MACHINE LEARNING FOR HEALTH CARE SYSTEM: A PREDICTIVE ANALYSIS OF HEART DISEASES

P. Sripalreddy [1], Rashi Agrawal [2]

[1] Research Scholar, Department of Computer Science Engineering, Chhatrapati Shahu Ji Maharaj University, Kanpur.
Email: sripal.patlola@gmail.com

[2] Research guide, Department of Computer Science Engineering, Chhatrapati Shahu Ji Maharaj University, Kanpur.
Email: dr.rashiagrawal@gmail.com

ABSTRACT

Worldwide, machine learning (ML) is applied in the healthcare industry. In the medical data set, ML techniques aid in the prevention of cardiac conditions and motor impairments. Finding such crucial information gives researchers important new understanding on how to apply their diagnosis and treatment for a specific patient. To help medical professionals forecast diseases, researchers analyze vast volumes of complex healthcare data using a variety of Machine Learning techniques. We are using an open UCI dataset with 303 rows and 76 attributes for this study. Of these 76 qualities, about 14 are chosen for testing, which is required to verify how well various approaches work. The isolation forest method standardizes the data for increased accuracy by utilizing the most important attributes and metrics from the data collection.

Keywords: Heart Disease, Health Care, Machine Learning, Naive Bayes, Decision Tree Classifier, SVM, K-Nearest Neighbor, Logistic Regression, Random Forest.

1.0 INTRODUCTION

Machine learning is a technique used in data science study to learn from past research experiences. The traditional approaches run into a number of problems in trying to address every topic raised by different scholars. Data assessment serves the objective of assessing real models, which calls for the application of a stable, strong, and reliable framework, including such Machine Learning techniques. After discovering the fundamental patterns in the data, the machine learning approach prefers to work with immediate input from the training sample [1]. The entire training step may result in the creation of an automatic framework. The suggested system works well with both dynamic and static datasets. One result of the training and testing phase is a data prediction. A testing phase makes use of certain data gathering, such as training datasets,[2].

Over the past decade, heart disease (cardiovascular) has been the leading factor of fatality globally. According to the World Health Organization, about 23.6 million people die primarily as a cause of cardiovascular disorders, with coronary heart disease and cerebral stroke accounting for 82 percent of these deaths. The ML methods are more precise and efficient than other methods without any human assistance [3].

The ML model mainly takes input data, i.e., any text, images. After that, split the dataset into two sections: training dataset and testing dataset. The training model is created by using a training dataset. Then we can apply the testing dataset over the trained model. This trained model will produce the results, i.e., outcome A, B, and C. There have been several applications of Machine Learning methods to everyday life. The high dimensional of records is a widely known issue in Machine Learning; the sets of data researchers utilize enormous amounts of information, and researchers occasionally cannot see it even in 3D, known as the dimensional curse [4]. A diversity of Machine Learning methods is used in the classification and forecasting of various cardiovascular based diseases. An automatic classification model that distinguishes betweenpatients with heart disease who may be at extremely high risk and those at a low level of risk can also be usedto identify them [5].

Higher blood pressure, diabetes, high cholesterol, obesity, tobacco, and a family history of heart disease are potential causes. According to the research [4], the significant factors which cannot be modified are gender, hereditary factors, and age. Another factor in this data source is thalassemia, which is determined byhereditary factors. Other factors include higher blood pressure, tobacco regularity, high blood cholesterol,lack of physical activity, overweight, physical sickness, stress level, alcohol consumption, and an irregular diet [6].

The proposed study of article [7] makes significant contributions by developing machine-learning-based health care innovative methods for predicting heart disease. Various ML prediction methods, i.e., regression models, Random Forest, SVM, Decision Tree Classifier, KNN, and Naive Bayes, were utilized in theresearch [8] to classify the patients, i.e., no heart disease(healthy) and with heart disease(unhealthy). All the relevant and interrelated functions that significantly affect the anticipated significance were determined using minimal redundancy maximal relevance, shrinkage, relief, and selection operators. Techniques of cross- validation, i.e., the k-fold validation method, were used. Different performance metrics, i.e. precision, F1- Score, AUC, and recall, were determined to measure and analyse the efficiency of the various classification algorithms [9].

The following are the critical contributions created by the research proposal:

- This study thoroughly investigated the accuracy, precision and processing performance of many ML classification methods.

- The classifiers' effectiveness is evaluated using k-fold cross-validation by selecting efficient feature selection methods (FSM), i.e., filter, wrapper.

- The classifiers' effectiveness is evaluated using k-fold cross-validation by selecting efficient feature selection methods (FSM), i.e., filter, wrapper.

- The research will also reveal which feature extraction method can be used to develop a massive learning algorithm for diagnosing heart disease using classification models.

- The proposed forecasting system will also remove the various anomalies and missing values from the heart disease dataset.

- The proposed system will also find efficient and precise data pre-processing models with better accuracy.

The research article contains various sections. These sections mainly cover literature review, the challenges in current work, existing Machine Learning methods and critical features, description of heart disease data sets, proposed model, implementation process, simulation results, result comparisons, and, finally, the conclusions.

2.0 LITERATURE REVIEW

Heart disease is the most serious illness in the modern era. Healthcare professionals must diagnose heart illnesses early in order to save lives and keep their patients from developing the condition. To predict every case of heart disease in the dataset, a number of classification techniques were assessed [10]. Many other characteristics, such as fatigue, lipid swings, renal problems, and hypertension, make. For many years, a variety of data analytics methods were put forth to assist medical practitioners in identifying the majority of the initial signs of heart disease [11]. The literature review summary and the efficiency level of Ml Algorithms, which are analyzed to identify the research gap, are displayed in Table 1. Additionally, it helps in pinpointing crucial details in the database on heart disease and forecasting enhanced.

Table 1. Comparison of various existing ML methods

Reference	Method	Key Findings	Dataset	Challenges
[12]	Feature selection algorithm (FCMIM), SVM	Improved accuracy results for heart disease dataset	Heart disease dataset (Cleveland)	Perform better onlyfor small dataset
[13]	Machine Learning,iPSCs and omics	Better throughput	Clinical dataset of heart disease	Only a few parameters wereimplemented.
[14]	Hybrid Machine Learning, Hybrid Random Forest	Better accuracy (87.8 %)	Cardiovascular disease dataset	Limited features
[15]	Machine Learning method, i.e. logistic regression, SVM	Accuracy (85%), sensitivity (89%), and Specificity (81 %)	Kaggle Online dataset	Accuracy can be improved.
[17]	Machine Learning methods	Higher Accuracy	Locomotor disorders, Heart diseases dataset (UCI)	Limited features dataset.
[18]	Machine Learning Models (KNN, Logistic regressionNeural Network)	Improved precision, F1-score	UCI Online Heart (Cleveland) disease dataset	Limited features
[19]	Machine Learning Methods	Higher Precision, and Accuracy	Online dataset	Limited features
[20]	IoT, Machine Learning Methods, SVM	Accuracy (97.5 %)	Heart Disease dataset	Limited features
[21]	Deep Learning, Neural Network	Better precision	Online UCI Dataset	Limited dataset
[22]	The statistical model X2 wasused with DNN and ANN	Predictions were aligned using clinical data parameters	UCI dataset	Accuracy and timecan be improved
[23]	Various Machine Learning classification techniques and Principal component analysis have been used to anticipate heart disease	Better measurementsand select characteristics results	Hungarian-Cleveland data	Dimension issues, accuracy
[24]	The feature extraction wasdone with an Adaboost classifier anda PCAcombination	Prediction accuracy improved	Online clinical dataset	Time and accuracy
[25]	SVM, KNN	Accurate prediction ofheart disease	UCI dataset	Precision can be improved
[26]	Naive Bayes and SVM wereused as classifiers	Classification of heart disease dataset, causeof heart disease, diabetes	Kaggle dataset	Features selection and classification performs slower
[27]	Naive Bayes, Decision Tree Classifiers, RF, Neural Networks	Classify heart diseases, better accuracy results	Online dataset UCI	Accuracy can be improved
[28]	k-NN algorithm	Feature selection, Classification	Kaggle dataset	Feature Categorization can be improved.

A variety of ML methods were used to classify and anticipate the heart disease source data, including Stochastic Gradient Descents method, KNN, Naive Bayes method, Support Vector Machine method, Ada boost method, Decision Tree Classifier method, J48 method, JRIP method, and some others [12].

