach is to choose the highest score from the decision function results and use it as the Decision threshold value. Any decision score values below this threshold can be considered as a negative in value class 0, while any values above the threshold can be considered as a positive class 1. Typically, machine learning strategies are employed to estimate classes, and a common approach is to use a decision threshold to determine if a person has experienced a coronary event. Given the unevenness in the dataset, this work opts to employ all of the models to produce probability forecasts instead of class predictions. By leveraging the disparity in the number of participants with CVD and those without the disease, can effectively capitalize on this opportunity. Given that certain algorithms do not provide direct probabilistic predictions, and considering that the probabilities predicted by these models may be not considered, so this technique employed the ability to enhance the capacity for future predictions.

3.7 Classification Algorithms:

Logistic Regression: Logistic regression is a valuable predictive classification strategy that may be used when doing analysis on a dataset that has a dependent variable that is a category and several independent factors that impact the outcome. In contrast to the technique known as linear regression, the approach known as logistic regression makes use of the logistic sigmoid function in order to transform its output into a probability value that may be assigned to two or more distinct categories [36]. In the realm of medicine, one of the most common applications of the statistical method known as logistic regression is to forecast the incidence of heart disease. In order to provide accurate forecasts about cardiac disease, a machine learning model based on logistic regression is used. At the outset, the models of logistic regression are trained with the assistance of five distinct splitting criteria. After that, these models are assessed with the use of test data in order to achieve the best degree of accuracy possible and to assess the operation of the algorithms. The approach divides occurrences into two distinct categories: category 1 indicates the presence of heart disease, while category 0 indicates that no indicators of cardiac sickness are present.

K-Neighbors Classifier: The K-Nearest Neighbors approach categorizes new data or cases based on their presumed resemblance to previously collected information or examples. Using the information that has been saved by the K-NN algorithm, a recently gathered statistic is categorized according to the similarities that it has with other statistics. This suggests that the K-NN approach may be used to efficiently classify new data as it becomes available, which is a significant implication [37]. Training a classifier using the K-Nearest Neighbors algorithm. The KNN algorithm classifier will now undergo training using the data that is currently accessible. Importing the K-Neighbors Classifier class from the Sklearn Neighbors library is going to be done in order to do this. After the class has been imported, the Classifier object will be generated automatically. The number of neighbors that are required by the algorithm will serve as the defining parameter for this class.

Support Vector Classifier: It is referred to as a support vector when the data points that are closer together and have a higher influence on the location and orientation of the hyperplane [38]. In order to enhance the margin of the classifier that makes use of these support vectors, the position of the hyperplane will change if the support vectors are removed without a reason. These components make it possible for us to design SVC in order to categorize data points. One of the most distinctive features of the support vector machine approach is that it searches for a hyperplane in a space with N dimensions, where N is the number of characteristics. The support vector machine (SVM) has a high computational cost, despite the fact that it has a high accuracy in classification. The large range of benchmark problems demonstrates that the SVC method has greater generalization performance when compared to the techniques that are considered to be state-of-the-art. Additionally, the SVC approach has quicker convergence and a lower number of support vectors while maintaining the same level of pattern classification quality.

Gaussian-Naive Bayes: It is assumed in Gaussian Naive Bayes that the continuous numerical characteristics follow a normal distribution. First, divide the property into subsets according to the kind of output we're interested in. Each parameter (also known as features or predictors) in Gaussian Naive Bayes is treated as though it were capable of predicting the output variable on its own. The final prediction, which provides a probability for the dependent variable to be placed in each group, is the sum of the predictions for all parameters. The Naive Bayes classifier works on the assumption that feature importance is unrelated to its neighbors. In order to estimate the parameters necessary for classification, Naive Bayes classifiers need data to use as training. Naive Bayes classifiers are useful in many practical settings because of their straightforward design and implementation [39].

Decision Tree Classifier: The controlled approach of decision tree is utilized for both unconditional and numerical value prediction. Data occurrences and their associated class labels are shown in a tree structure. The tree may be used to infer a set of rules that can be applied to the unknown data record to determine how it should be ranked in relation to the output value. The central node undergoes an attribute test. A tree branch with a class name at its leaf node represents the test's outcome. The entire data collection or set of sample points is divided into two or more groups using this method. The ratio of splits is determined by the parameter or factor that is predicted to be the most effective separator or differentiator. The Decision Tree Classifier is a class that has the capability to perform multiple classifications on a given dataset. If there are multiple classes with the same highest probability, the algorithm will predict the category with the lowest ranking among them [40].

