A Experimental Investigation of A Generic Method for Early Detection of Heart Disease

Venkateswara Rao Cheekati[1†], **Dr. D. Natarajasivan**[2†] and **Dr. S. Indraneel**[3]

[1]Research Scholar, Department of CSE, Annamalai University, Annamalai Nagar, Tamil Nadu
[2]Asst.Professor, Department of Computer Science and Engineering, Annamalai University, Tamilnadu.
[3] Professor, Department of Computer science and Engineering, St. ANN'S college of Engineering and Technology, Chirala, AP.

Abstract

The heart is the second most important organ in the body after the brain. Any trouble in the heart will eventually cause trouble in the rest of the body. As people who live in the modern world, we experience huge changes every day that affect us in some way or another. Heart disease, which kills people all over the world, is one of the top five diseases that kill the most people. So, being able to predict this disease is very important, as it will allow people to take the right steps at the right time. Data mining and machine learning are ways to find useful information in a huge amount of data and make it better. It is the first and most important step in figuring out how to define and find useful information and hidden patterns in databases. Optimization algorithms can be used to solve complex, non-linear problems because they are flexible and can be changed. Machine learning techniques are used in the medical sciences to help solve real health problems by predicting and treating diseases early on. In this paper, we use six different machine learning algorithms and then compare them based on how well they work. With a testing accuracy of 97.29%, decision tree is the best classifier out of all the others.

Keywords:
Heart Disease, Machine Learning Models, Python, Spyder

1. INTRODUCTION

Heart disease is thought to be one of the most dangerous diseases that can happen to anyone at any time. It is a fatal disease because, on average, hundreds of people die every day from it and other similar diseases [1-3]. According to a study by the Indian Heart Association, more than 17 million people die every year from heart disease, and that number is expected to rise to 2.3 crore deaths by 2030 [4]. Heart diseases are caused by a wide range of conditions that affect how the heart works. The different Some kinds of heart disease are:

• Atherosclerosis is a heart disease that affects the heart's arteries.

*Diseases of the heart's valves can cause problems with the pumping action of the heart and the regulation of blood flow.

*Conditions like cardiomyopathy can alter the normal contraction of heart muscle.

*The heart's electrical conduction can be disrupted by arrhythmias.

*Infections of the heart and structural defects in the heart can occur before birth.

• Coronary arteries deliver blood to the heart, and when there are too many lipids in the blood, a cholesterol plaque forms inside the artery walls and narrows the passageway, leading to coronary artery disease. When a cholesterol plaque ruptures and forms a clot in the arteries, blocking blood flow, it causes a heart attack, the most common type of cardiovascular disease.

A person's mental and emotional well-being directly affects their actions in both their personal and professional lives. Since cardiovascular disease is on the rise, it is important to develop a model that incorporates symptoms and dietary habits to predict the onset of heart disease with minimal effort. The following characteristics may be present in people who are at risk for cardiovascular disease:

*Greater Quantity of Cholesterol.

• Hypertension, or high blood pressure.

• Smoking

• A large amount of lipids.

Excessive body fat storage is a problem.

Manuscript received September 5, 2022
Manuscript revised September 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.9.87

- A history of cardiovascular disease in the family.

Predicting and diagnosing cardiovascular disease is possible through the use of multiple attributes, such as age, sex, cholesterol, resting blood pressure, etc., and analysing them in a way that permits experts to make better and more accurate knowledge-based results. This explosion in data has sparked a new field called Machine Learning, which can glean insights from massive datasets. Data mining's purpose in medicine is to aid in the identification and treatment of actual medical conditions. This paper aims to do just that, comparing the accuracy of various machine learning algorithms that make use of various tools and techniques. It also emphasises the potential of a healthcare prediction model based on data mining and extensive machine learning analysis in the future. IPSOS [5] reports that when compared to other leading causes of death, heart disease has the highest percentile death rate. Figure 1 shows the number of deaths expected to result from other diseases in 2019.

