Analysis and Prediction of Heart Disease and Pathological Suspects using Classification Algorithms in R

1. Abstract

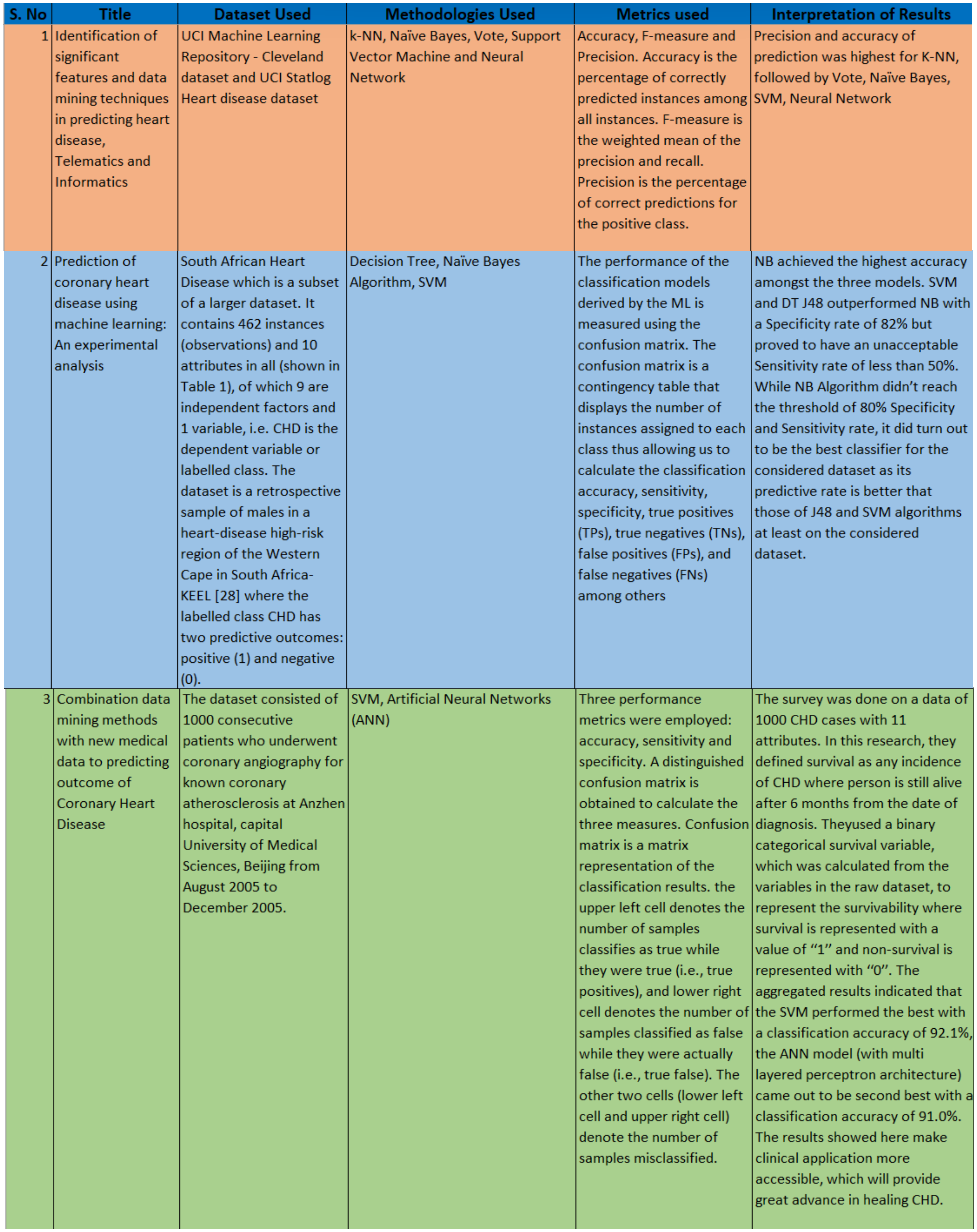
With the new and sedentary lifestyle which a large majority of the population across the world is being exposed to, there has been a significant increase in heart diseases like those of heart failure, pericardial diseases, heart stroke etc. In order to analyse and portray the severity of this issue, we have decided to conduct heart disease data analysis. and implement classification algorithms using cardiotocographic data using essential parameters like fasting blood sugar, serum cholesterol, max heart rate, etc. Through this analysis, we will be able to identify if a person is prone to having a heart disease or not using classification algorithms for gaining accurate and precise results. These classifications will be used to classify patients as pathological and non-pathological suspects of heart diseases. The motive behind choosing this problem statement for data analysis is to create awareness on the severity of heart diseases today and also provide a substantial solution in predicting if a person is likely to have a heart ailment or not.

2. Introduction

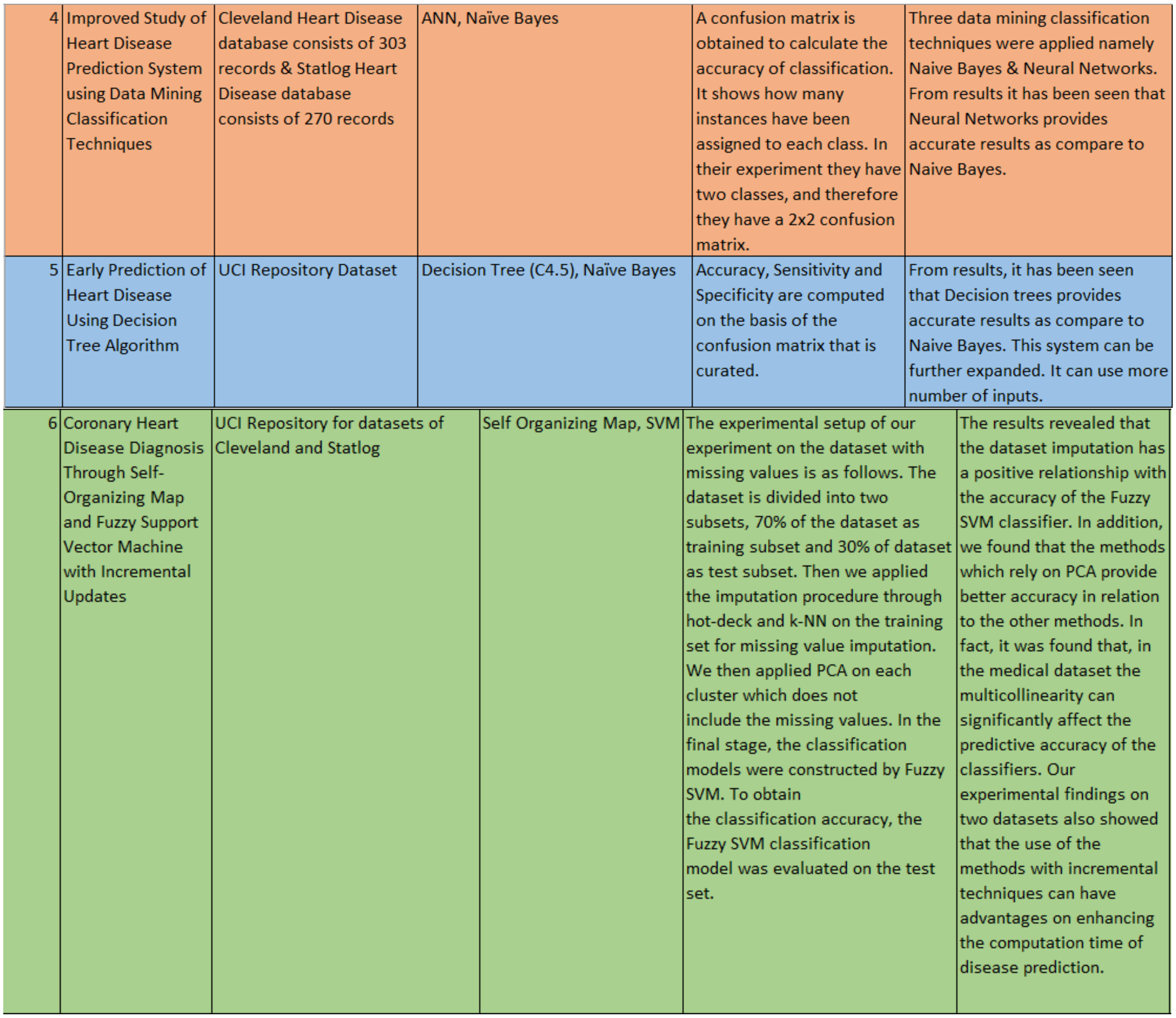
According to surveys conducted by health centres like Centres for Disease Control and Prevention (CDC), heart disease has turned out to be the main leading cause of death especially in the United States of America. Nearly one out of four deaths in the USA occur due to heart diseases and so is the case in most of the other countries. Heart disease occur in may forms and types like those of coronary artery disease, congenital heart effects, arrythmia, dilated cardiomyopathy, etc. Heart disease symptoms also have a wide range. These symptoms are likely to indicate a severe heart problem – anginal/chest pain, difficulty in breathing, fatigue and light headedness, edema or retention of fluid in the body. These are caused due to a slight or major damage in the part of the heart, an issue with blood vessels connected to the heart, age, dietary choices and options, too much stress and anxiety levels. Early detection of hear diseases can often turn out to be a life saver and in order to serve a noble cause, we have selected the problem statement of analysing and predicting heart diseases and pathological suspects using the various forms of classification algorithms and techniques in R Studio.