Random Forest Classifier: The Random Forests method makes use of the prediction subset that is the most accurate overall, and this subset is picked at random from each individual node. Both regression and classification tasks may be accomplished with the assistance of this tool. In addition to that, this algorithm stands out as the solution that is the most accessible and adaptable. A forest is made up of many different kinds of trees. The presence of a higher number of trees in a forest has been shown to have a direct correlation with an increase in the forest's overall strength [41]. The Random module is put to use in the context of data analysis in order to provide data samples that are chosen at random. These samples are chosen in a way that is neither deterministic nor random in order to provide an accurate and objective portrayal of the whole dataset. A random forest is a kind of meta-estimator that makes use of averaging in order to boost projected accuracy and control over-fitting. This is accomplished by training several decision tree classifiers on various subsamples of the dataset.

IV. Results & Discussions

In order to accomplish the goals of this inquiry, a dataset that was related to heart disease was analyzed. In addition to the identification and removal of outliers, a number of classification techniques, such as SVC, KNN, GNB, DT, RF, and LR, have also been used. The confusion matrix, the area under the curve (AUC), and the probability ratio curve (PRC), as well as the clustering comparison, and the statistics that were obtained from the dataset, were the subsequent steps in the comparison study, which was based on all of the parameters. In conclusion, the findings of the best classifier, the ROC Curve Analysis, and the best classification model were provided.

Evaluation criteria are numerical measures employed to gauge the performance and efficiency of a statistical or machine learning model. The selection of evaluation metrics is influenced by the particular problem, the nature of the information and the intended outcome. Accuracy, Precision, Recall, and F1-score are the metrics that are used to assess how well the evaluation metrics are doing. The abbreviations RPV and RNV stand for "True Positive Value" and "True Negative Value," respectively, whereas SPV and SNV refer to "False Positive Value" and "False Negative Value," respectively.

“Accuracy” of a binary categorization may also be determined by calculating the number of both positive and negative outcomes.

$$\textbf{Accuracy} = \frac{\mathbf{RPV+RNV}}{\mathbf{RPV+RNV+SPV+SNV}} \quad \textbf{(1)}$$

"Precision" denotes the percentage of total number of information (data) points that have been appropriately adjusted for each individual example.

$$\textbf{Precision} = \frac{\textbf{RPV}}{\textbf{RPV+SPV}} \qquad \textbf{(2)}$$

The term "Recall" refers to the process of comparing an expected positive number against one that actually occurs.

$$\textbf{Recall} = \frac{\textbf{RPV}}{\textbf{RPV+SNV}} \qquad \textbf{(3)}$$

Machine learning uses the "F1-score" to assess model accuracy. This technique combines model accuracy and recall evaluations into one value. The accuracy metric counts the number of times a model correctly predicted something across a dataset.

$$\textbf{F1-score} = \frac{\textbf{2}\times\textbf{Recall}\times\textbf{Precision}}{\textbf{Recall+Precision}} \qquad \textbf{(4)}$$

Confusion matrix contains the two dimensions, "Actual class" and "Predicted class," in each dimension. The actual categories of heart disease are listed in the rows, while those that may be anticipated are shown in the columns. The dataset has two classes: Class 0 and Class 1.

Both of these classes are present. The confusion matrix from fig.10 offers a graphical depiction of the performance of the algorithm.

A straightforward way to visually examine the prediction mistakes is provided by the confusion matrix table. There are 93 positive values and 100 negative values in the Logistic Regression model, with 45 errors. Similarly, the K-Nearest Neighbors model has 102 right predictions, 99 wrong predictions, and 102 positive values. There are 98 valid predictions and 104 wrong ones in the Support Vector Classifier model.

The Gaussian Naive Bayes model, on the other hand, predicts just 41 out of 100 accurate outcomes. The decision tree model has 22 wrong predictions among its 112 accurate ones and 104 false ones. Finally, 110 positive values, 114 negative values, and 14 erroneous predictions are shown using the Random Forest Classifier model.

Fig.11 shows Thalach (exercise maximum heart rate) which indicates cardiovascular fitness and heart disease risk. Increasing thalach/ through activity decreases heart disease risk, whereas high levels may indicate cardiac problems. ST segment depression occurs when a ventricle is resting and repolarized. An exceptionally low ST segment trace relative to the base may cause this cardiac condition. This supports the concept that low ST Depressive disorders enhance coronary heart disease risk. A lot of ST depression is normal. The color "slope" represents the peak workout ST segment and has values of 0 for uphill, 1 for flat, and 2 for downhill. With favorable and bad coronary artery disease, the three slope groups were distributed similarly.