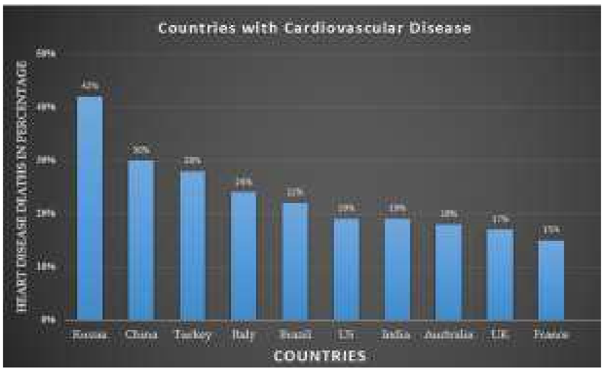


Figure 1 shows the top ten countries in terms of the prevalence of cardiovascular disease. [5]

2. LITERATURE SURVEY

There has been a lot of effort put into predicting heart disease using Machine Learning, Deep Learning, and Data mining techniques. A variety of datasets, algorithms, and methods have been employed by researchers, with results observed and future work carried out in the pursuit of efficient medical diagnostic methods for Cardiovascular disease. For the past decade, studies into improving CVD diagnosis through the use of prediction models have been ongoing. Diagnosing CVD automatically is an important clinical challenge. Better treatment is possible if heart disease is caught early on. Only by implementing Machine Learning and Deep Learning, the foundations of predictive analytics, will this goal be realised. Researchers from all over the world have tried out different methods for diagnosing and predicting cardiovascular illness. These are among the most common methods:

Authors Jaymin Patel et al. (2015): The study's intended outcome is the development of a model for predicting cardiovascular diseases. The training data underwent preprocessing classification, regression, clustering, association, and visualisation before being used in the experiment, which was conducted in WEKA using the Cleave Land UCI dataset. It is common practise to employ WEKA's Explorer mode when evaluating classifiers. The analysis made use of a 10-fold cross-validation and reduced-error pruning to further refine the performance of decision tree classifiers such as J48, Logistic Model Tree Algorithm, and Random Forest. The best performance was by j48 with less aggressive error pruning. If alternative discretization methods, multiple classifiers, voting procedures, and other decision tree Algorithms (Gain ratio, Gini index) had been considered, the accuracy would have been improved. The lack of proper combination and more complex models prevents greater sensitivity, specificity, and accuracy from being attained. Algorithms developed in this study can be used to create a system for cross-referencing data mining models in the future [6].

Sonam Nikahr et al. (2016): Through the removal of superfluous data points and attributes, this research aims to improve the accuracy of Naive Bayes classifiers. eliminating all but the most illuminating ones. Cleave Land Heart Disease Database data included 303 records and 76 attributes, from which 19 were used for analysis. It does so by employing a Selective Nave Bayes classifier, in which c4.5 trees are built. Nave Bayes and Decision Tree, when combined with Information Gain calculations, outperform alternative classifiers in this study. The decision tree, which is more accurate than Nave Bayes, is induced using the greedy algorithm [7].Based on the work of Syedahamin Pouriyeh et al. (2017) In this paper, we intend to do just that: compare a variety of machine learning methods using a sample dataset. The research used a dataset culled from the cleave land database, which included 303 instances and a maximum of 76 attributes. Since the variance of the 10-fold cross-validation method is lower than that of other estimators, such as the single-fold method, it has been used to divide up datasets. Decision trees, Naive Bayes, Multi-Layer Perceptrons, K-Nearest Neighbors, Single Conjunctive Rules, Radial Basis Functions, and Support Vector Machines are just some of the many machine learning classifiers put to use. The aforementioned machine learning methods have been used in conjunction with bagging, boosting, and stacking. The primary metrics of performance in this analysis are the areas of Precision, Recall, F-measure, and ROC. We split the experiment into two parts; in the first, we used the entire dataset and applied machine learning algorithms with 10-Fold cross-validation. Based on the findings, SVM performed with the highest accuracy of any method tested (89.12 percent). An experiment involving Bagging, Boosting, and stacking was performed in the second case. To improve DT's accuracy, we used bagging, and it went from 77.55% to 78.54%. The accuracy of DT went up from 77.55% to 82.17% after the boost. In spite of this, an accuracy of 84.15% was achieved by stacking multiple SVMs together. According to Youness Khourdifi et al. Machine Learning Algorithms optimised with particle swarm and ant colony techniques were used to propose heart disease prediction models. A technique known as Fast Correlation Based Feature Selection (FCFS) has been used to improve the quality of heart disease classification by removing unnecessary features. A wide range of algorithms, including KNN, SVM, NB, RF, and MLP (Artificial Neural Networks), as well as hybrid methods like Particle Swarm Optimization and Ant-Colony Optimization, are used for classification (ACO). The information in the dataset comes from the UCI machine learning repository. WEKA is used to classify the new features after a training dataset is prepared using a binary type classification problem. The proposed hybrid approach is used to process the data set. FCBF, PSO, and ACO's optimised model had 99.65% precision. As a result, the hybrid model outperformed the other methods of classification considered. The author's expertise, the study's tools, and the time frame all act as constraints [9].Alex Mamta et al. This research aims to establish a system for early disease detection and accurate disease forecasting. In this article points out the shortcomings of data mining algorithms, i.e. the diagnostic accuracy is better than the predictive accuracy. precise, albeit laborious. The data comes from "jubilee Mission College And Research Thrissur," and it was gathered through interviews and records of patients' discharges from a hospital for heart disease. We have collected 2200 records with 20 attributes and sorted them into a usable format. SVM, RF, KNN, and ANN were all fed the attributes; however, the best accuracy (92.21%) was achieved by KNN. [10]