In the research article [13], the researcher suggested an efficient method to findout the presence of heart disease utilizing the Back-Propagation (BP) feature extractor of ANN on available online heart disease database categorizations. In research [14], ML learning methods with ANN were used to predict heart disease cases. In this work, researchers have created an automation application to anticipate the sensitivity of a heart disease trained by using some primary symptoms, i.e., disease period, gender, heartbeat rate, and history. The outcomes of the ANN model were shown as the most accurate and precise algorithm for the forecasting of heart disease contrasted to various ML models.

In the research [15], a hybrid algorithm was recommended to accurately predict and classify the risk of heart disease in various age groups. This research introduced an automation model for precise data analysis using a neural network. The proposed model predicts heart conditions inside the initial stages accurately. It has been demonstrated that assessing an individual's threat quality by utilizing ML techniques, i.e., DT, KNN, NB, and GA, can be more significant when using characteristics and variations of the methods mentioned above.

The researcher [16] utilized a dimension reduction procedure to pre-process the data source and eliminate the values and outliers. The dataset has fourteen disorder characteristics, but researchers used four variables for performance monitoring: Confusion Matrix, Sensitivity, Specificity, and Accuracy. The researchers obtained the highest classification precision of 93%, better than earlier described classification methods.

3.0 METHODOLOGY

Supervised and unsupervised learning are subcategories of Machine Learning and AI. These methods help in the training process. It mainly utilized two classes of the dataset, i.e., labelled, unlabeled. This research presents an automation ML model to predict and classify heart disease [29]. Figure 1 shows the architecture of the proposed ML-based model; it includes phases of data pre-processing, assessment, feature selection, model training, testing, and comparing outcomes.

3.1 DECISION TREE CLASSIFIER

It mainly represents a process incorporating a tree-like growth model of pronouncements and their promising outcomes, including occurrence implications, cost elements, and performance characteristics. It is yet another strategy for demonstrating a preliminary optimization approach. A Decision Tree Classifier is among the most frequent types of classification models. It is just a flowchart-like structure with a network that represents a test on a function. A Decision Tree Classifier, as stated earlier, splits the classifications into sub-sets (i.e., root, left child, right child). In the sample group chosen, it is the most widely used approach [30].

Iterative Dichotomiser-3 (ID-3) designed by the researcher [31] is among the most prevalent Decision Tree Classifier methods, as it produces all possible intelligent Decision Tree Classifiers and chooses the best. When particularly in comparison to the learning algorithm, the learning time is shorter. The number of items and characteristics in the training data set determines the precise complexity of Decision Tree Classifiers. It is not based on any assumptions about probability distributions. To great accuracy, a Decision Tree Classifier algorithm can manage high-dimensional statistics. Information Gain & Gain Ratio are the essential attribute selection measures (ASM) [32].

3.1.1 INFORMATION GAIN VALUE (IGV)

It mainly represents the reduction in entropy value. To analyze various given features and information gain method measures the variation among entropy values before and after partitioning the dataset. The information gain (Decision Tree Classifier model) is first used by the ID-3 process ("Iterative Dichotomiser-3") decision three techniques [33]. Equation (1) describes the formula for information.

$$Information(D) = \sum_{i=1}^{m}$$

$Information(D)$, represents the information for tuple D.

$$Information\,\mathrm{A}(D) = \sum_{i=1}^{V}$$

$$\frac{|Dj|}{|D|}$$

$$X * Information\ (Dj)$$

$$(2)$$

$$Gain(A) = Information\,\mathrm{A}(D) \qquad (3)$$

The Equation (2) shows the information gain value, and Equation (3) shows the gain value.

Where:

- $Information\,\mathrm{A}(D)$ Represents the mean quantity of data required to determine a tuple's classification model in D.
- The weight of the jth division is represented by $[|Dj|\,/|D|]$.
- Data A(D) is the predictive information necessary to classify a tuple from D based on A's extraction of features.
- The characteristic A of Node N (ij) has the maximum information gain; hence Gain (A) is chosen as the dividing feature.

3.1.2 Gain Ratio Value (GRV)

The attribute with several consequences has a bias throughout information gain. It indicates that it favours the attribute with some of the most decision variables. Study the possibility of a unique identification number, such as customer ID, which has nil information (D) based on high purity separation. It thus maximizes the quantity of knowledge gained while also introducing unnecessary partitioning. A gain ratio is an extended version of information gain used primarily by Decision Tree Classifier algorithm C4.5 [17], which improves existing ID3.0. A standardized division method is used to improve the information gain. A J-48 method is a Java integration of the Decision Tree Classifier method C4.5, available in the WEKA data mining platform [34].

3.2 NaiveBayes

It is a classifier based on supervised learning that solves regression and classification tasks. It is based on Bayesian statistics. The basic concept of the least-squares underpins a binary, i.e., two-class and multi-class classification. The technique is most intuitive for binary classification and also for variable input data [35]. The Naive Bayes framework is straightforward and well-suited to large data sets. As compared to other Machine Learning methods, it produces a higher precision.

A Bayes theorem calculates the probability of an incident happening based on the possibility of a previous incident. Equation (4) expresses the mathematical model-

$$Prob\,(\ Ai$$

Bi

$$\frac{Prob\left(\frac{Bi}{Ai}\right) * Prob(Ai)}{Prob(Bi)}$$

Where Ai and Bi represent events, here, once event Bi is true, we need to find the probability of the event setAi. Evidence is another term for event Bi.

- Ai's priority is $Prob\ (Ai)$ (The prior probability, such as the possibility of an event occurring prior to actually confirmation, has been received.).
- Here $Prob\ (Ai, Bi)$ is the possibility of the number of factors determining of B, i.e. the possibility of activity continuing once reported.
- The proof is a consistent value to an unidentified instance's characteristic (here, it is event$B$i).

3.3 Random Forest

It is a supervised technique based ML classifier that designs and builds models using Decision Tree Classifiers. Trees, in general, learn abnormal behaviour and overfit the trained model with minor differences and bias. It is usedto decrease the variance among features in a given dataset. It also helps in classification same training dataset and testing datasets and emerges at the cost of a modest bias increase. Various companies such as banking and online use this method to estimate objectives. An ensemble methodology is used to classify, predict the future, and perform specific activities. If researchers try to classify something, the Random Forest will generate a class that almost all trees have chosen [36]. Random Forests give the effects of K-fold cross-validation.

Both Scikit-learn and Spark provide details on the impurity factor equations in their evidence. Users can use both parameters of Gini impurity by default and set their variance as a substitute for categorization. In regression, both of the parameters use mean square error to determine variance reduction. Variability reduction can also becalculated in Scikit-learn using mean absolute error [37].

$$Gini\ Impurity = 1 - Gini \qquad (5.1)$$

$$Gini\ = \ \mathrm{P1}^2 + \mathrm{P2}^2 * \mathrm{P3}^2 \qquad \dots\dots.. + \mathrm{Pn}^2 \qquad (5.2)$$

The Equation (5.1) represents the Gini impurity formula. **Where** P1 … … Pn represents the probabilities of each possible class in solution space, $Gini$ represents the purity, and $Gini\ Impurity$ represents the impurity of a particular node. Here $Gini$ works only for categorical targets.

3.4 K-Nearest Neighbor

It is a supervised Machine Learning methodology for the solution of regression as well as classification complications. It is simple to set up and acknowledge, but it does have the disadvantage of being noticeably slower as the volume of data in use expands. As a result of its high level of accuracy, the kNN method can directly compete with precise existing models. If people need high precision and yet do not need a human-readable method, the KNN algorithm is perfect. We can mainly evaluate forecasts based on distance metrics [38].

The best algorithm for the given data set is complicated and depends on several samples, features, and dimensions. Datasets are used in Machine Learning as they need an intelligent analysis. For a search location, the degree of neighbours is demanded. [39].

- Most of the time, the value of k has little effect on the brute force search period.
- The time it takes to search a ball tree, or even a KD tree, slows down as k grows due to the presence of two components, higher k and a substantial chunk of the dimensional space. Second, the tree is traversed using (k>1) and needs an internal queue for results.

When the k is more than N, the capacity to prune branches in a tree-based query is limited. Brute force queries may be more efficient in this circumstance. A construction step is required for both the ball tree and the KD Tree. When amortized across a large number of queries, the cost of this architecture becomes trivial. On the other hand, construction can account for a large portion of the total cost if only a few queries are run. A brute force method is superior to a tree-based method when only a few query points are required [40].