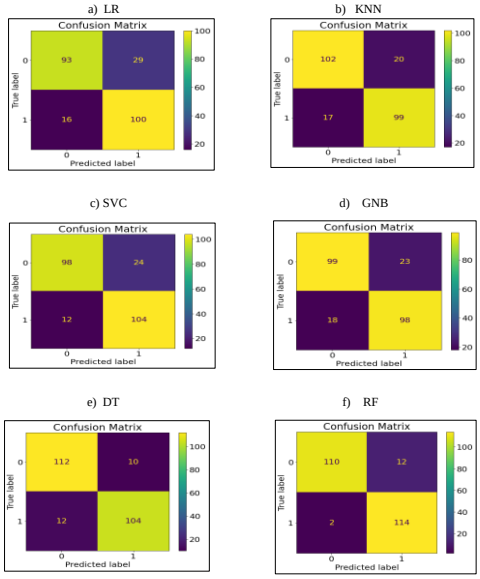
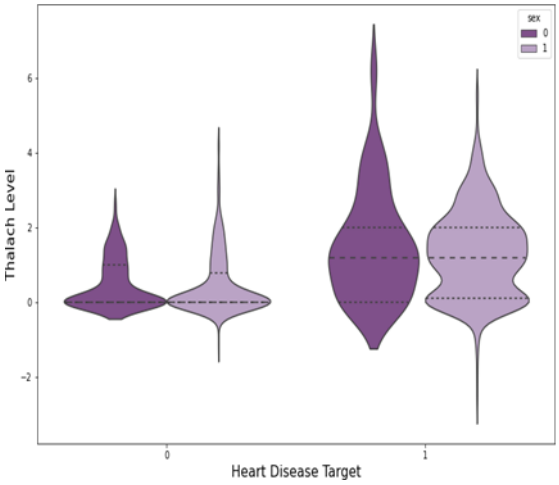
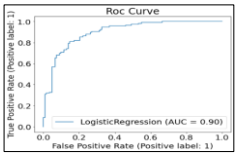


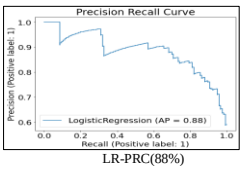
Fig 7. Confusion Matrix

The fig.7,10 shows the representation of confusion matrix, roc curve, precision recall curve. The effectiveness of the algorithm may be seen in the confusion matrix. The prediction mistakes may be quickly and graphically inspected using a confusion matrix table. In this particular investigation, the system used the following criteria for evaluation: recall, precision, accuracy and F1-score. The aforementioned metrics were determined by employing model performance estimations in the calculation process. The above fig.8,9 shows Thalach (exercise maximum heart rate) which indicates cardiovascular fitness and heart disease risk. Increasing thalach/ through activity decreases heart disease risk, whereas high levels may indicate cardiac problems. ST segment depression occurs when a ventricle is resting and repolarized.

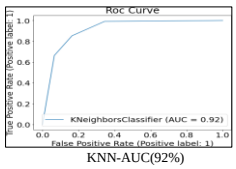




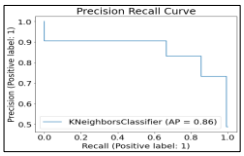
LR-AUC(90%)



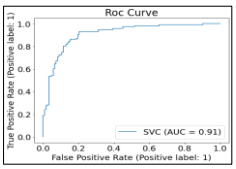
LR-PRC(88%)



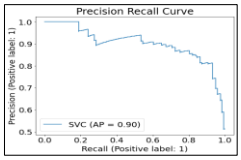
KNN-AUC(92%)



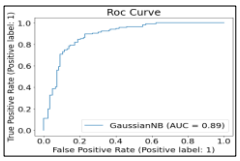
KNN-PRC(86%)



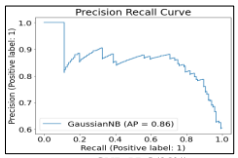
SVC-AUC(91%)



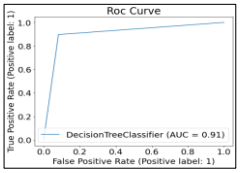
SVC-PRC(90%)



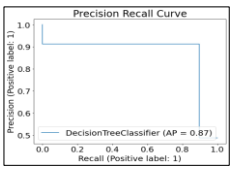
GNB-AUC(89%)



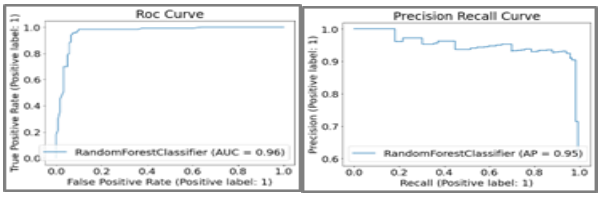
GNB-PRC(86%)



DT-AUC(91%)



DT-PRC(87%)



RF-AUC(96%) RF-PRC(95%)

Fig 10. AUC&PRC Curve

A rating of 1.0 is considered perfect in the model evaluation. Through the use of a learning curve, it's possible to determine how much additional training data will improve the model. This showcases scores for a machine learning model, considering varying numbers of training samples. The cross-validation procedure is carried out with tact while invoking the learning curve. By plotting recall on the x-axis and precision on the y-axis, it is possible

to derive the precision-recall curve. This curve presents the relationship between false positives and false negatives. The precision-recall curve is not constructed by considering the number of true negative results.