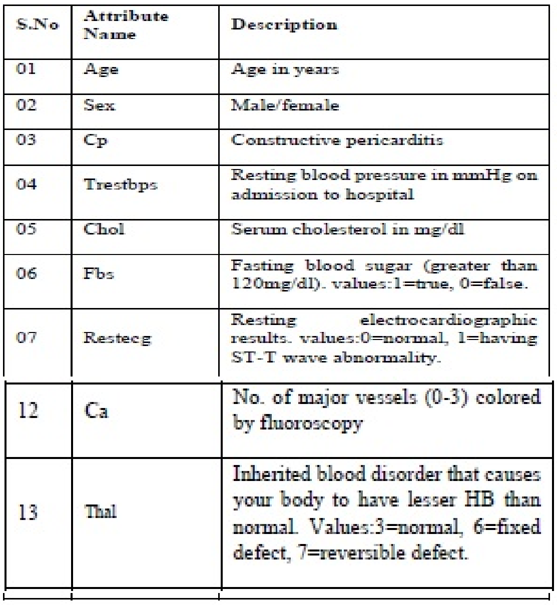
3. MACHINE LEARNING TOOLKIT

Anaconda 2020[11], a free and open-source programme, has been used to process all of the experimental data. Python (version 3.7.6) and Spyder, an integrated development environment (IDE), are also used for various programming-related tasks and evaluations. Additionally, to pandas, a number of other standard packages are used for importing databases, preprocessing data, and running machine learning algorithms on the dataset.

4. DESCRIPTION OF THE DATASET

The data was downloaded from the machine learning repository at UCI. Two datasets, one with 1026 instances and 14 attributes and the other with 303 instances and 14 attributes, have been collected. After merging, the resulting dataset has 1329 instances and 14 characteristics. Table 1.2 details the attributes being described.

Table 1: Description of Attributes of Heart Diseases



S.No	Attribute Name	Description
01	Age	Age in years
02	Sex	Male/female
03	Cp	Constructive pericarditis
04	Trestbps	Resting blood pressure in mmHg on admission to hospital
05	Chol	Serum cholesterol in mg/dl
06	Fbs	Fasting blood sugar (greater than 120mg/dl). values:1=true, 0=false.
07	Restecg	Resting electrocardiographic results. values:0=normal, 1=having ST-T wave abnormality.
12	Ca	No. of major vessels (0-3) colored by fluoroscopy
13	Thal	Inherited blood disorder that causes your body to have lesser HB than normal. Values:3=normal, 6=fixed defect, 7=reversible defect.