3.5 SVMMethod

It is a supervised classification based ML method that can be utilised to find solutions for different regression and classification tasks. Moreover, this is often used to overcome regression and classification tasks. The method is analyzed as a transition phase within an n-dimensional neighbourhood (n represents the number of features) [41]. It is used to solve complex problems that cannot be solved linearly. A "kernel trick" function can be used to find non-linear solutions to any problems using SVM efficiently. In SVM, the statistics are plotted into a high-dimensional area where the challenge can be separated sequentially. Classifier chooses a division line with the highest deficit. Every point is considered as a support vector. A fully trained classification algorithm can categorize any test sequence and anticipate the training examples (generalization) [42].

A Discriminant function feature can be used to determine the classification model. Equation (6) represents the formula for the classification model function.

$$F(Xi) = [WXi + b] \tag{6}$$

Here Xi, denotes a training or test pattern, w represents the weight vector value, and b represents the bias correction factor related to the function. A total of the product line of vector elements can be determined using the input space and vector combination $[Wi * Xi]$ and is denoted by the symbol WXi [43].

The regression model in the system of two functionalities is asfollows Equation (7)-

$$F(Xi) = [W1 * X1 + W2 * X2 + b] \tag{7}$$

After training, the SVM provides us with estimates for $W1$,$W2$, and b.

3.6 LogisticRegression

It is yet another method that Machine Learning has obtained out from the domain of statistics. It is mainly preferred for classification problems based on two-class features. It is a statistical technique that predicts received data derived from previous findings from the given data set. A logistic regression method anticipates multiple data parameters by studying the connection among one or many pre-existing predictor factors [44]. The logistic regression equation can be defined as described in the equation-

$$y = \frac{[e^{(\beta 0 + \beta 1 * x + \beta 2)}]}{[(1 + e^{(\beta 0 + \beta 1 * x + \beta 2)})]} \tag{8}$$

Where:

The predicted value represents by variable, A the bias term represented by β0, and β1 represents the single data value coefficient (x), and β2 represents the double value. The training examples should learn the β coefficient (the actual statistic which is stable) for each section in the data input.

4.0 RESULTS AND DISCUSSION

This section provides an analysis of different attributes associated with various heart diseases. This experimental analysis is based on a machine learning supervised classification technique, i.e., SVM, KNN, Decision Tree Classifier, RF, and Naïve Bayes.

In this research, we utilise an online UCI heart disease dataset [21]. The implementation of various ML techniques has been done using python programming under Anaconda distribution.

4.1 DatasetDescription

The data set for this study is retrieved from the UCI Center Machine Learning database (online data set) [21, 25]. The database consists of a total of 4 data sources collected from four healthcare institutions. Compared to other data sets, the Cleveland data source has very few missing features and more records.

This dataset contains the record of 303 patients. Although there are 76 characteristics in this data set, all released studies only use a subcategory of fourteen of them. It has a range of zero (no existence of disease) to four (existence of disease). An experimental study using the Cleveland data set focuses on differentiating between appearance (attributes 1, 2, 3, and 4) and absence (attribute 0). Table 2 shows the various parameters related to the heart disease data set. Table 2. Attributes in Heart disease dataset (UCI)

S. No.	Parameter	Description	Values
I.	Age	Age in years	Numeric values
II.	Sex/Gender	Gender type, i.e., Male or Female	1: Male, 2:Female
III.	CP	Chest Pain Level	Four types of Chest pain (0,1,2,3)
IV.	Trestbps	Blood pressure vale at the time of rest	< or > 120 Mg/DL
V.	Chol	Represents the level of Serum Cholesterol	Numeric values
VI.	FBS	Represents the level of sugar in fasting blood	Numeric values
VII.	Restecg	Represents the level of Resting electrocardiographic	Five types of Values (0,1,2,3,4)
VIII.	Thalach	Maximum heart rate level	Numeric values
IX.	Exang	Exercise enduced level	Yes / No
X.	Oldpeak	ST level during the workout, compared with the results of rest taken	Numeric values
XI.	Slope	level of peak exercise in ST-segment	Three values (0:up, 1:flat, 2:down)
XII.	CA	Reprenets the number of flourosopy vessels	Four values (o to 3)
XIII.	Thal	Used for Defect classes (4 classes) Normal; fixed; reversible; Non-reversible	Four values (0 to 3)
XIV.	Class	Representing the Target	Two classes (0,1)

4.2 Pre-ProcessingofHeartDiseaseDataset

There are no apparent limitations in the dataset, and they are also not evenly distributed. The dataset includes a significant number of incomplete as well as noisy values. These statistics are pre-processed to tackle missing value difficulties and generate accurate forecasts. The pre-processing data phase consists of several stages, including data cleaning, transformation (normalization and aggregation), data integration, and reduction [39]. The proposed system is utilized two different approaches for data pre-processing, i.e., normalization, aggregation. However, the obtained results after using a customarily distributed dataset to overcome the overfitting issues and then utilizing Exclusion Forest for feature extraction and anomalies removal are pretty enticing. The skewness of statistics, outlier recognition, and data distributions are checked using numerous plotting methods [32]. Figure 2 shows the Correlation between Heart diseases and numeric features. This matrix shows the summary of the overall dataset as input towards a more advanced analysis and identification of more investigation. Each of the cells here represents the Correlation between variables.

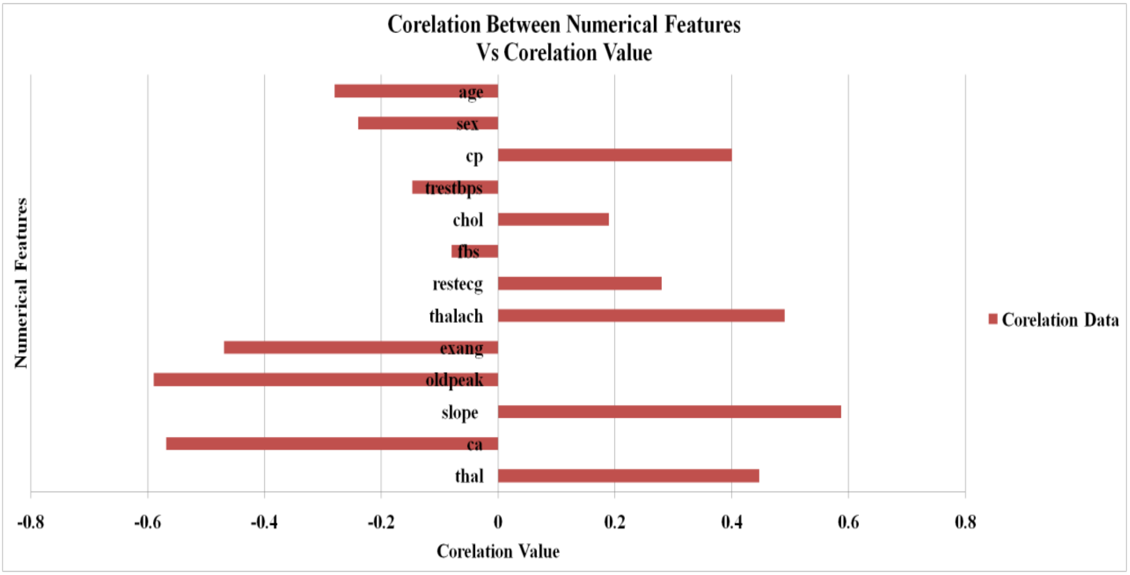


Fig. 2. Correlation between Heart Diseases and Numeric Features

4.2.1 Verifying Data Distribution

We need to analyse the data pattern in order to categorise the heart disease dataset and ensure precise predictions about heart disease levels. Heart disease positive cases are 54% in this heart disease dataset, whereas no heart disease cases are 45%. Perhaps the dataset must be balanced to prevent overfitting problems. Figure 3 provides the heart disease dataset's representation with disease statistics. Here 1 represents data with heart disease, and 0 represents data with no heart disease. This dataset contains 165 candidates with heart disease and 138 without any heart disease, showing the distribution of positive and negative heart disease cases. It will help an ML method identify which patterns in the dataset can be best for the process [40].