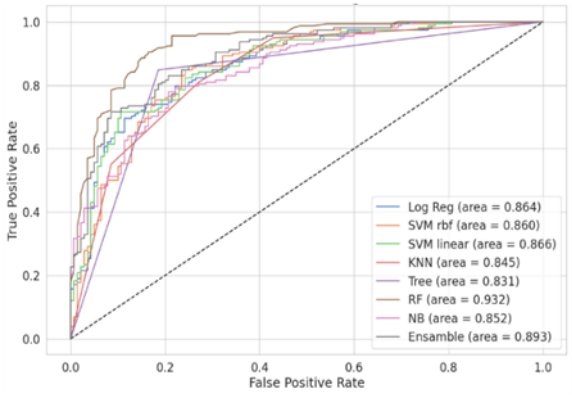
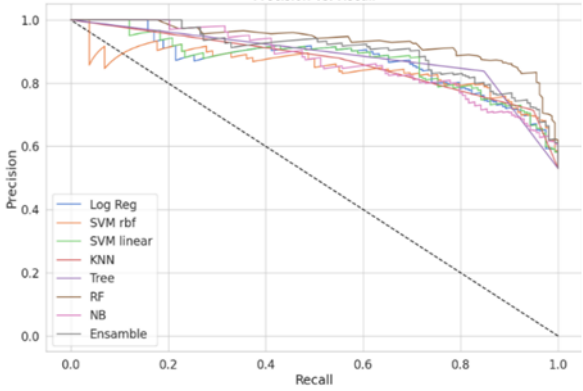
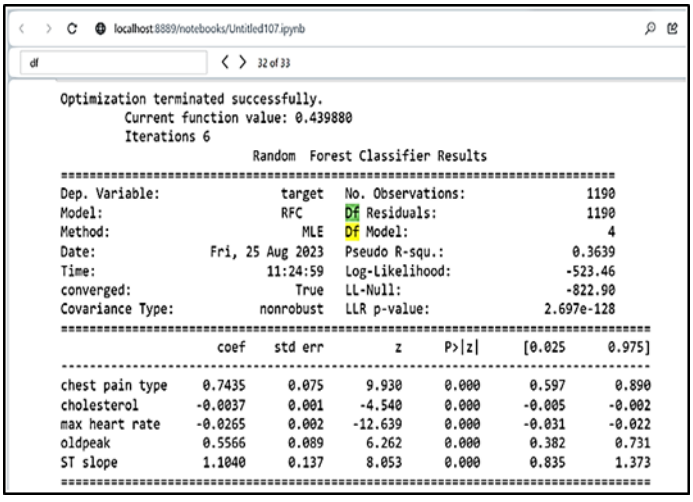


Figure.11 Radial Basis function curve



From fig.11,12 it is inferred that the mathematical approach for logistic regression curves is used for the goal of establishing the optimal cut-off value for predicting whether a new observation is a "failure" (0) or a "success" (1). This is done in order to determine the optimal cut-off value. This conclusion is arrived at by doing research on the models that were covered in the previous phase of the discussion. In order to categorize patients into those who have heart disease and those who do not have heart disease, the system had created and analyzed a linear kernel support vector machine (SVM-linear), a radial basis function (RBF) kernel support vector machine and a k-nearest neighbor (k-NN) model. All of these models were used to categorize patients. In order to accomplish the objective of accurately categorizing patients, this was carried out. It is important to keep in mind that each of these models represents a different parameter that has to be taken into consideration while analyzing all of these models. The linear kernel support vector machine model has an efficiency of 0.89, which is also its efficiency. Naive Bayes captures an area of 85%, Random Forest has a reasonable cut-off curve of 93, and the linear support vector machine (SVM) for the radial basis function kernel has a value of 0.86. All of these results are based on the radial basis function kernel.



localhost:8889/notebooks/Untitled107.ipynb

df 32 of 33

```
Optimization terminated successfully.
         Current function value: 0.439880
         Iterations 6
                         Random  Forest Classifier Results
==============================================================================
Dep. Variable:                 target   No. Observations:                 1190
Model:                            RFC   Df Residuals:                     1190
Method:                           MLE   Df Model:                            4
Date:                Fri, 25 Aug 2023   Pseudo R-squ.:                  0.3639
Time:                        11:24:59   Log-Likelihood:                -523.46
converged:                       True   LL-Null:                       -822.90
Covariance Type:            nonrobust   LLR p-value:                2.697e-128
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
chest pain type     0.7435      0.075      9.930      0.000       0.597       0.890
cholesterol        -0.0037      0.001     -4.540      0.000      -0.005      -0.002
max heart rate     -0.0265      0.002    -12.639      0.000      -0.031      -0.022
oldpeak             0.5566      0.089      6.262      0.000       0.382       0.731
ST slope            1.1040      0.137      8.053      0.000       0.835       1.373
==============================================================================
```