Classification algorithms form a major part in various applications today, especially in cases where we need to group a set of data or certain attributes into different classes or categories. It is one of the supervised learning techniques wherein a model is generated based on a suitable algorithm after which it is trained using pre-set classes which each of the category or attribute belongs to. Once that's done, the testing phase begins wherein sample attributes and data is passed to the model and the model classifies it into suitable classes based on the training it has received. Throughout this process, identifiers and boundary conditions are predicted in order to classify and categorize easily. Clustering is also another technique to categorize data objects into different classes but clustering is an unsupervised technique as grouping takes place on the basis of similarity of objects. Some of the commonly used classification algorithms in R include: Logistic Regression, Decision Trees, Support Vector Machines (SVM), Naïve Bayes Classifier, k-Nearest Neighbours, Artificial Neural Networks (ANN), and so on. In the project implemented, we have analysed and predicted heart disease occurrence through the following algorithms: Linear Discriminant Analysis (LDA_, Quadratic Discriminant Analysis (QDA), K Nearest Neighbours (KNN), SVM, Random Forests and Gradient Boosting.

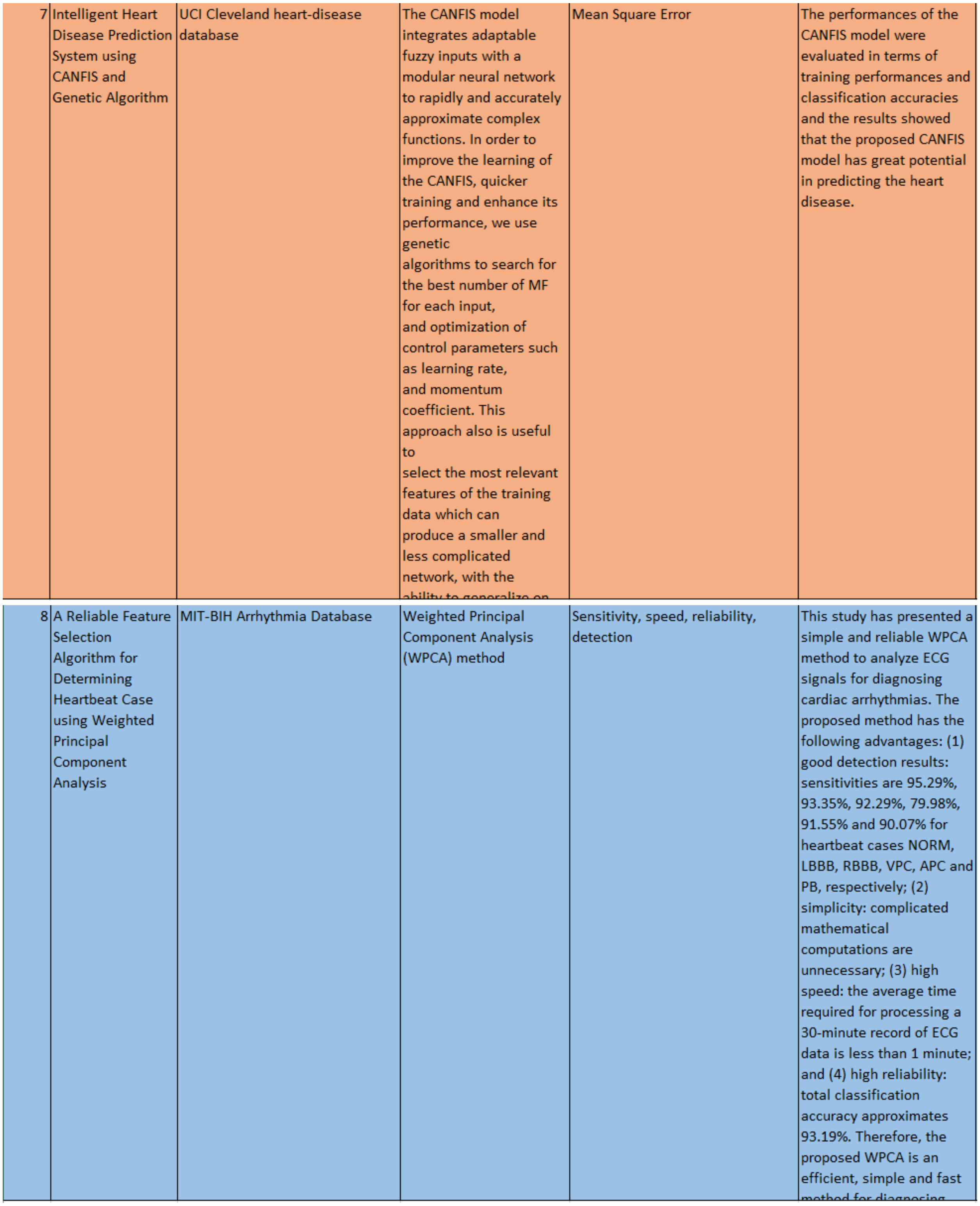
3. Literature Survey



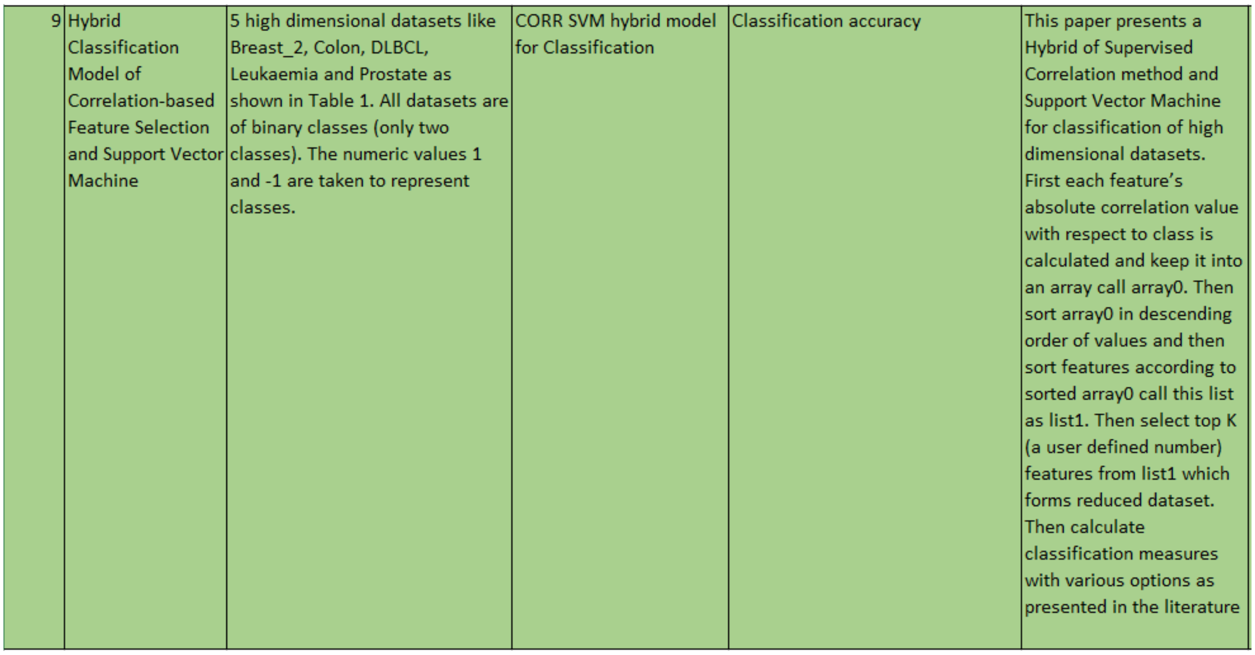
S. No	Title	Dataset Used	Methodologies Used	Metrics used	Interpretation of Results
1	Identification of significant features and data mining techniques in predicting heart disease, Telematics and Informatics	UCI Machine Learning Repository - Cleveland dataset and UCI Statlog Heart disease dataset	k-NN, Naïve Bayes, Vote, Support Vector Machine and Neural Network	Accuracy, F-measure and Precision. Accuracy is the percentage of correctly predicted instances among all instances. F-measure is the weighted mean of the precision and recall. Precision is the percentage of correct predictions for the positive class.	Precision and accuracy of prediction was highest for K-NN, followed by Vote, Naïve Bayes, SVM, Neural Network
2	Prediction of coronary heart disease using machine learning: An experimental analysis	South African Heart Disease which is a subset of a larger dataset. It contains 462 instances (observations) and 10 attributes in all (shown in Table 1), of which 9 are independent factors and 1 variable, i.e. CHD is the dependent variable or labelled class. The dataset is a retrospective sample of males in a heart-disease high-risk region of the Western Cape in South Africa-KEEL [28] where the labelled class CHD has two predictive outcomes: positive (1) and negative (0).	Decision Tree, Naïve Bayes Algorithm, SVM	The performance of the classification models derived by the ML is measured using the confusion matrix. The confusion matrix is a contingency table that displays the number of instances assigned to each class thus allowing us to calculate the classification accuracy, sensitivity, specificity, true positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs) among others	NB achieved the highest accuracy amongst the three models. SVM and DT J48 outperformed NB with a Specificity rate of 82% but proved to have an unacceptable Sensitivity rate of less than 50%. While NB Algorithm didn't reach the threshold of 80% Specificity and Sensitivity rate, it did turn out to be the best classifier for the considered dataset as its predictive rate is better that those of J48 and SVM algorithms at least on the considered dataset.