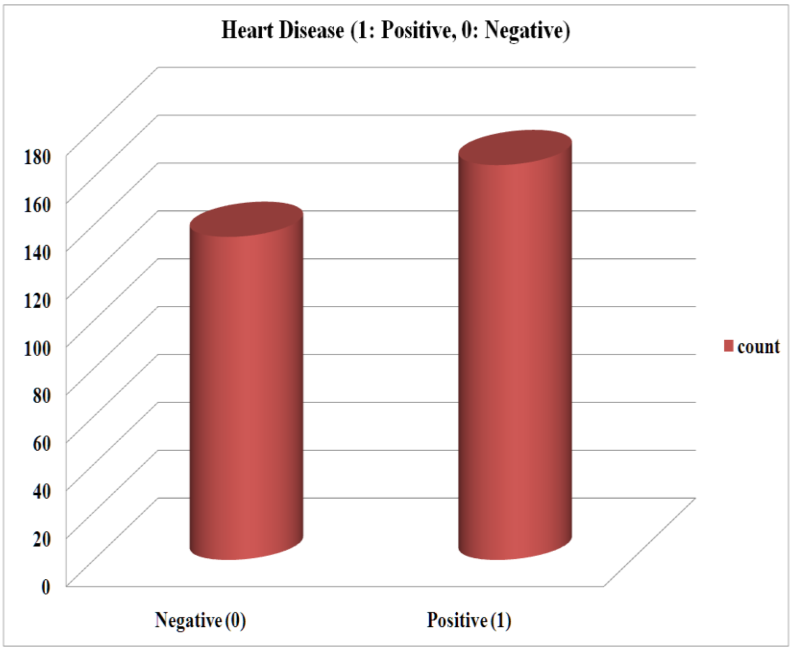


Fig. 3. Distribution of data (1: Positive, 0: Negative)

Figure 3 shows the representation of the candidate count of heart disease in the dataset, here 1 represents data with heart disease, and 0 represents data with no heart disease. This dataset contains 165 persons with heart disease and 138 persons without heart disease. Now we apply the Isolation Forest method to remove the anomalies.

Isolation Forest (Anomalies Removal): We use an Isolation forest method for anomalies removal from the dataset. Randomly sub-sampled data is processed in an Isolation Forest in a hierarchical structure depending on random selection characteristics. Anomalies are less likely to appear in samples extending further down thetree since they need more branches to separate individuals.

Algorithm 1: Isolation Forest

```
This algorithm will remove the anomalies from the dataset.
Input: Data with Anomalies
Output: Data with no anomalies
1. A random sub-sample of the dataset is chosen and allocated to a binary tree while supplied data.
2. The tree is first split by choosing a random characteristic (from collection among all N characteristics). Then there is a branching randomly sampled threshold.
3. If a data point's quantity is below the threshold, it flows to the left new branch; otherwise, it travels towards the right-side tree. As a result, a node is broken into two parts, i.e., left subtree and right subtree.
4. This method is repeated until each piece of data is entirely separate or until the maximum level will not be achieved.
5. To create random binary forests, repeat all procedures above.
```

Verifying the Asymmetry of a Distribution (Skewness)

Various methods and graphs obtain numerous distribution results. To verify various parameters and also determine the asymmetry (skewness) of the heart diseases data set. Separate plots are displayed so that a comprehensive review of the information will be further examined. The potting includes various distributions: **a)** age and gender distributions, **b)** trestbps and chest pain distributions, **c)** fasting blood and cholesterol distributions, **d)** thalach and electrocardiogram distribution, **e)** resting electrode and distributions, **f)** old peak and exang distributions, **g)** ca and slope distributions, and **h)** target and thaI distributions [28].

Figure 4 represents the ratio based on sex (1: male and 0: female). It represents the heart disease count based on Sex, 1: Have heart disease and 0: have no disease. The graph indicates that 114 females and 93 males have no heart disease, whereas 23 females and 72 males have heart disease out of 303 patient datasets. These findings show that males are more likely than females to suffer from cardiac disease.

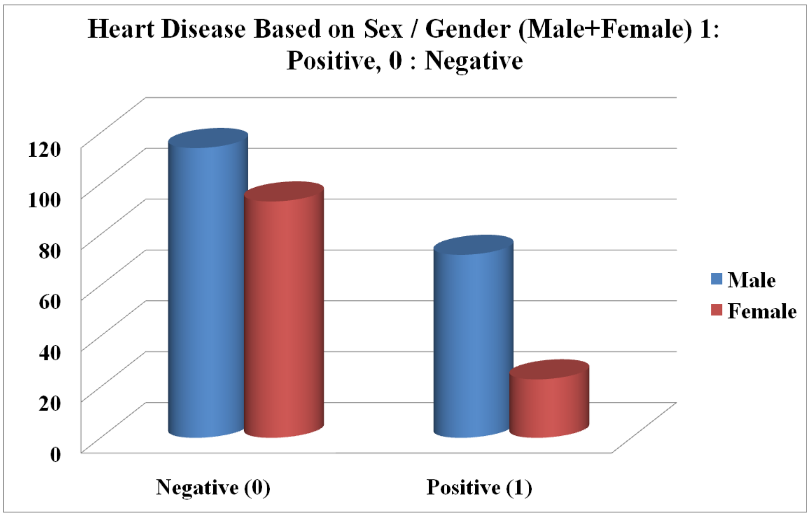


Fig. 4. Heart disease ratio based on sex (1: Have heart disease and 0: Have no heart disease)

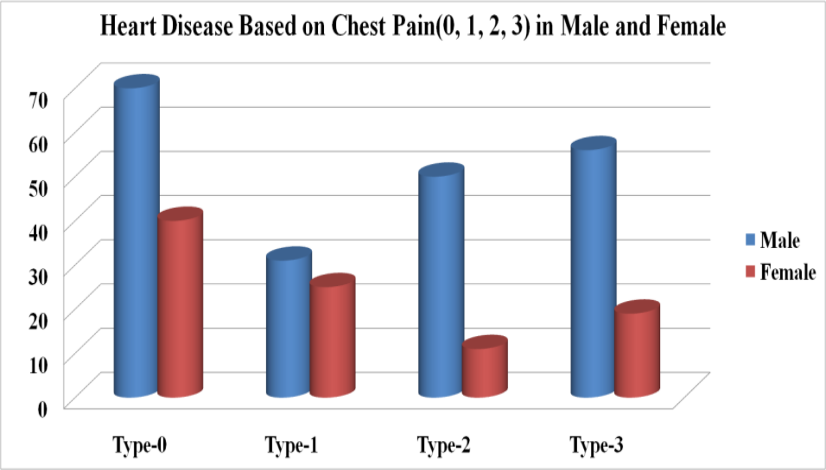


Fig. 5. Heart disease ratio based on Chest pain level

Figure 5 represents Heart Disease by Chest Pain Type (0, 1, 2 and 3). This graph indicates that in male type-0, chest pain cases are 70, type 1: 31, type 2: 50 and type 3: 56 cases. The results indicate chest pain type -0; patients are higher as compared to other possible types. Similar in type-0 chest pain cases 40, type-1 chest pain cases 25, type 2 chest pain cases 11 and type 3 chest pain cases are 19. The results indicate chest pain type -0; patients are higher as compared to other possible types.

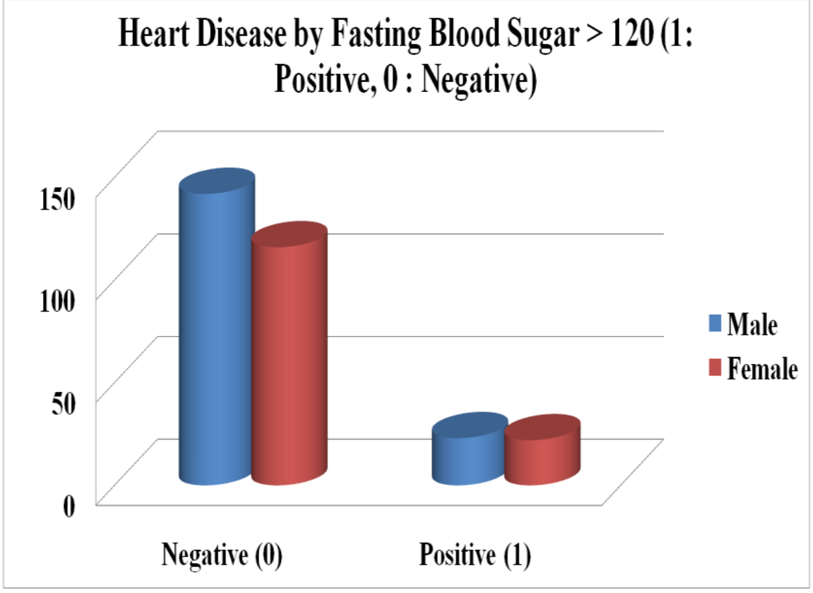
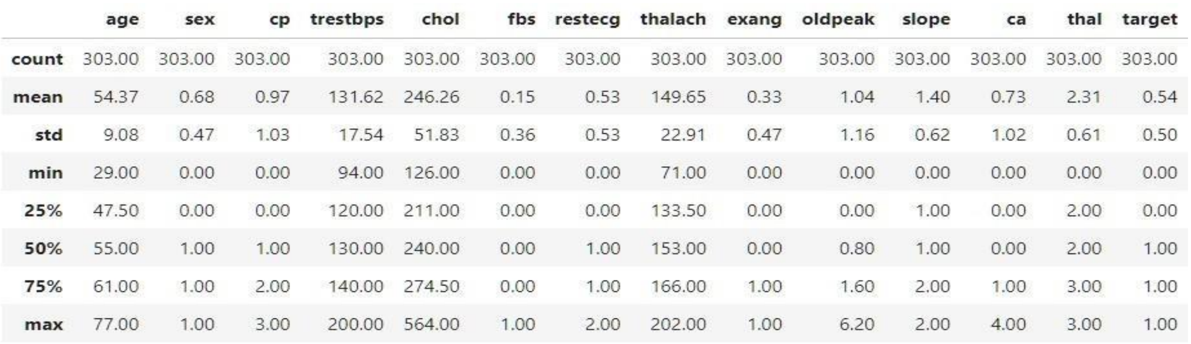