Fig.13 Random Forest Classifier Results

From fig.13 coefficient's standard error provides an indication of how well the model predicts the value of the unknown variable. The standard error of the coefficient is almost always significant. The standard error is a metric that may be used to evaluate the accuracy of the coefficient prediction. The accuracy of estimates is improved, and there are fewer standard deviations. It has been noticed that the cholesterol coefficient is -0.0037, which is a lesser value and the standard deviation error has been observed to be 0.001. When the z-score is positive, it means that the raw score is greater than the average. The number of total iterations is 6 with the current function value of 0.439880. Positively, the LL - Null value should be as high as possible. Low LLR p - values of 0.05 percentage points imply that the null hypothesis that the model based on the intercept is superior than the entire model may be rejected. This is because the null hypothesis may be rejected. The p-value is an indicator of statistical significance that was created to help researchers decide whether to accept or reject the null hypothesis. When the p-value is trying to figure out is the chance that there is no connection among the factors. A low p-value means that there i*s proof that the null hypothesis is false. Therefore, in this particular instance, the Z value that has positive numbers is the chest pain type, the old peak, and the ST slope. However, the p-value that is provided for the Z- statistic needs to be considered as the probability that an outcome that is just as severe or even more extreme than the one that was observed would have occurred if the null hypothesis had been true.

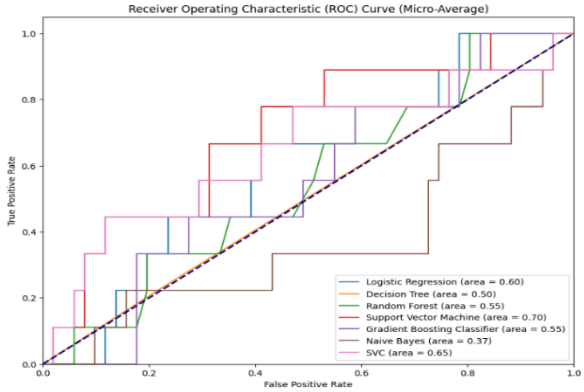


Fig 14 ROC Curve- UCI Dataset

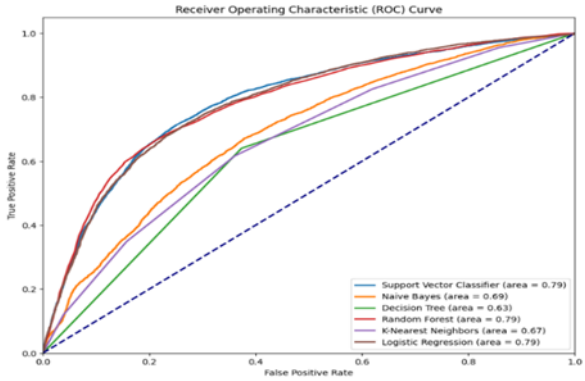


Fig 15 ROC Curve-Cleavand Dataset

In order to test how well a classification model worked, the validation study used two separate datasets. Visual representations of the resulting metrics and coverage are shown in Figures 14 and 15, allowing for a thorough evaluation of the diagnostic capabilities. In the context of predictive modeling, there is a significant difference in the performance of Support Vector Classifier (SVC) models on the UCI Heart Disease dataset and the Cleveland Heart Disease dataset. The SVC on the UCI dataset reaches an accuracy level of about 70% after extensive hyperparameter adjustment. On the Cleveland dataset, the SVC, Random Forest, and Logistic Regression models achieve a combined accuracy of 79%. This mismatch highlights the need of fine-tuning model parameters and using feature selection methods. In the case of the UCI dataset, hyperparameter modification considerably improves the SVC's effectiveness, resulting in increased accuracy and overall model performance.

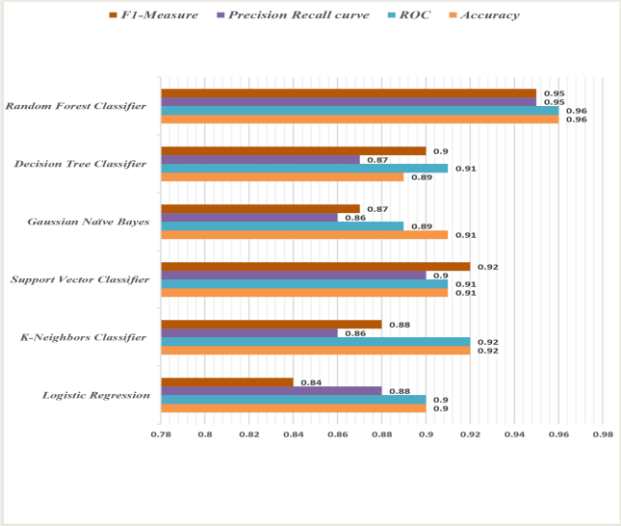


Fig. 16 Classifier Performance (Proposed Work)

Fig.16 shows learning and deployment accuracy data for machine learning classifiers. Comparing these classifiers is essential for choosing the best model, understanding algorithmic behaviour, and assuring machine learning system performance and dependability. The best prediction model is the Random Forest, with 96% accuracy across all categories. Although it ranks lower than the Random Forest, the Decision Tree Classifier has 89% Precision, Recall, and Accuracy. Gaussian Naïve Bayes classifier achieves 91% accuracy and excels in all measures. In all criteria, the Support Vector Classifier accurately detects instances with 91% accuracy. K-Nearest Neighbor's identifies events by proximity to surrounding data points with 92% accuracy and finally Logistic Regression classifier predicts outcomes based on input features with 90% accuracy.