3	Combination data mining methods with new medical data to predicting outcome of Coronary Heart Disease	The dataset consisted of 1000 consecutive patients who underwent coronary angiography for known coronary atherosclerosis at Anzhen hospital, capital University of Medical Sciences, Beijing from August 2005 to December 2005.	SVM, Artificial Neural Networks (ANN)	Three performance metrics were employed: accuracy, sensitivity and specificity. A distinguished confusion matrix is obtained to calculate the three measures. Confusion matrix is a matrix representation of the classification results. the upper left cell denotes the number of samples classifies as true while they were true (i.e., true positives), and lower right cell denotes the number of samples classified as false while they were actually false (i.e., true false). The other two cells (lower left cell and upper right cell) denote the number of samples misclassified.	The survey was done on a data of 1000 CHD cases with 11 attributes. In this research, they defined survival as any incidence of CHD where person is still alive after 6 months from the date of diagnosis. Theyused a binary categorical survival variable, which was calculated from the variables in the raw dataset, to represent the survivability where survival is represented with a value of "1" and non-survival is represented with "0". The aggregated results indicated that the SVM performed the best with a classification accuracy of 92.1%, the ANN model (with multi layered perceptron architecture) came out to be second best with a classification accuracy of 91.0%. The results showed here make clinical application more accessible, which will provide great advance in healing CHD.



4	Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques	Cleveland Heart Disease database consists of 303 records & Statlog Heart Disease database consists of 270 records	ANN, Naïve Bayes	A confusion matrix is obtained to calculate the accuracy of classification. It shows how many instances have been assigned to each class. In their experiment they have two classes, and therefore they have a 2x2 confusion matrix.	Three data mining classification techniques were applied namely Naive Bayes & Neural Networks. From results it has been seen that Neural Networks provides accurate results as compare to Naive Bayes.
5	Early Prediction of Heart Disease Using Decision Tree Algorithm	UCI Repository Dataset	Decision Tree (C4.5), Naïve Bayes	Accuracy, Sensitivity and Specificity are computed on the basis of the confusion matrix that is curated.	From results, it has been seen that Decision trees provides accurate results as compare to Naive Bayes. This system can be further expanded. It can use more number of inputs.
6	Coronary Heart Disease Diagnosis Through Self-Organizing Map and Fuzzy Support Vector Machine with Incremental Updates	UCI Repository for datasets of Cleveland and Statlog	Self Organizing Map, SVM	The experimental setup of our experiment on the dataset with missing values is as follows. The dataset is divided into two subsets, 70% of the dataset as training subset and 30% of dataset as test subset. Then we applied the imputation procedure through hot-deck and k-NN on the training set for missing value imputation. We then applied PCA on each cluster which does not include the missing values. In the final stage, the classification models were constructed by Fuzzy SVM. To obtain the classification accuracy, the Fuzzy SVM classification model was evaluated on the test set.	The results revealed that the dataset imputation has a positive relationship with the accuracy of the Fuzzy SVM classifier. In addition, we found that the methods which rely on PCA provide better accuracy in relation to the other methods. In fact, it was found that, in the medical dataset the multicollinearity can significantly affect the predictive accuracy of the classifiers. Our experimental findings on two datasets also showed that the use of the methods with incremental techniques can have advantages on enhancing the computation time of disease prediction.