5. MOTIVATION FOR THE STUDY

The primary goal of this study is to create a model for detecting the onset of cardiovascular disease. This study also seeks to determine the most effective classification algorithm for predicting the aforementioned illness. This research is validated by a comparative analysis of six algorithms: Logistic Regression, Decision Tree, Random Forest, Naive Bayes, Support Vector Machine, and Artificial Neural Network. These algorithms have been applied to a wide range of data and analysed in a variety of ways. The findings of this study will help researchers zero in on the most accurate and effective method for predicting cardiovascular disease.

6. PROPOSED METHODOLOGY

The complete experimental procedure, from data collection to the production of results, is shown in Figure 2 below. First, data is gathered from the aforementioned locations, and then pre-processing occurs. Pre-processing the data helps eliminate bias and removes noise. After the data has been cleansed and organised, it is split into a training set and a test set. Separately, training and testing data are put through machine learning algorithms. This procedure concludes with the production of accuracy-related results that are compared across various machine learning algorithms.

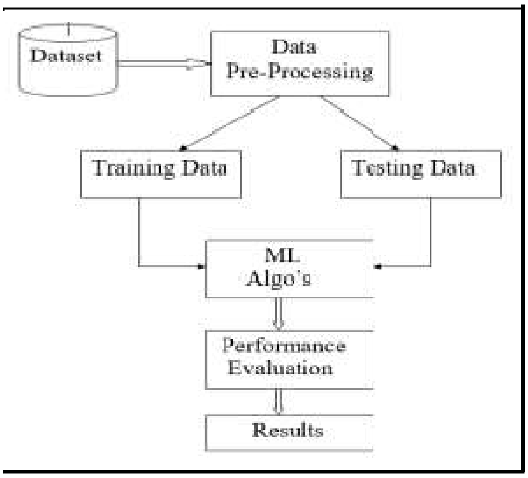


Figure 2: The Workflow of the Experimental Study

7. MACHINE LEARNING ALGORITHM

Machine learning algorithms have great potential in healthcare because they can process much larger and more diverse datasets than humans can. Better outcomes, lower healthcare costs, and higher patient satisfaction can be achieved when data is transformed into clinical insights that aid physicians in planning and providing care. Some of the algorithms used in this study are described below.

1.) Regression Analysis Using Logistic Models [12] It's a type of regression analysis in statistics, and it's widely employed for making forecasts and calculating probabilities. With the help of a set of independent variables, a prediction can be made about a binary dependent variable belonging to a categorical variable. This requires adjusting the data to fit the following equation:

$$\mathbf{Y = \beta 0 + \beta 1x1 + \beta 2x2 + \ldots + \beta nxn} \qquad \mathbf{(1)}$$

2) Nodes at the ends of branches make up the "leaves" of a decision tree, making it a flowchart-like diagram [13]. It has become well-known for the discovery of knowledge in general, rather than in any specific domain, thanks to this property. Decision trees are of the j-48 type, which allows for the generation of both pruned and unpruned trees during the classification process. This algorithm eventually comes to be able to deal with both discrete and continuous features [14]. Down below in figure 3[15] is the basic decision tree.

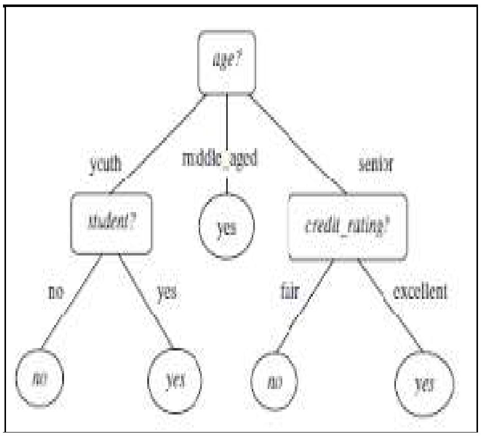


Figure 3: The Decision Tree

3) There is a mathematical model called artificial neural networks that is based on the incredibly basic neuron in the human brain. Nodes in an ANN represent individual processing units, and the lines connecting them serve as the network's communication channels. Each layer of a neural network, beginning with the input layer, represents a different predicate variable [16]. The input layer's nodes are linked to those of the hidden layer, and the hidden layer's nodes are linked to those of the output layer. Figure 4[17] depicts a basic ANNs.