Fig. 6. Heart disease based on fasting blood sugar >120 (1: Heart disease and 0: No heart disease)

Similar Figure 6 represents Heart disease based on fasting blood sugar >120 (1: Heart disease and 0: No heart disease). This graph indicates that positive heart disease cases based on FBS value (Greater than 120) are 23 Male and 22 Females, and type negative (no heart disease) cases are 142 males and 116 females. The results indicate that male patients are due to FBS level as compared to females.

4.2.2 Exploratory Data Assessment

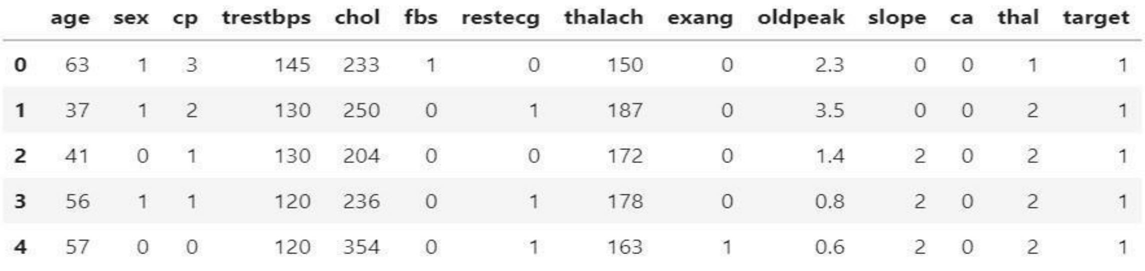
During this procedure, we obtained important measurement information from the statistics, so we have checked the data distributions process of the different characteristics, such as the correlations of variables with the destination variable, various probabilities. Figure 7 shows the cardiovascular disease dataset’s overview. This dataset contains 14 attributes, and each attribute has different values, including count, mean, std, min, and max. This dataset contains 165 persons with heart disease and 138 persons without heart disease. Similar Figure 8 shows the parameter distribution and their values in the dataset.



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
count	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00
mean	54.37	0.68	0.97	131.62	246.26	0.15	0.53	149.65	0.33	1.04	1.40	0.73	2.31	0.54
std	9.08	0.47	1.03	17.54	51.83	0.36	0.53	22.91	0.47	1.16	0.62	1.02	0.61	0.50
min	29.00	0.00	0.00	94.00	126.00	0.00	0.00	71.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	47.50	0.00	0.00	120.00	211.00	0.00	0.00	133.50	0.00	0.00	1.00	0.00	2.00	0.00
50%	55.00	1.00	1.00	130.00	240.00	0.00	1.00	153.00	0.00	0.80	1.00	0.00	2.00	1.00
75%	61.00	1.00	2.00	140.00	274.50	0.00	1.00	166.00	1.00	1.60	2.00	1.00	3.00	1.00
max	77.00	1.00	3.00	200.00	564.00	1.00	2.00	202.00	1.00	6.20	2.00	4.00	3.00	1.00

Fig. 7. Analysis of data of all numeric parameters in the data set

Figure 9 shows the distribution of heart disease parameters restecq, exang and slope. This graph is plotted based on heart disease state, i.e., have heart disease and no heart disease. Figure 10 and Figure 11 show the distribution of heart disease parameters ca, thal, target, sex, cp, FBS. This graph is plotted based on heart disease state, i.e., have heart disease and no heart disease.



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Fig. 8. Heart disease data set parameters and their values

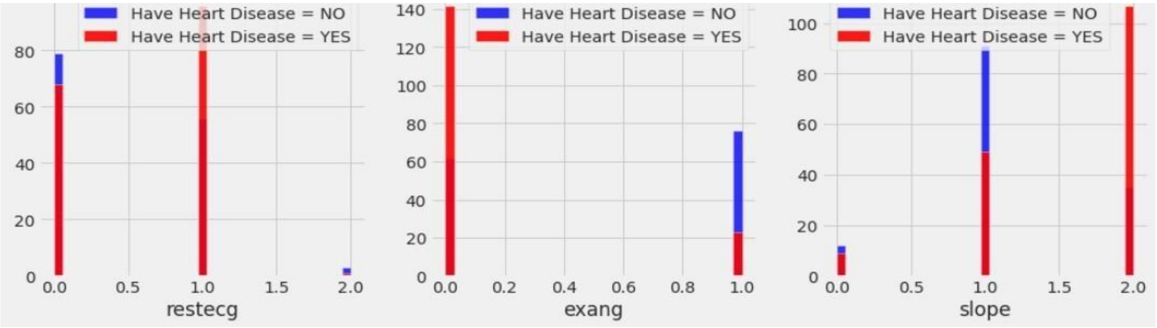


Fig. 9. Distribution of parameters based on restecq, exang and, slope

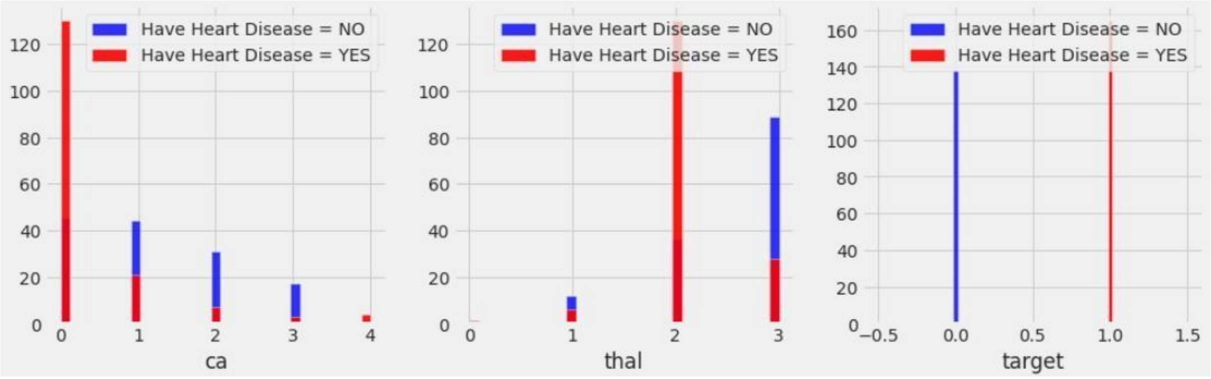


Fig. 10. Distribution of parameters based on restecq, exang and, slope

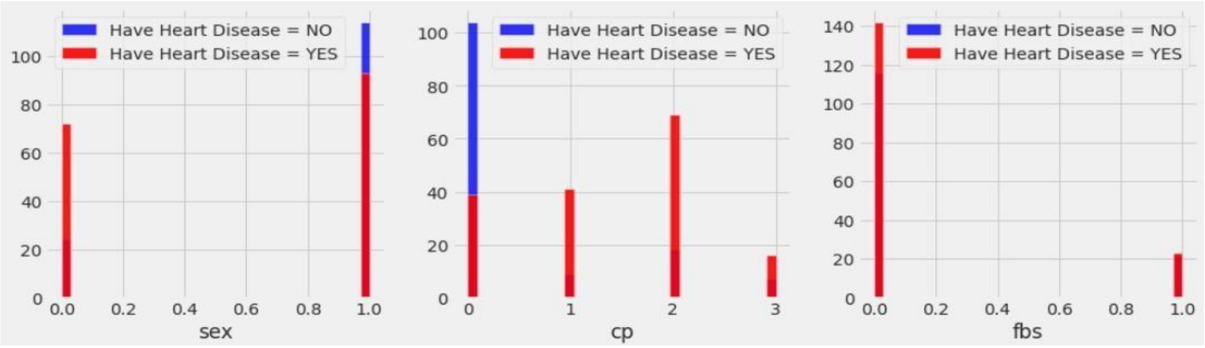


Fig. 11. Distribution of parameters based on sex, cp and fbs

4.3 Comparison Parameters

To assess the effectiveness of the forecast modelling techniques, we compare them to each other. The following performance measuring parameters have been used in this research [30].