Table 6: Comparative Analysis

Dataset	Classifier	Precision	Recall	Accuracy
UCI Heart Disease Dataset	SVC	0.85	0.8	0.82
	NB	0.79	0.82	0.81
	RF	0.86	0.78	0.8
	KNN	0.81	0.75	0.77
	LR	0.83	0.79	0.81
	DT	0.8	0.82	0.81
Cleveland Heart Disease Dataset	SVC	0.78	0.75	0.77
	NB	0.8	0.82	0.81
	RF	0.75	0.78	0.76
	KNN	0.79	0.76	0.78
	LR	0.77	0.8	0.79
	DT	0.73	0.72	0.74
Statlog, Cleavand,	SVC	0.92	0.91	0.91

Hungarian,	GNB	0.87	0.89	0.91
Swiss,	RF	0.95	0.95	0.96
VA (Proposed Work)	KNN	0.88	0.92	0.92
	LR	0.84	0.9	0.9
	DT	0.9	0.91	0.89

V. Conclusion

To achieve its basic aims, the study used a variety of research approaches. The findings highlight the need for future research into the possible advantages of integrating additional factors to improve prediction accuracy. The effectiveness of data preparation and feature selection approaches was tested by comparing the Random Forest classifier to others both before and after deployment. The implications were produced after a comprehensive study of the acquired data, with an emphasis on repeated patterns and trends. Using just existing data, the model demonstrated a very high diagnosis rate for heart disease. Future research should broaden the study's scope to include developing firms and modify components important to specific organizations in order to improve model efficacy. This study demonstrates that machine learning algorithms can effectively handle the issue of cardiovascular disease risk assessment, even when resources are restricted. The most accurate prediction model was discovered after thoroughly assessing six classification algorithms utilizing the novel Grid-Search CV. The Random Forest Classifier outperformed the other approaches, with an accuracy rate of 0.96%. To make well-informed healthcare choices, the research emphasizes the need of employing diverse techniques to assess patients' risk factors using their medical data. These findings illustrate that machine learning has the potential to change healthcare delivery and improve patient outcomes throughout the globe. The study suggests that utilizing powerful machine learning technologies to address complex healthcare issues such as cardiovascular disease may lead to more effective risk assessment and management strategies.

References:

[1] Abdollahi, Jafar, and Babak Nouri-Moghaddam. “A Hybrid Method for Heart Disease Diagnosis Utilizing Feature Selection Based Ensemble Classifier Model Generation.” Iran Journal of Computer Science, no. 3, Springer Science and Business Media LLC, May 2022, pp. 229–46. Crossref, doi:10.1007/s42044-022-00104-x.

[2] Islam, Muhammad Nazrul, et al. “Predicts “: An IoT and Machine Learning-Based System to Predict Risk Level of Cardio-Vascular Diseases.” BMC Health Services Research, no. 1, Springer Science and Business Media LLC, Feb. 2023. Crossref, doi:10.1186/s12913-023- 09104-4.

[3] Mohi Uddin, Khandaker Mohammad, et al. “Machine Learning-Based Approach to the Diagnosis of Cardiovascular Vascular Disease Using a Combined Dataset.” Intelligence-Based Medicine, Elsevier BV, 2023, p. 100100. Crossref, doi: 10.1016/j.ibmed.2023.100100.

[4] Louridi, Nabaouia, et al. “Machine Learning-Based Identification of Patients with a Cardiovascular Defect.” Journal of Big Data, no. 1, Springer Science and Business Media LLC, Oct. 2021. Crossref, doi:10.1186/s40537-021-00524-9.

[5] Muktevi Srivenkatesh. “Prediction of Cardiovascular Disease Using Machine Learning Algorithms.” International Journal of Engineering and Advanced Technology, no. 3, Blue Eyes Intelligence Engineering and Sciences Engineering and Sciences Publication - BEIESP, Feb.2020,pp.2404–14.Crossref, doi:10.35940/ijeat.b3986.029320.

[6] Ali, Farman, et al. “A Smart Healthcare Monitoring System for Heart Disease Prediction Based on Ensemble Deep Learning and Feature Fusion.” Information Fusion, Elsevier BV, Nov. 2020, pp. 208–22. Crossref, doi:10.1016/j.inffus.2020.06.008.