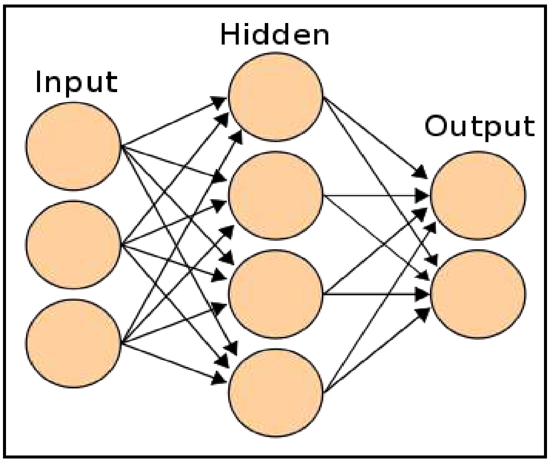


Figure 4: The Artificial Neural Network

4) The Naive Bayes Classifier, which uses the conditional probability rule and Bayes' theorem [18] as its foundation. All attributes in a dataset can be extracted independently for use and analysis [19]. According to [20], a conditional probability measures how likely it is that some conclusion C is to be correct given some evidence or observation E.

$$P(H|E) = \frac{P(E|H).P(H)}{P(E)}$$

5) In contrast to the decision tree, the accuracy of predictions made with Random Forest is higher because the algorithm uses a larger number of trees. Both trees are trained separately, and then their predicted values are averaged. A forest tree classifier's generalisation error is proportional to the significance of each tree in the forest and to the strength of their correlation [21]. Figure 5 below from [22] depicts a basic random forest.

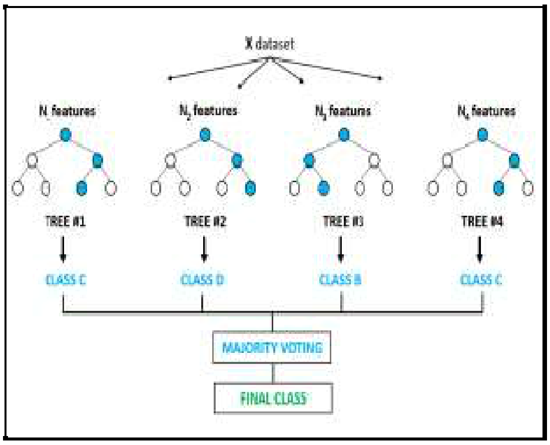


Figure 5: The Artificial Neural Network

6) Support Vector Machine (SVM): SVM is a supervised machine learning algorithm that uses a hyper plane to categorise and partition the heart disease dataset [23]. As a result of its memory-intensive nature, SVM is notoriously difficult to interrupt and tune. This supervised classifier's primary goal is to properly categorise the provided data points in a dimensional space. The SVM classifier makes use of various kernels, such as the linear kernel, polynomial kernel, and radial basis function kernel, to function in high and complex dimensions. Multiple kernel equations are given in [24].

8. MEASUREMENT AND COMPARATIVE ANALYSIS

In this study, we developed an experimental and analytic framework, and then tested and compared six different machine learning algorithms for making accurate predictions about cardiovascular disease. Figure 6 below displays the resulting histogram, showing the plots across which each attribute of the dataset is distributed.

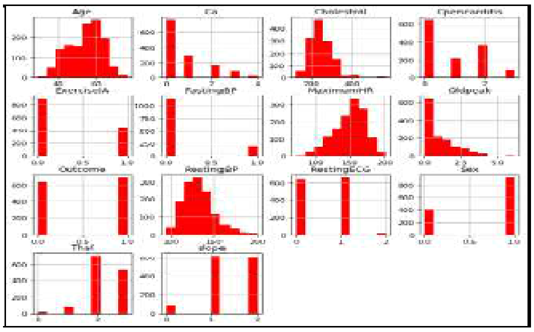


Figure 6: Histogram for each attribute in the dataset.