7	Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm	UCI Cleveland heart-disease database	The CANFIS model integrates adaptable fuzzy inputs with a modular neural network to rapidly and accurately approximate complex functions. In order to improve the learning of the CANFIS, quicker training and enhance its performance, we use genetic algorithms to search for the best number of MF for each input, and optimization of control parameters such as learning rate, and momentum coefficient. This approach also is useful to select the most relevant features of the training data which can produce a smaller and less complicated network, with the ability to generalize on	Mean Square Error	The performances of the CANFIS model were evaluated in terms of training performances and classification accuracies and the results showed that the proposed CANFIS model has great potential in predicting the heart disease.
8	A Reliable Feature Selection Algorithm for Determining Heartbeat Case using Weighted Principal Component Analysis	MIT-BIH Arrhythmia Database	Weighted Principal Component Analysis (WPCA) method	Sensitivity, speed, reliability, detection	This study has presented a simple and reliable WPCA method to analyze ECG signals for diagnosing cardiac arrhythmias. The proposed method has the following advantages: (1) good detection results: sensitivities are 95.29%, 93.35%, 92.29%, 79.98%, 91.55% and 90.07% for heartbeat cases NORM, LBBB, RBBB, VPC, APC and PB, respectively; (2) simplicity: complicated mathematical computations are unnecessary; (3) high speed: the average time required for processing a 30-minute record of ECG data is less than 1 minute; and (4) high reliability: total classification accuracy approximates 93.19%. Therefore, the proposed WPCA is an efficient, simple and fast method for diagnosing



9	Hybrid Classification Model of Correlation-based Feature Selection and Support Vector Machine	5 high dimensional datasets like Breast_2, Colon, DLBCL, Leukaemia and Prostate as shown in Table 1. All datasets are of binary classes (only two classes). The numeric values 1 and -1 are taken to represent classes.	CORR SVM hybrid model for Classification	Classification accuracy	This paper presents a Hybrid of Supervised Correlation method and Support Vector Machine for classification of high dimensional datasets. First each feature's absolute correlation value with respect to class is calculated and keep it into an array call array0. Then sort array0 in descending order of values and then sort features according to sorted array0 call this list as list1. Then select top K (a user defined number) features from list1 which forms reduced dataset. Then calculate classification measures with various options as presented in the literature

The field of Heart-Disease analysis and predictions is one of the fields which has been widely researched upon be it in terms of data analysis and visualizations, predictions, etc. Table 1 given below gives a detailed study of the survey conducted by our team on the existing works and algorithms related to heart disease predictions.

4. Motivation

The main motivation behind the curation of this project is because of the impact and relevance of the problem statement we identified: to conduct heart disease data analysis and implement classification algorithms using cardiotocographic data and to classify patients as pathological and non-pathological suspects of heart diseases. Based on the literature review done so far, we have observed that most of the research works have limited themselves to implementing two or three classification/regression algorithms. In order to provide a holistic and detailed comparative analysis on the various available ML classification algorithms with respect to their accuracy and precision, we have decided to take up this problem domain and fill in the research gap.

5. Methodology

In order to curate the implementation of the project, the first and foremost step was to acquire the data needed for the analysis and execution of algorithms. Dataset acquisition is extremely important and a crucial step in our project. The dataset we have used is the dataset obtained from the UCI Global repository for heart disease analysis. It has the following attributes (the respective indication of each attribute is also mentioned beside each attribute):

- slope (type: int) - the slope of the peak exercise ST segment, an electrocardiography read out indicating quality of blood flow to the heart – Value 1: upsloping, Value 2: flat, Value 3: downsloping
- thal (type: categorical) - results of thallium stress test measuring blood flow to the heart, with possible values – 1: normal, 2: fixed defect, 3: reversable defect
- trestbps (type: int) - resting blood pressure
- cp (type: int): chest pain type (4 values), Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic
- fbs (type: binary) - fasting blood sugar > 120 mg/dl
- restecg (type: int) - resting electrocardiographic results (values 0,1,2)
- oldpeak (type: float) - oldpeak = ST depression induced by exercise relative to rest, a measure of abnormality in electrocardiograms
- sex (type: binary) - 0: female, 1: male
- age (type: int) - age in years
- exang: exercise_induced_angina (type: binary) - exercise-induced chest pain (0: False, 1: True)
- chol (type: int) - serum cholestoral in mg/dl
- thalach (type: int) - maximum heart rate achieved
- ca (type: categorical) - number of major vessels (0-3) colored by flourosopy
- target (type: categorical) - diagnosis of heart disease (angiographic disease status) – Value 0: < 50% diameter narrowing, Value 1: > 50% diameter narrowing

Following the data acquisitions, the next step is to load all suitable libraries and packages which would provide all the required functions. The dataset is then loaded into the rmd file and data cleaning takes place. Data cleaning is an important and essential step before moving on to proceed with the further operations. The data is checked for any empty fields and data cells and a few columns and records are mutated without any data loss for easy comprehension and readability of data. Once the data transformation and cleaning are completed, the next step is to visualize the data in order to deduce inferences from the same. Figure 1 to Figure 5 represent the visualization of certain important aspects and parameters of the dataset.

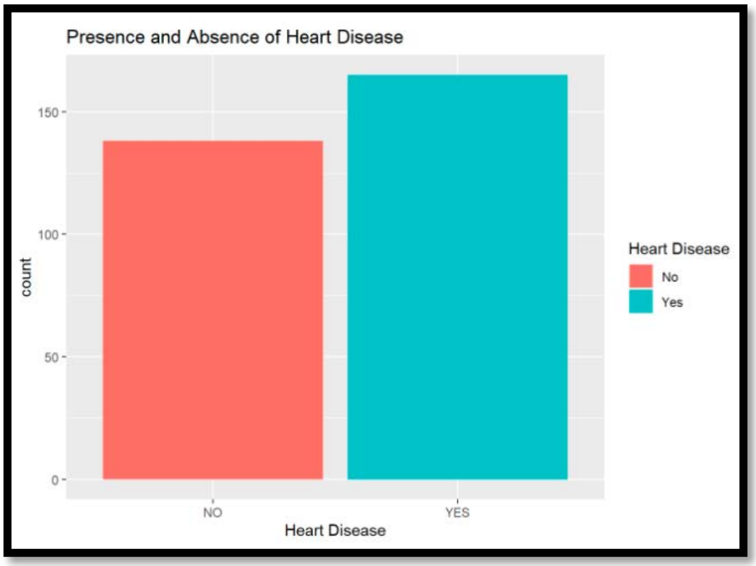


Figure 1. Number of people with and without heart disease

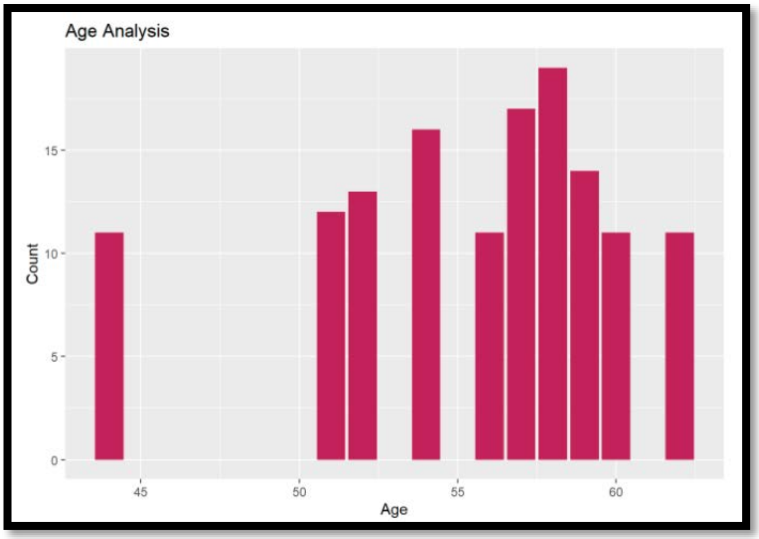


Figure 2. Age distribution of people with heart disease

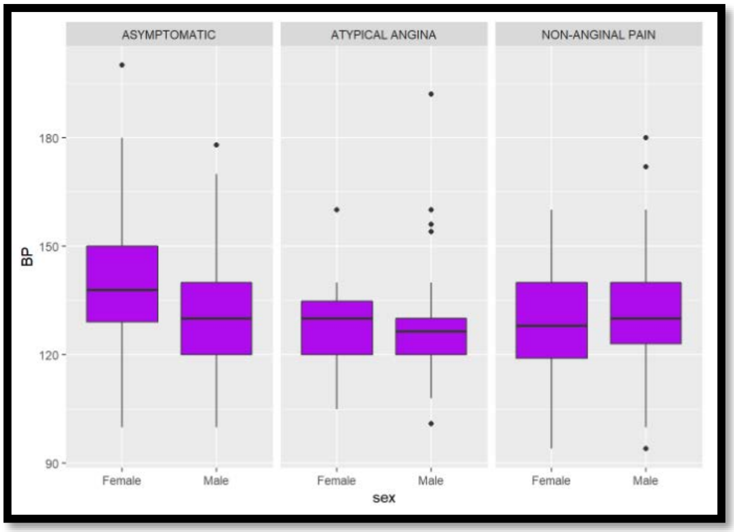


Figure 3. Box Plot of types of chest pain based on BP levels between male and female