Here TP= (true positives), FP= (false negatives), TI = Total Instance.

- **Accuracy:** the fraction of occurrences in the complete dataset has been successfully forecast out of all incidences, as described in Equation (9)

$$\frac{TP + TN}{(9)} \qquad Accuracy =$$

(TI)

- **Sensitivity / Recall:** the fraction of correctly predicted set of data occurrences that are positive [29]. Equation (10) describes the recall equation.

Sensitivity or Recall =

TP

$(TP + FP)$ (10) ————

- **Precision:** It measures how many positive forecasting have been accurate. The precision determines the accurateness for such a minority class as a result. It is measured by deducting the count of positive predictions productive instances by the total number of samples instances anticipated [30]. Equation (11) describes the equation for precision.

———— $Precision = [$

TP

$]$ (11)

$TP + FP$

- **F1-Score:** Precision and recall are combined in a composite harmonic mean (mean of reciprocals). Researchers initially assess the model's accuracy or ability to recognize only each relevant data source to do the same [22].

$$F1Score = 2 \times [\frac{precision \times recall}{(precision + recall)} \qquad (12\,)$$

- **UC (Area under the curve):** A probability approximate for a framework ranking a random selection positive case larger than a selected randomly negative specific instance. The filled area under the curve receiver required to operate characteristic curves is drawn to visual elements to evaluate the models' effectiveness.

$$y = f(x)\ between\ x = a\ ,\ \&\ x = b \qquad (13\,)$$

4.4 Experimental Findings

In this section, the classification techniques and consequences have been discussed from different viewpoints. The experimental analysis has been performed by using Python programming language in Anaconda distribution. Experimental results represent the correlation measure results. We are using 60% dataset for training and 40% for testing.

Correlation parameters are determined to measure the strength of the association among multiple numerically measured dependent variables. The effectiveness of classification models has been evaluated based on a set of criteria. Furthermore, a k-fold cross-validation procedure has been applied. Performance measurements have been used to monitor the efficiency of classification methods. All the essential features are normalized and standardized to categories the actual data.

4.4.1 K-Fold Cross-Validation Outcomes

The exclusive features of the input data were examined by applying 10-fold cross-validation approaches on several ML classification models. There were two components to the dataset: the training dataset and the testing dataset. About 90% of the data was utilised for training and only 10% for testing. Table 3 shows the results of 10-fold cross-validation of classification models with exclusive features.

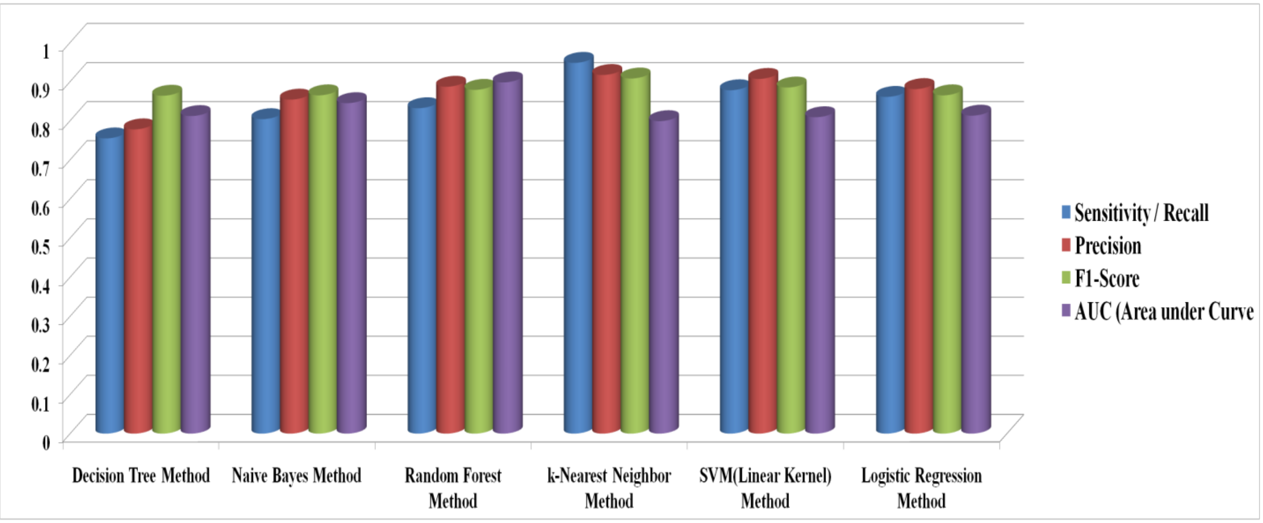


Fig. 12. Accuracy % of various ML Methods

Table 3. Experimental Results for various ML Methods

ML Method	Sensitivity / Recall	Precision	F1-Score	AUC (Area under curve)
Decision Tree (DT) Method	0.7548	0.778	0.864	0.812
Naive Bayes (NB) Method	0.804	0.854	0.865	0.845
Random Forest (RF) Method	0.832	0.887	0.879	0.898
k-Nearest Neighbor (KNN) Method	0.948	0.917	0.908	0.799
SVM(Linear Kernel) Method	0.878	0.907	0.885	0.809
Logistic Regression (LR) Method	0.861	0.881	0.865	0.813

Table 3 and Figure 12 show the experimental results for various Machine Learning methods. The KNN classification achieved the maximum score when the total number of the nearest Neighbors was eight. The linear kernel performed the best for this data set with the maximum score among the four kernels; Linear, Poly, RBF, and Sigmoid. In the Decision Tree Classifier algorithm, the score is maximum, with the total number of features to be selected being either 4 or 18. The research involved the analysis of patient data with pre-processing. After that, the classification models, i.e. KNN, SVM, Decision Tree Classifier, and Random Forest, were trained and tested with maximum scores.

The experimental result for Decision Tree Classifier method is showing accuracy (0.7898), recall (0.7548), precision (0.778), F1 score (0.864) and ACU (0.812). Similarly, the Naive Bayes Machine Learning method is showing accuracy (0.8678), recall (0.804), precision (0.854), F1-score (0.865) and AUC (0.845). The experimental result for the Random Forest method is showing accuracy (0.879), recall (0.832), precision (0.887), F1-score (0.879) and AUC (0.898).

Similarly, for k-Nearest Neighbor method, the experimental results are accuracy (0.941), recall (0.948),precision (0.917), F1-Score (0.908) and AUC (0.799). Similar for SVM method experimental results are accuracy(0.891), recall(0.878), precision(0.907), F1-Score(0.885) and AUC(0.809) and for Logistic regression method, the experimental results are accuracy(0.871), recall(0.861), precision(0.881), F1-Score(0.865) and AUC(0.813).

4.4.2 Confusion Matrix for the Classifiers

Different performance evaluation metrics have been implemented in this research work to assess the classifier performance. Each observation in the test dataset is forecast in a precisely single box using the confusion matrix. Here are two responding subclasses; hence the matrix becomes [2*2]. Table 4 shows a sample confusion matrix for actual and predicted data samples. Table 5 and Table 6 shows the confusion matrix results for various methods during the training and testing phases.