As shown in the Heat Map format of Figure 7, below, the correlation between these specific parameters has been used to predict the occurrence of heart disease.

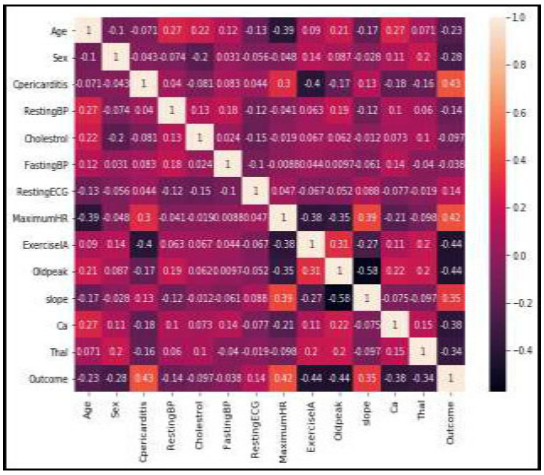
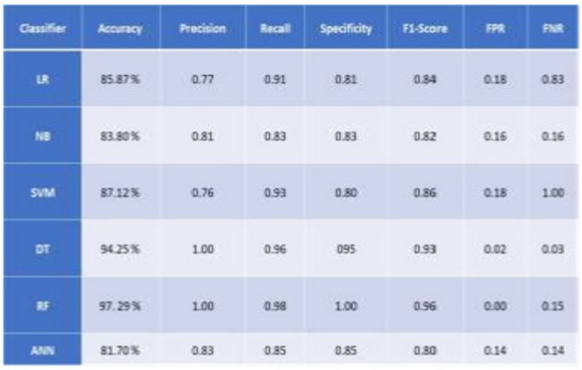


Figure 7: The Heatmap of parameters.

After that, we'll implement these six classification algorithms and see what the statistical results are. The effectiveness of these models is measured using the confusion matrix. Each model's final accuracy was determined using the following metrics.

Accuracy, Recall, Specificity, F1-score, False-Positive Rate, False-Negative Rate, and Negative Predicted Value can all be seen in Table 2 below.

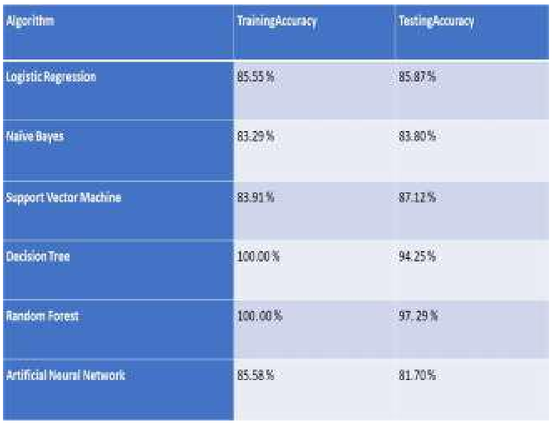
Table 2: The measurement evaluation of ML models



Classifier	Accuracy	Precision	Recall	Specificity	F1-Score	FPR	FNR
LR	85.87%	0.77	0.91	0.81	0.84	0.18	0.83
NB	83.80%	0.81	0.83	0.83	0.82	0.16	0.16
SVM	87.12%	0.76	0.93	0.80	0.86	0.18	1.00
DT	94.25%	1.00	0.96	095	0.93	0.02	0.03
RF	97.29%	1.00	0.98	1.00	0.96	0.00	0.15
ANN	81.70%	0.83	0.85	0.85	0.80	0.14	0.14

The results of six Machine Learning Algorithms used for cardiovascular disease prediction are shown in Table 3 below. The testing accuracy of the random forest is 97.29%, which is significantly higher than the testing accuracy of the decision tree (94.25%).

Table 3: The Accuracy of ML models.



Algorithm	TrainingAccuracy	TestingAccuracy
Logistic Regression	85.55%	85.87%
Naive Bayes	83.29%	83.80%
Support Vector Machine	83.91%	87.12%
Decision Tree	100.00%	94.25%
Random Forest	100.00%	97.29%
Artificial Neural Network	85.58%	81.70%

Below in Figure 8, we see a visual representation of the accuracy (in both Training and Testing) of six different machine learning algorithms.