[7] Nashif, Shadman, et al. “Heart Disease Detection by Using Machine Learning Algorithms and a Real-Time Cardiovascular Health Monitoring System.” World Journal of Engineering and Technology, no. 04, Scientific Research Publishing, Inc., 2018, pp. 854–73. Crossref, doi:10.4236/wjet.2018.64057.

[8] Mohan, Senthilkumar, et al. “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques.” IEEE Access, Institute of Electrical and Electronics Engineers (IEEE), 2019, pp. 81542–54. Crossref, doi:10.1109/access.2019.2923707.

[9] Ram, Sigdel Shree, et al. “A Machine Learning

Framework for Edge Computing to Improve Prediction Accuracy in Mobile Health Monitoring." Computational Science and Its Applications – ICCSA 2019, Springer International Publishing, 2019, pp. 417–31, http://dx.doi.org/10.1007/978-3-030-24302-9_30.

[10] U. Umar, M. A. Khan, R. Irfan and J. Ahmad, "IoT-based Cardiac Healthcare System for Ubiquitous Healthcare Service," 2021 International Congress of Advanced Technology and Engineering (ICOTEN), Taiz, Yemen, 2021, pp. 1-6, doi: 10.1109/ICOTEN52080.2021.9493478.

[11] Karthick, T., and M. Manikandan. "Fog Assisted IoT Based Medical Cyber System for Cardiovascular Diseases Affected Patients." Concurrency and Computation: Practice and Experience, no. 12, Wiley, Oct. 2018. Crossref, doi:10.1002/cpe.4861.

[12] Ahamed, Jameel, et al. "CDPS-IoT: Cardiovascular Disease Prediction System Based on IoT Using Machine Learning." International Journal of Interactive Multimedia and Artificial Intelligence, no. 4, Universidad Internacional de La Rioja, 2022, p. 78. Crossref, doi:10.9781/ijimai.2021.09.002.

[13] M. A. Khan, "An IoT Framework for Heart Disease Prediction Based on MDCNN Classifier," in IEEE Access, vol. 8, pp. 34717-34727, 2020, doi: 10.1109/ACCESS.2020.2974687.

[14] Özcan, M., & Peker, S. (2023, November 1). A classification and regression tree algorithm for heart disease modelling and pre diction. Healthcare Analytics; Elsevier BV. https://doi.org/10.1016/j.health.2022.100130.

[15] C. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, "Effective Heart Disease Prediction Using Machine Learning Techniques," Algorithms, Feb. 06, 2023. https://doi.org/10.3390/a16020088.

[16] M. Mandava, S. R. Vinta, H. Ghosh, and I. S. Rahat, "An All-Inclusive Machine Learning and Deep Learning Method for Forecasting Cardiovascular Disease in Bangladeshi Population," EAI Endorsed Transactions on Pervasive Health and Technology, Oct. 03, 2023. https://doi.org/10.4108/eetpht.9.4052.

[17] U. Nagavelli, D. Samanta, and P. Chakraborty, "Machine Learning Technology-Based Heart Disease Detection Models," Journal of Healthcare Engineering, Feb. 27, 2022. https://doi.org/10.1155/2022/7351061.

[18] ElSeddawy, A. I., Karim, F. K., Hussein, A. M., & Khafaga, D. S. (2022, October 11). Predictive Analysis of Diabetes-Risk with Class Imbalance. Computational Intelligence and Neuroscience; Hindawi Publishing Corporation. https://doi.org/10.1155/2022/3078025.

[19] N. Nandal, L. R. Goel, and R. Tanwar, "Machine learning-based heart attack prediction: A symptomatic heart attack prediction method and exploratory analysis," F1000Research, Sep.29, 2022.https://doi.org/10.12688/f1000research.123776.1.

[20] Doki, S., Devella, S., Tallam, S., Gangannagari, S. S. R., Reddy, P. S., & Reddy, G. P. (2022). Heart disease prediction using XGBoost. 2022 Third International Conference on Intelligent Computing Instrumentation and Control-Technologies(ICICICT). https://doi.org/10.1109/icicict54557.2022.9917678.

[21] Barhoom, A. M. A. (2022). Prediction of Heart Disease Using a Collection of Machine and Deep Learning Algorithms. https://philarchive.org/rec/BARPOH-4.

[22] Khan, M. I. H. (2020.). Data-Driven Diagnosis of Heart Disease. International Journal of Computer Applications - IJCA. https://www.ijcaonline.org/archives/volume176/number41/31477-2020920549.

[23] Yuvraj Nikhate and M. V. Jonnalagedda. "Survey On Heart Disease Prediction using Machine Learning." International Journal of Creative Research Thoughts (2020) :2320-2882.

[24] J.Jeyaganesan, "Diagnosis and Prediction Of Heart Disease Using Machine Learning Techniques," Elementary Education Online, Jan. 08, 2022. https://doi.org/10.17051/ilkonline.2020.02.696765.