Table 4. Sample Confusion Matrix

Class	Actual Positive	Actual Negative
Predicted Positive	TP	FP
Predicted Negative	FN	TN

Table 5. Confusion Matrix Results for Training (Heart Disease Dataset)

Method Name	TP	FP	FN	TN
LR Method	80	17	11	104
SVM(Linear Kernel) Method	89	8	6	109
KNN Method	82	15	13	102
RF Method	97	0	0	115
DT Classifier Method	97	0	0	115
NB Method	97	0	0	115

Table 6. Confusion Matrix Results for Testing (Heart Disease Dataset)

Method Name	TP	FP	FN	TN
LR Method	34	7	5	45
SVM(Linear Kernel) Method	36	5	6	44
KNN Method	35	6	6	44
RF Method	33	8	8	42
DT Classifier Method	34	7	13	37
NB Method	33	8	8	42

4.4.3 Experimental Results Based on Selected Features using Filter Method

In clinical diagnosis, the feature selection techniques support the selection of accurate health care decisions. Table 7 shows the relevant seven features which are selected using the filter feature selection technique. Figure 13 shows the feature score of the essential features. Chest pain seems to be a significant aspect for predicting heart disease in the rankings graph. We did the test on a range of different amounts of persons. The results demonstrate the strength of SVM and Random Forest methods with the selected features.

Table 7. Experimental results Based on Selected Features using the Filter method

Feature	Name of the Feature	Code	Feature Score
4	Age	Ag	0.231
8	Sex(Male, Female)	Sex	0.114
11	Chest Pain	CP	0.148
12	Fasting Blood Sugar	FBS	0.151
3	Cholesterol Level	CL	0.221
10	slope	slope	0.232
9	Ca	CA	0.114

4.4.3 Experimental results for Machine Learning methods

The final prediction rate (heart disease prediction) results for all the Machine Learning classifier techniques are presented in Figure 14. The SVM method shows 83.25%, Decision Tree Classifier 83.89 %, KNN 86.45, Random Forest 88.35, Logistic Regression 84.22% and Naive Bayes 84.69% prediction score. The experimental result demonstrates that the Random Forest classifier technique has a better prediction rate for detecting heart disease than other models.

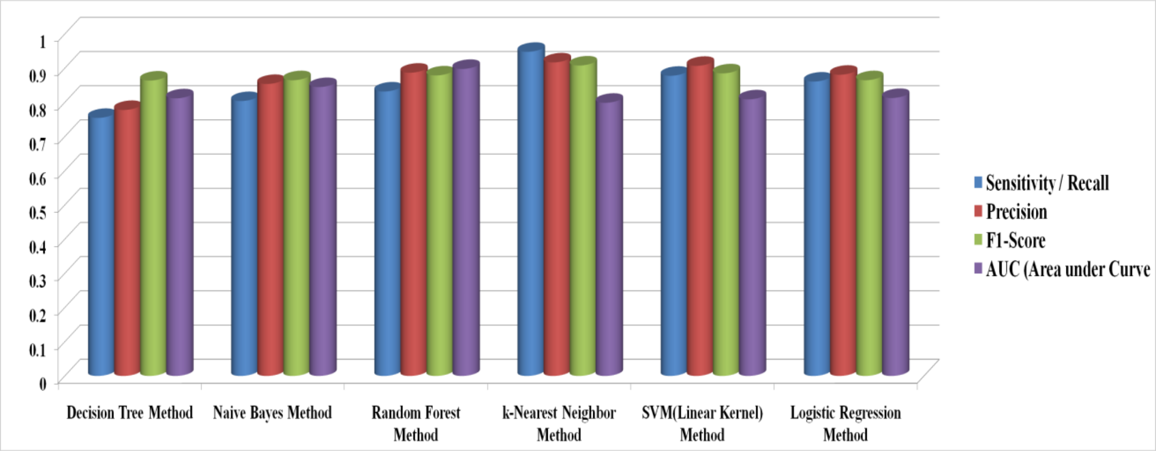


Fig. 13. Performance-based on selected features for various classifiers

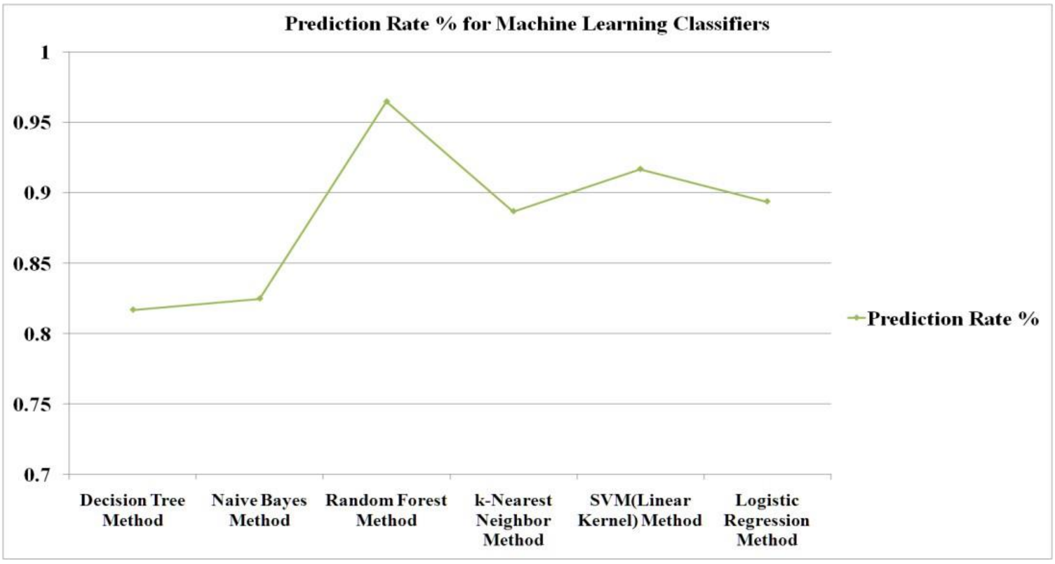


Fig. 14. Prediction Rate % for Machine Learning Classifiers

5.0 CONCLUSION

This study has focused on four distinct approaches to doing comparative evaluation and achieving genuine good performance. In this study, we found that machine learning techniques performed noticeably better than statistical methods. This study supports other researchers' findings that, despite smaller databases, machine learning (ML) models are the most effective method for predicting and categorizing cardiac disease. On the UCI online heart disease data set, a number of performance metrics, including precision, F1 score, accuracy, and recall, have been compared for all Machine Learning classification approaches. Compared to the fourteen accessible parameters, the KNN classification algorithm fared better. The future scope of this work, and as limitations imposed to the efforts made above, the emerging classification models can be included for complex time series data set. In addition to this, the research can be applied to other biological diseases.

Conflicts of Interest

Concerning the publication of this study, there are no conflicts of interest among the authors.

Availability of Data

Upon request, the statistics used to verify the findings of this research can be obtained from the corresponding author.

REFERENCES

[1] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, "Heart disease identification method using Machine Learning classification in e-healthcare," *IEEE Access*, Vol. 8, 2020, pp. 107562–107582.

[2] M. Mullen, A. Zhang, G. K. Lui, A. W. Romfh, J. W. Rhee, and J. C. Wu, "Race and Genetics in Congenital Heart Disease: Application of iPSCs, Omics, and Machine Learning Technologies," *Frontiers in Cardiovas- cular Medicine*, Vol. 8, 2021, pp. 37–51.

[3] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid Machine Learning techniques," *IEEE Access*, Vol. 7, 2019, pp. 81542–81554.

[4] N. K. Trivedi, S. Simaiya, U. K. Lilhore, and S. K. Sharma, "An efficient credit card fraud detection model based on Machine Learning methods," *International Journal of Advanced Science and Technology*, Vol. 29, No. 5, 2020, pp. 3414–3424.

[5] A. K. Dwivedi, "Performance evaluation of different Machine Learning methods for prediction of heart disease," *Neural Computing and Applications*, Vol. 29, No. 10, 2018, pp. 685–693.

[6] S. K. Jonnavithula, A. K. Jha, M. Kavitha, and S. Srinivasulu, "Role of Machine Learning algorithms over heart diseases prediction," *AIP Conference Proceedings*, Vol. 2292, 2020, pp. 40013–40013.

[7] S. S. Yadav, S. M. Jadhav, S. Nagrale, and N. Patil, "Application of Machine Learning for the detection of heart disease," *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, 2020, pp. 165–172.

[8] V. Patil and U. K. Lilhore, "A survey on different data mining & Machine Learning methods for credit card fraud detection," *International Journal of Scientific Research in Computer Science, Engineering and Infor- mation Technology*, Vol. 3, No. 5, 2018, pp. 320–325.

[9] M. Poongodi and S. Bose, "Detection and Prevention system towards the truth of convergence on decision using Aumann agreement theorem," *Procedia Computer Science*, Vol. 50, 2015, pp. 244–251.

[10] S. Nashif, M. R. Raihan, M. R. Islam, and M. H. Imam, "Heart disease detection by using Machine Learning algorithms and a real-time cardiovascular health monitoring system," *World Journal of Engineering and Technology*, Vol. 6, No. 4, 2018, pp. 854–873.