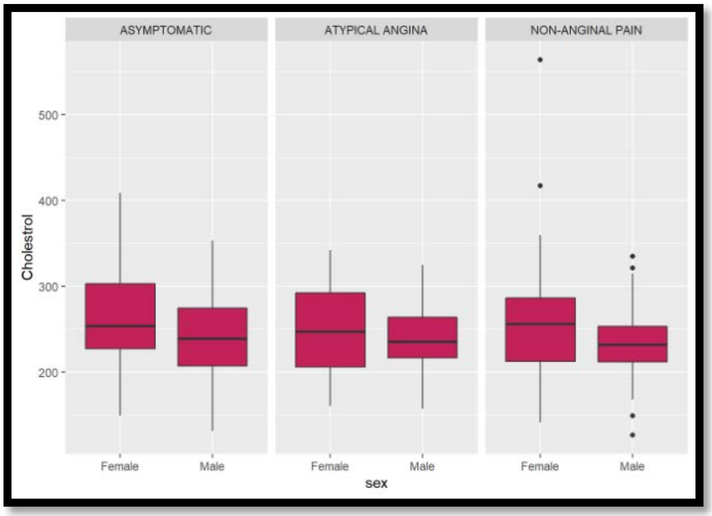


Figure 4. Box Plot of types of chest pain based on cholesterol levels between male and female

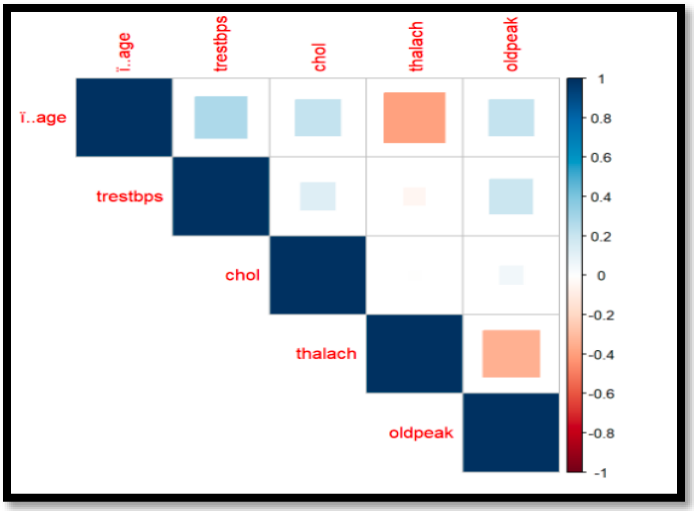


Figure 5. Correlation Plot between various attributes of the dataset

After the analysis and visualization of the data, the next step is to create models for each of the classification algorithms and to train and test the models. These have been further explained in the following section.

6. Proposed Work / Implementation

In order to analyse and comprehend the various classification algorithms and prediction accuracy, we have chosen six suitable classification algorithms. The concept and computation behind each of the algorithms has been elaborated as follows:

i. Linear Discriminant Analysis (LDA)

LDA is an algorithm which is used to acquire group means and averages and for each individual data object, it computes the probability with which it belongs to a certain group which has been predicted. The necessary functions of it can be found in the MASS package of R Studio. It was initially founded by R A Fisher in 1936 to classify the subjects into one of the two possible groups. However, over the years, various developers and researchers have expanded this and it is possible to classify into more than two classes. The key concept behind LDA is to apprehend the linear combination of original variables that would provide the most likely separation and classification between groups. Figure 6 shows the step-by-step process of the LDA algorithm starting from acquiring the input samples, splitting them into training and testing, extracting necessary features, projection to lower dimensional spaces and then classifying.

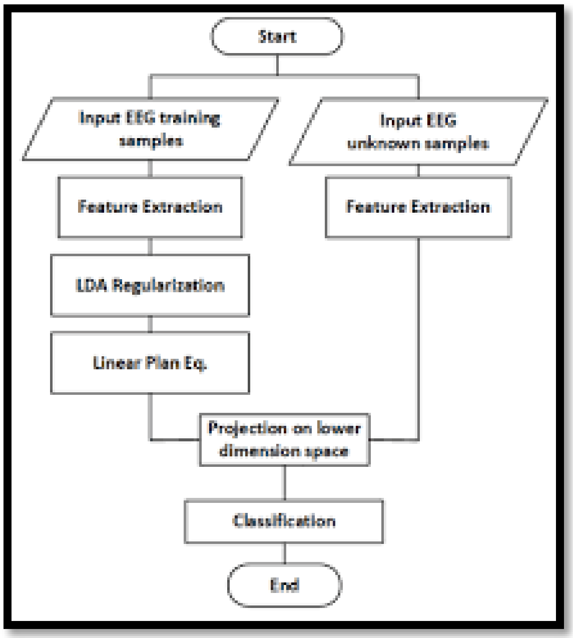


Figure 6. Work Flow of Linear Discriminant Analysis

ii. Quadratic Discriminant Analysis (QDA)

QDA is an enhanced variant of LDA which makes use of an individual covariance matrix for estimating the class of each data object. It proves to be an extremely useful technique especially when there's a predefined notion regarding the covariances of each of the distinct classes. However, this technique cannot be used for dimensionality reduction. In QDA, there's a p x p transformation matrix for each class k and it is these matrices that ensure the within-group covariances are spherical in nature. Figure 7 shown below, gives the step-by-step work flow of QDA starting from data acquisition to feature extraction to pattern recognition and then classification.

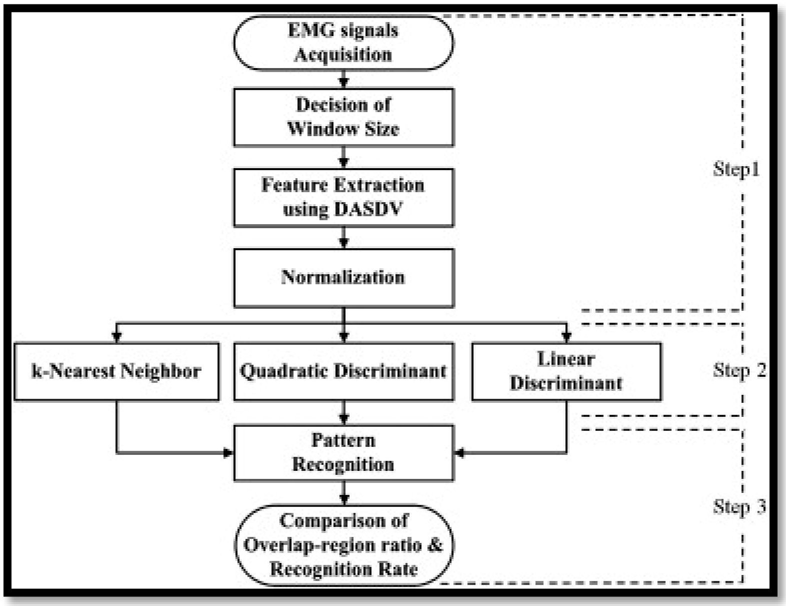


Figure 7. Work Flow of Quadratic Discriminant Analysis

iii. K-Nearest Neighbour (KNN)

This is one of the lazy learning algorithms which maps and keeps a track of all the objects in the training set which usually belong to an n-dimensional space. With the help of the labelled data objects, it labels the other unclassified objects using how closely similar their attributes and features are. It's a highly robust technique which is apt for data which is likely to contain more noise and works great with massive datasets too even though it may be computationally heavier and complex. Figure 8 elaborates on the work flow and procedures of using KNN Classification algorithm.