[25] Shah, D., Patel, S., & Bharti, S. K. (2020, October 16). Heart Disease Prediction using Machine Learning Techniques. SN Computer Science; Springer Science+Business Media. https://doi.org/10.1007/s42979-020-00365-y.

[26] J. S. Sonawane and D. R. Patil, "Prediction of heart disease using learning vector quantization algorithm," Mar. 01, 2014. https://doi.org/10.1109/csibig.2014.7056973.

[27] Vembandasamy, K., Sasipriya, R. and Deepa, E. (2015) Heart Diseases Detection Using Naive Bayes Algorithm. IJISET-International Journal of Innovative Science, Engineering & Technology, 2, 441-444.

[28] Otoom, A. F., Abdallah, E. E., Kilani, Y., & Ashour, M. (2015, January 1). Effective diagnosis and monitoring of heart disease. ResearchGate. https://doi.org/10.14257/ijseia.2015.9.1.12.

[29] Jabbar, M. A., Deekshatulu, B. L., & Chandra, P. (2013). Heart Disease Prediction System using Associative Classification and Genetic Algorithm. arXiv (Cornell University). https://arxiv.org/1303.5919.

[30] X. Li and Q. Yian-Fang, "A Data Preprocessing Algorithm for Classification Model Based on Rough Sets," Physics Procedia, Jan. 01, 2012. https://doi.org/10.1016/j.phpro.2012.03.345.

[31] Pathan, M. S., Nag, A., Pathan, M. M., & Dev, S. (2022, November 1). Analyzing the impact of feature selection on the accuracy of heart disease prediction.Healthcare-Analytics.https://doi.org/10.1016/j.health.2022.100060.

[32] Xavier, A. (2021, February 22). Heart Disease Prediction using Machine learning and Data Mining Technique. IJERT. https://doi.org/10.17577/IJERTCONV9IS03065.

[33] Patil, S., & Bhosale, S. (2023, August 31). Improving Cardiovascular Disease Prognosis Using Outlier Detection and Hyperparameter Optimization of Machine Learning Models. Revue D'intelligence Artificielle. https://doi.org/10.18280/ria.370429.

[34] Valarmathi, R., & Sheela, T. (2021, September 1). Heart disease prediction using hyper parameter optimization (HPO) tuning. Biomedical Signal Processing and Control. https://doi.org/10.1016/j.bspc.2021.103033.

[35] Prabu, S., Thiyaneswaran, B., Sujatha, M., Nalini, C., & Sujatha, R. (2022, January 1). Grid Search for Predicting Coronary Heart Disease by Tuning Hyper-Parameters. Computer Systems Science and Engineering. https://doi.org/10.32604/csse.2022.022739.

[36] Ambrish, G., Ganesh, B., Ganesh, A., Srinivas, C., D., & Mensinkal, K. (2022, June 1). Logistic regression technique for prediction of cardiovascular disease. Global Transitions Proceedings. https://doi.org/10.1016/j.gltp.2022.04.008.

[37] Jabbar, M. A., Deekshatulu, B. L., & Chandra, P. (2013, January 1). Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm.ProcediaTechnology.https://doi.org/10.1016/j.protcy.2013.12.340

[38] Elsedimy, E. I., AboHashish, S. M. M., & Algarni, F. (2023, August 5). New cardiovascular disease prediction approach using support vector machine and quantum-behaved particle swarm optimization. Multimedia Tools and Applications. https://doi.org/10.1007/s11042-023-16194-z.

[39] Ali, L., Khan, S. U., Golilarz, N. A., Imrana, Y., Qasim, I., Noor, A., & Nour, R. (2019, November 20). A Feature-Driven Decision Support System for Heart Failure Prediction Based on $\chi 2$ Statistical Model and Gaussian Naive Bayes. Computational and Mathematical Methods in Medicine. https://doi.org/10.1155/2019/6314328.

[40] Assegie, T. A., Kumar, R. P., Kumar, N. K., & Vigneswari, D. (2022, September 1). An empirical study on machine learning algorithms for heart disease prediction. IAES International Journal of Artificial Intelligence. https://doi.org/10.11591/ijai.v11.i3.pp1066-1073.

[41] Sumwiza, K., Twizere, C., Rushingabigwi, G., Bakunzibake, P., & Bamurigire, P. (2023, January 1). Enhanced cardiovascular disease prediction model using random forest algorithm. Informatics in Medicine Unlocked. https://doi.org/10.1016/j.imu.2023.101316.

[42] Ansari, G. A., Bhat, S. S., Ansari, M. D., Ahmad, S., Nazeer, J., & Eljialy, A. E. M. (2023, May 29). Performance Evaluation of Machine Learning Techniques (MLT) for Heart Disease Prediction. Computational and Mathematical Methods in Medicine. https://doi.org/10.1155/2023/81912.