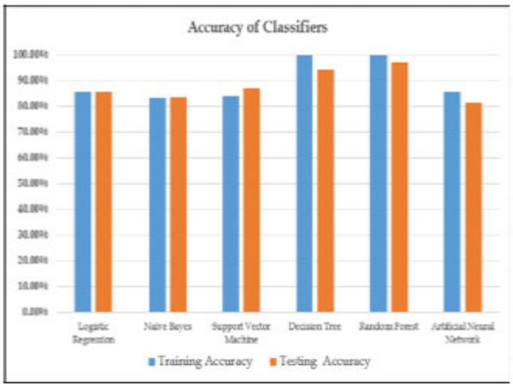


Figure 8: Graphical representation of Accuracy

In figure9 below, we see the receiver operational curve (ROC). Relative Operating Characteristic (ROC) shows how well a model performs across a range of cutoffs. Two values, True Positive RATE and False Positive RATE, are plotted on the curve. Increasing the number of False Positives and True Negatives occurs when the classification threshold is lowered, making more items Positive.

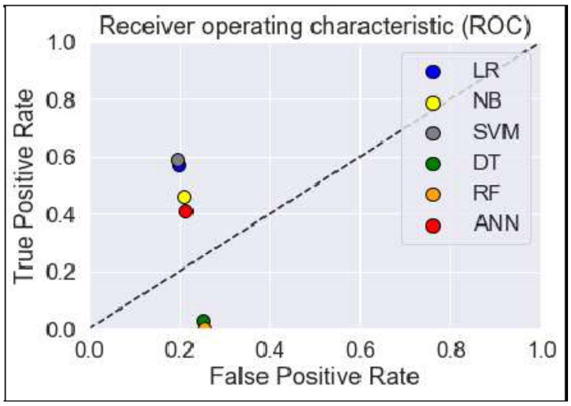


Figure 9: The measurement evaluation of ML models

9. CONCULSION AND FUTURE SCOPE

The primary goal of the study, or at least an attempt to do so, was to compare different ML algorithms in order to better understand how they can be used to predict the onset of heart disease. The elimination of outliers and the treatment of corrupted and missing values in pre-processing procedures remained the primary focus in the quest to enhance data-set quality. But we also suggested a block diagram approach to model creation in machine learning. These models can categorise a person's propensity for developing (heart) disease based on the parameters of the proposed dataset. Experimental results reflected accuracy of Random Forest for our dataset is (100% for training set) and (97.29% for the testing set), which is expected to be the highest among all classifiers, after we ran six different machine learning algorithms to predict the disease and compared the results with various statistical measures. To test the reliability of various machine learning algorithms, a 10-fold cross-validation procedure was used (LR, NB, SVM DT, RF, ANN).

Statistics such as accuracy, recall, specificity, F1-score, false-positive rate, false-negative rate, and negative predicted value were all improved with the aid of the RF model. One gets the impression that in the future, the ML classifiers will need to be trained and tested with large data-sets in order to determine the extent to which they can improve the prediction of this disease. Throughout the entire research process, from obtaining the data to presenting the findings, more work could be done to improve the quality of this Study. Hybridization or ensemble approaches are being explored in current research for their potential to improve efficiency, reliability, and validity, all of which are crucial for saving human lives at an early stage.

References

[1] Sikkandar, Mohamed Yacin. "Design a Contactless Authentication System Using Hand Gestures Technique in COVID-19 Panic Situation." Annals of the Romanian Society for Cell Biology (2021): 2149-2159.

[2] Behera, Santosh K., Pradeep Kumar, Debi P. Dogra, and Partha P. Roy. "A Robust Biometric Authentication System for Handheld Electronic Devices by Intelligently Combining 3D Finger Motions and Cerebral Responses." IEEE Transactions on Consumer Electronics 67, no. 1 (2021): 58-67.