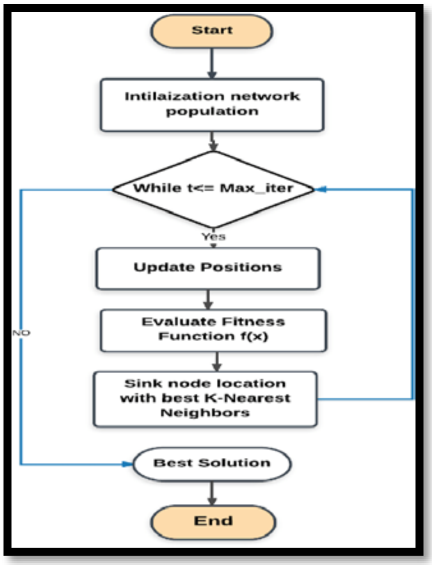

Start
Intilaization network population
While t<= Max_iter
Yes
NO
Update Positions
Evaluate Fitness Function f(x)
Sink node location with best K-Nearest Neighbors
Best Solution
End

Figure 8. Work flow diagram of KNN Algorithm

iv. Support Vector Machines (SVM)

SVM is used to represent data objects in the form of points in an empty space. Based on this representation, the algorithm then computes a function which would split the space as per the target classes and labels which have been identified. The dataset is split into training and testing and SVM uses the training set to plot the objects in space and to tweak the function in such a way that it splits the space accurately. Once the function is finalized, it places the objects in different parts of the space depending on which class they fall into. SVM's are very lightweight and highly efficient in higher dimensional spaces. Figure 9 shows the executional steps of SVM algorithm.

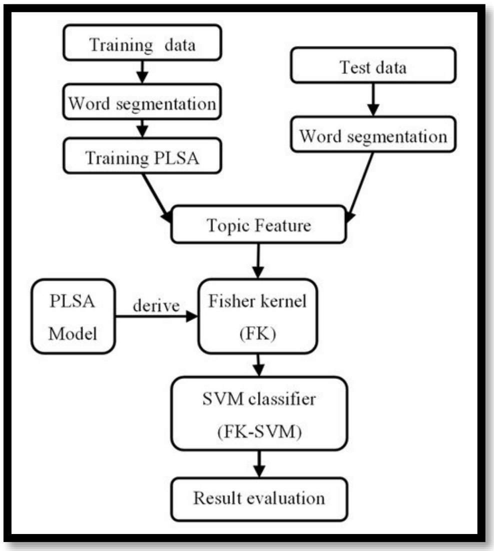


Figure 9. Work flow diagram of SVM Algorithm

v. Random Forests (RF)

RF algorithms is essentially an ensemble of decision tree algorithms as it typically combines and juxtaposes various decision trees in order to get an accurate and precise prediction of classes. This algorithm is one of the non-linear techniques. A novel observation is fed to each of the created trees individually and based on the majority of the class predicted for the observation, the final class is decided upon. An error estimate is made for the cases which were not used while building the tree. That is called an OOB (Out-of-bag) error estimate which is mentioned as a percentage. Figure 10 details out the execution of steps involved in the random forest algorithm starting from training each of the decision tree models to testing the decision tree with a sample data and finally voting on the most predicted class for the observation.

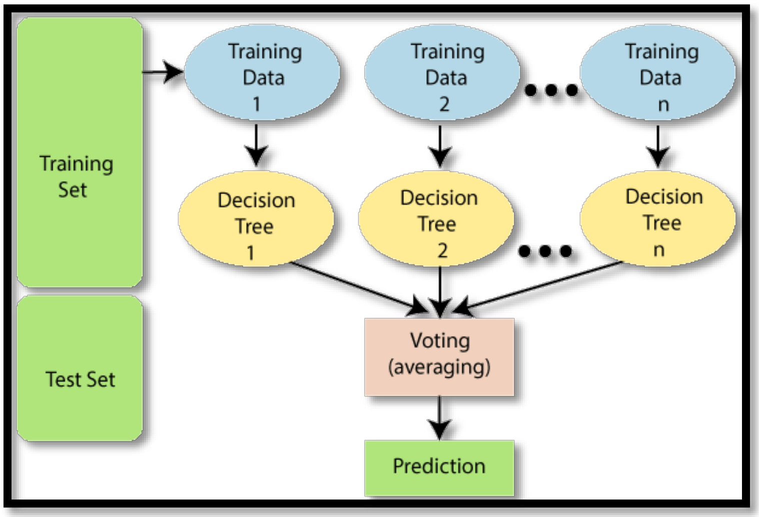


Figure 10. Work flow diagram of Random Forests Algorithm

vi. Gradient Boosting (GB)

Boosting is one of the other commonly used ensemble techniques which is not concerned with the reduction of variance of the learning phase. Boosting may often tend to be inclined towards biasness based on simple or weak learners. Weak learners consistently learn because of which there's a higher chance for it to improve its accuracy and performance with time. Gradient boosting is one such algorithm which has its underlying foundation of the boosting concept with the soul difference being that it identifies hard objects and examples by computing the large residual values calculated in the preceding iterations. It initially builds one learner to predict the values/labels of samples and calculate the loss (the difference between the outcome of the first learner and the real value). It will build a second learner to predict the loss after the first step. The step continues to learn the third, fourth etc until a certain threshold. Figure 11 elaborates further on the process involved in gradient boosting classification.

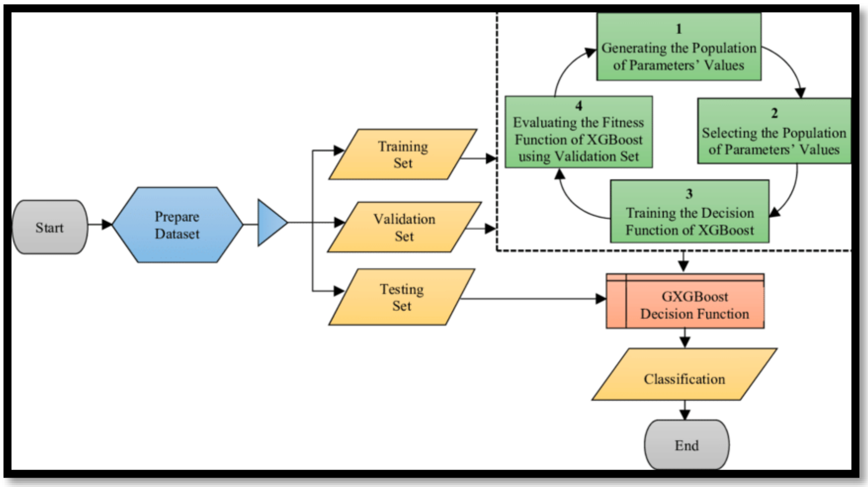

1
Generating the Population of Parameters' Values
2
Selecting the Population of Parameters' Values
3
Training the Decision Function of XGBoost
4
Evaluating the Fitness Function of XGBoost using Validation Set
Start
Prepare Dataset
Training Set
Validation Set
Testing Set
GXGBoost Decision Function
Classification
End