[11] N. Pawar, U. K. Lilhore, and N. Agrawal, "A hybrid ACHBDF load balancing method for optimum resource utilization in cloud computing," *International Journal of Scientific Research in Computer Science, Engineer- ing and Information Technology*, Vol. 3307, 2017, pp. 367–373.

[12] S. Sharma and M. Parmar, "Heart diseases prediction using deep learning neural network model," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, Vol. 9, No. 3, 2020, pp. 124– 137.

[13] D. Singh and J. S. Samagh, "A comprehensive review of heart disease prediction using Machine Learning," *Journal of Critical Reviews*, Vol. 7, No. 12, 2020, pp. 281–285.

[14] K. Guleria, A. Sharma, U. K. Lilhore, and D. Prasad, "Breast Cancer Prediction and Classification Using Supervised Learning Techniques," *Journal of Computational and Theoretical Nanoscience*, Vol. 17, No. 6, 2020, pp. 2519–2522.

[15] U. K. Lilhore, S. Simaiya, K. Guleria, and D. Prasad, "An Efficient Load Balancing Method by Using Machine Learning-Based VM Distribution and Dynamic Resource Mapping," *Journal of Computational and Theoretical Nanoscience*, Vol. 17, No. 6, 2020, pp. 2545–2551.

[16] S. K. Sharma, U. K. Lilhore, S. Simaiya, and N. K. Trivedi, "An Improved Random Forest Algorithm for Predicting the COVID-19 Pandemic Patient Health," *Annals of the Romanian Society for Cell Biology*, 2021, pp. 67–75.

[17] U. K. Lilhore, S. Simaiya, D. Prasad, and D. K. Verma, “Hybrid Weighted Random Forests Method for Prediction & Classification of Online Buying Customers,” *Journal of Information Technology Management*, Vol. 13, No. 2, 2021, pp. 245–259.

[18] P. S. Kohli and S. Arora, “Application of Machine Learning in disease prediction,” *4th International conference on computing communication and automation (ICCCA)*, 2018, pp. 1–4.

[19] A. Ismail, S. Abdlerazek, and I. M. El-Henawy, “Big data analytics in heart diseases prediction,” *Journal of Theoretical and Applied Information Technology*, No. 11, 2020, pp. 98–110.

[20] A. N. Repaka, S. D. Ravikanti, and R. G. Franklin, “Design and implementing heart disease prediction using naives Bayesian,” *2019 3rd International conference on trends in electronics and informatics (ICOEI)*, 2019, pp. 292–305.

[21] V. V. Ramalingam, A. Dandapath, and M. K. Raja, “Heart disease prediction using Machine Learning tech- niques: a survey,” *International Journal of Engineering & Technology*, Vol. 7, No. 2, 2018, pp. 684–687.

[22] “ Heart Disease Data Set, UCI Machine Learning dataset, contains 4 databases, Access on 10th July 2021,htt ps://archive.ics.uci.edu/ml/datasets/heart+Disease.,” 2021.

[23] P. Singh, S. Singh, and G. S. Pandi-Jain, “Effective heart disease prediction system using data mining techniques,” *International journal of nanomedicine*, Vol. 13, 2014, pp. 121–128.

[24] J. Patel, D. Tejalupadhyay, and S. Patel, “Heart disease prediction using Machine Learning and data mining technique,” *Heart Disease*, Vol. 7, No. 1, 2015, pp. 129–137.

[25] Y. Khourdifi and M. Bahaj, “Heart disease prediction and classification 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, pp. 242–252.

[26] S. M. Nagarajan, V. Muthukumaran, R. Murugesan, R. B. Joseph, M. Meram, and A. Prathik, “Innovative feature selection and classification model for heart disease prediction,” *Journal of Reliable Intelligent Environments*, 2021, pp. 1–11.

[27] R. T. Selvi and I. Muthulakshmi, “An optimal artificial neural network-based big data application for heart disease diagnosis and classification model,” *Journal of Ambient Intelligence and HumanizedComputing*, Vol. 12, No. 6, 2021, pp. 6129–6139.

[28] A. Singh and R. Kumar, “Heart disease prediction using Machine Learning algorithms,” *2020 international conference on electrical and electronics engineering (ICE3)*, 2020, pp. 452–457.

[29] M. Poongodi, *et al.*, “Prediction of the price of Ethereum blockchain cryptocurrency in an industrial finance system,” *Computers & Electrical Engineering*, Vol. 81, 2020, pp. 106527–106539.

[30] M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, “Intelligent heart disease prediction system using Random Forest and evolutionary approach,” *Journal of network and innovative computing*, Vol. 4, 2016, pp. 175–184.

[31] J. Nahar, T. Imam, K. S. Tickle, and Y. P. P. Chen, “Computational intelligence for heart disease diagnosis: A medical knowledge-driven approach,” *Expert Systems with Applications*, Vol. 40, No. 1, 2013, pp. 96–104.

[32] A. M. Sagir and S. Sathasivam, “A Novel Adaptive Neuro-Fuzzy Inference System Based Classification Model for Heart Disease Prediction,” *Pertanika Journal of Science & Technology*, No. 1, 2017, pp. 25–30.

[33] M. Poongodi, M. Hamdi, V. Varadarajan, B. S. Rawal, and M. Maode, “Building an Authentic and Ethical Keyword Search by applying Decentralized (Blockchain) Verification,” *IEEE INFOCOM 2020- IEEE Con- ference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2020, pp. 746– 753.

[34] M. Poongodi, M. M. Hamdi, A. Malviya, G. Sharma, S. Dhiman, and Vimal, “Diagnosis and combating COVID-19 using wearable Oura smart ring with deep learning methods,” *Personal and ubiquitous*

computing, 2021, pp. 1–11.

[35] X. Liu, *et al.*, "A hybrid classification system for heart disease diagnosis based on the RFRS method.," *Computational and mathematical methods in medicine,* Vol. 24, No. 5, 2017, pp. 214–224.

[36] S. Simaiya, U. K. Lilhore, D. Prasad, and D. K. Verma, "MRI Brain Tumour Detection & Image Segmentation by Hybrid Hierarchical K-means clustering with FCM based Machine Learning Model," *Annals of the Romanian Society for Cell Biology*, 2021, pp. 88–94.

[37] M. Poongodi and S. Bose, "A novel intrusion detection system based on trust evaluation to defend against DDoS attack in MANET," *Arabian Journal for Science and Engineering*, Vol. 40, No. 12, 2015, pp. 3583– 3594.

[38] M. Poongodi, N. Tu, M. Nguyen, K. Hamdi, and Cengiz, "A Measurement Approach Using Smart-IoT Based Architecture for Detecting the COVID-19," *Neural Processing Letters,* 2021, pp. 1–15.

[39] N. K. Trivedi, S. Simaiya, U. K. Lilhore, and S. K. Sharma, "COVID-19 Pandemic: Role of Machine Learn- ing & Deep Learning Methods in Diagnosis," *Int J Cur Res Rev*—, Vol. 13, No. 06, 2021, pp. 150–156.

[40] S. Dhar, K. Roy, T. Dey, P. Datta, and A. Biswas, "A hybrid Machine Learning approach for prediction of heart diseases," *4th International Conference on Computing Communication and Automation (ICCCA),* 2018, pp. 1–6.

[41] A. Garg, U. K. Lilhore, P. Ghosh, D. Prasad, and S. Simaiya, "Machine Learning-based Model for Prediction of Student's Performance in Higher Education," *2021 8th International Conference on Signal* Processing and Integrated Networks (SPIN), 2021, pp. 162–168.

[42] F. Babicˇ, J. Olejaˊr, Z. Vantovaˊ, and J. Paralicˇ, "Predictive and descriptive analysis for heart disease diagnosis," 2017 federated conference on computer science and information systems (fedcsis), 2017, pp. 155–163.

[43] A. Hassan, D. Prasad, M. Khurana, U. K. Lilhore, and S. Simaiya, "Integration of Internet of Things (IoT) in Health Care Industry: An Overview of Benefits, Challenges, and Applications," *Data Science and Innova- tions for Intelligent Systems,* 2021, pp. 165–180.

[44] S. Simaiya, *et al.*, "EEPSA: Energy Efficiency Priority Scheduling Algorithm for Cloud Computing," *2021 2nd International Conference on Smart Electronics and Communication (ICOSEC),* 2021, pp. 1064–1069.