[3] Babu, Sarath, "Heart disease diagnosing using data mining technique." Electronics Communication and Aerospace Technology (ICECA),2017 International conference of vol.1.IEEE,2017

[4] MissChaitrali S. Dangare, Dr.Mrs. SulabhaS.Apte, A data mining approach for prediction of heart disease using neural networks, international journal of computer engineering and technology, 2012.

[5] https://www.statista.com/statistics/1108824/cardiovascular-as-cause -of-death-estimate-and-actual-worldwide/

[6] Sikkandar, Mohamed Yacin. "Design a Contactless Authentication System Using Hand Gestures Technique in COVID-19 Panic Situation." Annals of the Romanian Society for Cell Biology (2021): 2149-2159.

[7] SonamNikahr, A. M.Karandikhar, "Prediction Of Heart Disease Using Machine Learning Algorithms", International Journal Of Advanced Engineering, Management and Science, vol-2, June-2016.

[8] SeyedaminPouriyeh, Sara Vahid, Giovanna Sannino,"A Comprehensive Investigation and Comparison Of Machine LearningTechniques In The Domain Of Heart Disease", 22nd IEEEsymposium on computers and communication(ISCC 2017):workshop-ICTS4EHealth 2017.

[9] YounessKhourdifi, Mohamed Bahaj, " Heart Disease Prediction andClassification Using Machine Learning Algorithms Optimized By Particle Swarm Optimization and Ant Colony Optimization", international journal of intelligent engineering and systems, vol-12, No-1,2019.

[10] Mamta Alex P and Shaicy P Shaji,"Prediction and diagnosis of heart disease patients using data mining techniques", International Conference on Communication and Signal Processing, April 4-6,2019.

[11] Anaconda Inc., "Anaconda Distribution," Anaconda, 2019, [Online]. Available: https://www.anaconda.com/distribution/

[12] Mythili T., Dev Mukherji, Nikita Padalia, and Abhiram Naidu," A Heart Disease Prediction Model using SVM-Decision Trees-LogisticRegression (SDL)", International Journal of Computer Applications (0975 – 8887) Volume 68– No.16, April 2013.

[13]]Larose,D.," Discovering knowledge in data: an introduction to datamining''. New jersey: john wiley&sons, inc,2005.

[14] Umair Shafique, Faiz Majeed, Haseeb Qaiser,and Irfan UlMustafa,"Data Mining in Healthcare for Heart Disease", internationaljournal of innovation and applied sciences, vol.10 No.4 Mar,2015,pp.1312-1322

[15] Umair Shafique, Faiz Majeed, Haseeb Qaiser, and Irfan Ul Mustafa,"Data Mining in Healthcare for Heart Disease", internationaljornal of innovation and applied sciences, vol.10 No.4 Mar 2015,pp.1312-1322.

[16] https://www.analyticsvidhya.com/blog/wp-content/uploads/2014/1 0/ANN.png]GeorgeJonh and Pat Langley. "Estimating continuousdistributions in Bayesian classifiers" in proceedings of the Eleventh Conference OnUncertainty in Artificial Intelligence, Morgan Kaufman, pages 338-345,1995.

[17] Umair Shafique, Faiz Majeed, Haseeb Qaiser, and Irfan Ul Mustafa,"Data Mining in Healthcare for Heart Disease", internationaljornal of innovation and applied sciences, vol.10 No.4 Mar 2015,pp.1312-1322.

[18] Sonam Nikhar and A.M. Karandikar,"Prediction Of Heart Disease Using Machine Learning Algorithms", International Journal Of Advanced Engineering, Management and Science, vol-2,issue-6, June-2016.

[19] H.Benjamin Fedrick David and S. Antony Belcy, "Heart Disease Prediction Using Data Mining Techniques"

[20] https://miro.medium.com/max/1170/1*58f1CZ8M4il0OZYg2oR N4w.png

[21] D.K. Srivastava and L. Bhambhu, "Data classification using support vector machine," J.Theor.Appl.Info.Technol.,2009

[22] A. A. Abdillah and Suwarno, "Diagnosis of diabetes using support vector machines with radial basis function kernels," Int. J. Technol., vol. 7, no. 5, pp. 849–858, 2016, DOI: 10.14716/ijtech.v7i5.1370.