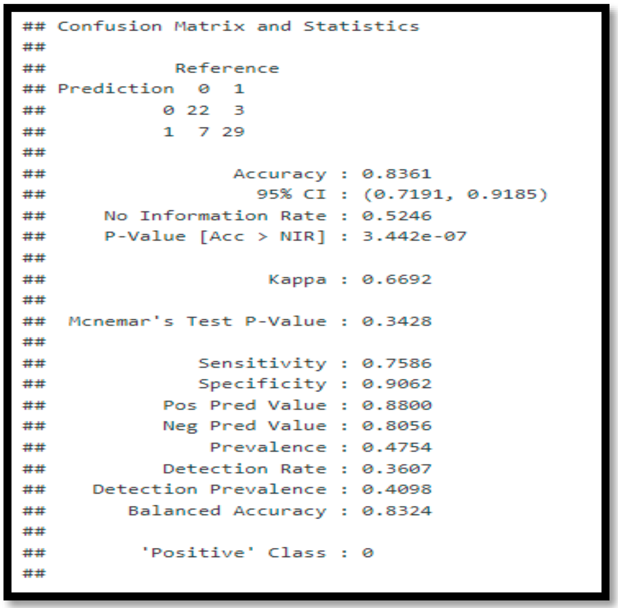
Figure 11. Work flow diagram of Gradient Boosting Algorithm

7. Results

After implementing the above-mentioned algorithms in order to analyse and predict if the data object as per our dataset is a pathological suspect or not, we were able to draw quite a few conclusions. For LDA, it was inferred that the accuracy is 0.8689 which is pretty good and the specificity and sensitivity values are 0.9062 and 0.8276 which means the true negative and true positive rates are good enough (results shown in Figure 12). In the case of QDA also we got the same result as that of LDA (results shown in Figure 13). For KNN algorithm, the accuracy is 0.8361 which is pretty good and the specificity and sensitivity values are 0.8750 and 0.7931 indicating a decent rate of true negatives and true positives (results shown in Figure 14). For SVM, as per the confusion matrix obtained, the accuracy rounded off to around0.8361 with a specificity and sensitivity of around 0.9375 and 0.7241 (results shown in Figure 15). Using RandomForest, the accuracy obtained is 0.7869 along with specificity and sensitivity values of 0.8438 and 0.7241 (results shown in Figure 16). Finally, with gradient boosting algorithm, we got an accuracy of 0.8033 and the specificity and sensitivity values are 0.8438 and 0.7586 (results shown in Figure 17). Table 2 given below shows the comparison of accuracy, specificity and sensitivity of each of the algorithms when tested with our dataset.

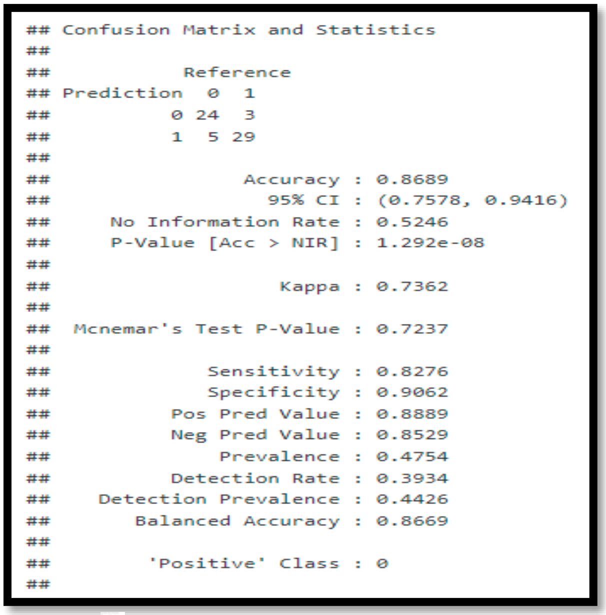
Algorithm Used	Accuracy	Specificity	Sensitivity
LDA	0.8361	0.7586	0.9062
QDA	0.8689	0.9062	0.8276
KNN	0.8361	0.8750	0.7931
SVM	0.8361	0.7241	0.9375
RF	0.7869	0.8438	0.7241
GB	0.8033	0.8438	0.7586

Table. 2 Results obtained from executing each algorithm



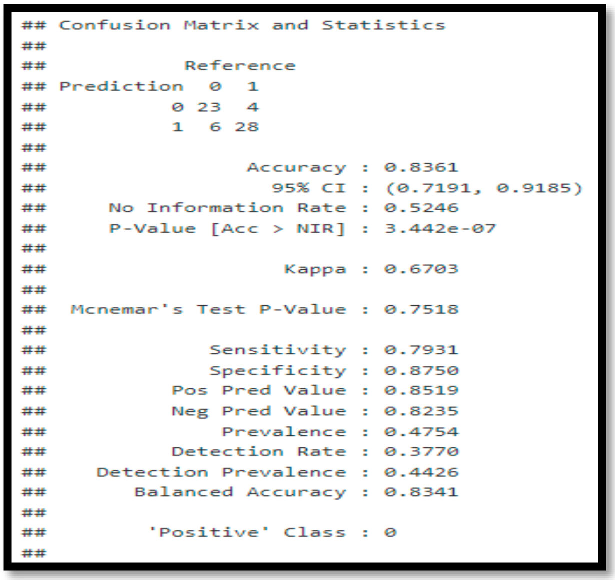
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 22  3
##          1  7 29
##
##                Accuracy : 0.8361
##                  95% CI : (0.7191, 0.9185)
##     No Information Rate : 0.5246
##     P-Value [Acc > NIR] : 3.442e-07
##
##                   Kappa : 0.6692
##
##  Mcnemar's Test P-Value : 0.3428
##
##             Sensitivity : 0.7586
##             Specificity : 0.9062
##          Pos Pred Value : 0.8800
##          Neg Pred Value : 0.8056
##              Prevalence : 0.4754
##          Detection Rate : 0.3607
##    Detection Prevalence : 0.4098
##       Balanced Accuracy : 0.8324
##
##        'Positive' Class : 0
##
```

Figure. 12 Results of LDA



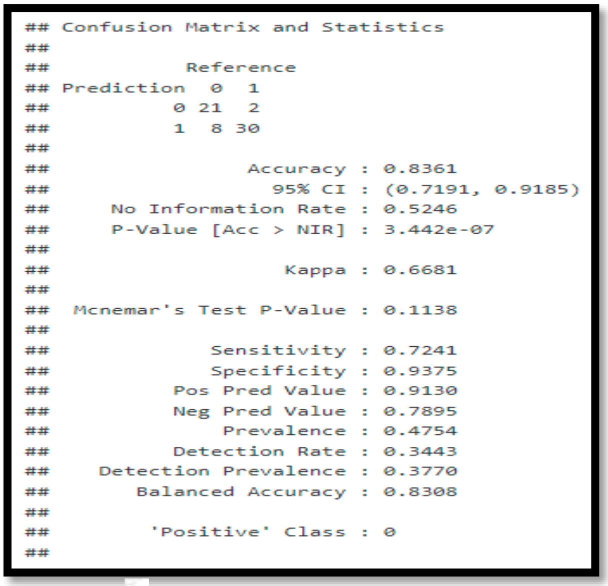
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 24  3
##          1  5 29
##
##                Accuracy : 0.8689
##                  95% CI : (0.7578, 0.9416)
##     No Information Rate : 0.5246
##     P-Value [Acc > NIR] : 1.292e-08
##
##                   Kappa : 0.7362
##
##  Mcnemar's Test P-Value : 0.7237
##
##             Sensitivity : 0.8276
##             Specificity : 0.9062
##          Pos Pred Value : 0.8889
##          Neg Pred Value : 0.8529
##              Prevalence : 0.4754
##          Detection Rate : 0.3934
##    Detection Prevalence : 0.4426
##       Balanced Accuracy : 0.8669
##
##        'Positive' Class : 0
##
```

Figure. 13 Results of QDA



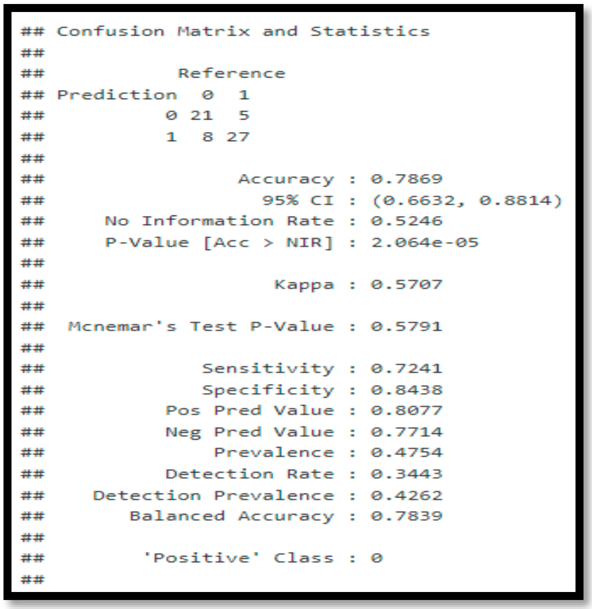
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 23  4
##          1  6 28
##
##                Accuracy : 0.8361
##                  95% CI : (0.7191, 0.9185)
##     No Information Rate : 0.5246
##     P-Value [Acc > NIR] : 3.442e-07
##
##                   Kappa : 0.6703
##
##  Mcnemar's Test P-Value : 0.7518
##
##             Sensitivity : 0.7931
##             Specificity : 0.8750
##          Pos Pred Value : 0.8519
##          Neg Pred Value : 0.8235
##              Prevalence : 0.4754
##          Detection Rate : 0.3770
##    Detection Prevalence : 0.4426
##       Balanced Accuracy : 0.8341
##
##        'Positive' Class : 0
##
```

Figure. 14 Results of KNN



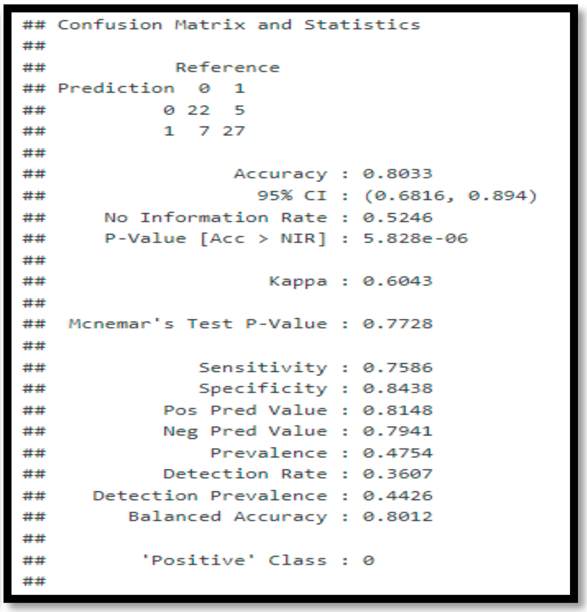
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 21  2
##          1  8 30
##
##                Accuracy : 0.8361
##                  95% CI : (0.7191, 0.9185)
##     No Information Rate : 0.5246
##     P-Value [Acc > NIR] : 3.442e-07
##
##                   Kappa : 0.6681
##
##  Mcnemar's Test P-Value : 0.1138
##
##             Sensitivity : 0.7241
##             Specificity : 0.9375
##          Pos Pred Value : 0.9130
##          Neg Pred Value : 0.7895
##              Prevalence : 0.4754
##          Detection Rate : 0.3443
##    Detection Prevalence : 0.3770
##       Balanced Accuracy : 0.8308
##
##        'Positive' Class : 0
##
```

Figure. 15 Results of SVM



```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 21  5
##          1  8 27
##
##                Accuracy : 0.7869
##                  95% CI : (0.6632, 0.8814)
##     No Information Rate : 0.5246
##     P-Value [Acc > NIR] : 2.064e-05
##
##                   Kappa : 0.5707
##
##  Mcnemar's Test P-Value : 0.5791
##
##             Sensitivity : 0.7241
##             Specificity : 0.8438
##          Pos Pred Value : 0.8077
##          Neg Pred Value : 0.7714
##              Prevalence : 0.4754
##          Detection Rate : 0.3443
##    Detection Prevalence : 0.4262
##       Balanced Accuracy : 0.7839
##
##        'Positive' Class : 0
##
```

Figure. 16 Results of Random Forests



```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 22  5
##          1  7 27
##
##                Accuracy : 0.8033
##                  95% CI : (0.6816, 0.894)
##     No Information Rate : 0.5246
##     P-Value [Acc > NIR] : 5.828e-06
##
##                   Kappa : 0.6043
##
##  Mcnemar's Test P-Value : 0.7728
##
##             Sensitivity : 0.7586
##             Specificity : 0.8438
##          Pos Pred Value : 0.8148
##          Neg Pred Value : 0.7941
##              Prevalence : 0.4754
##          Detection Rate : 0.3607
##    Detection Prevalence : 0.4426
##       Balanced Accuracy : 0.8012
##
##        'Positive' Class : 0
##
```

Figure. 17 Results of Gradient Boost

8. Conclusion

From the consolidated results and parameters obtained for analysis of these algorithms, it can be observed that in terms of accuracy, LDA, QDA and SVM leads the chart when compared to the others as they're able to classify the sample test data into their respective classes accurately. One of the other parameters considered is specificity. A test that has 100% specificity will identify 100% of patients who do not have the disease. A test that is 90% specific will identify 90% of patients who do not have the disease. Tests with a high specificity (a high true negative rate) are most useful when the result is positive. In our case, QDA tends to be the most precise algorithm in terms of specificity and SVM tends to be the least precise one. Sensitivity refers to a test's ability to designate an individual with disease as positive. A highly sensitive test means that there are few false negative results, and thus fewer cases of disease are missed. In terms of sensitivity, SVM tops the classification accuracy while LDA is at the bottom. Hence, we can conclude that based on the parameters the suitable algorithms can be used for prediction of pathological suspects.

References

1. https://ieeexplore.ieee.org/abstract/document/7567338
2. https://ieeexplore.ieee.org/abstract/document/7551594
3. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.451.9421&rep=rep1&type=pdf
4. https://link.springer.com/article/10.1007/s40815-020-00828-7

5. https://www.researchgate.net/profile/Safish-Mary/publication/315023624_Early_Prediction_of_Heart_Disease_Using_Decision_Tree_Algorithm/links/58c84b57aca2723ab16eba60/Early-Prediction-of-Heart-Disease-Using-Decision-Tree-Algorithm.pdf
6. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.258.8158&rep=rep1&type=pdf
7. https://ieeexplore.ieee.org/abstract/document/4420369
8. https://dl.acm.org/doi/pdf/10.1145/3342999.3343015?casa_token=EQ63fwch8XoAAAAA:L6EK91Udq48GClSOWPgHDdAnE1-PTEZd6ZZ416bs4dF67u92qLlSxUanXUaRd1ZuJlFd4nyto8MGyw
9. https://www.sciencedirect.com/science/article/abs/pii/S0736585318308876