Prediction of Heart Disease using Data Mining Techniques

Sujay Ravi, Binghamton University, sravi2@binghamton.edu

***Abstract*— Heart Disease contributes significantly towards the death of patients in the healthcare industry. There are tons of patient's data being collected every day. Now with the prominence of the machine learning algorithms it is easy to predict the heart disease at an earlier stage. In this paper, classification model specific feature selection method is applied to select the important features. The classification models used are Logistic Regression, Support vector machine, Naïve Bayes, Decision Tree, Random forest, Neural Networks and ensemble methods Boosting, Bagging, Vote, Stacking. The data set used for prediction are Cleveland and Statlog data set from UCI Machine Learning Repository. The software used for developing models is Python. The result showed stacking with logistic regression gave highest prediction accuracy of 85.48%.**

***Index Terms*— UCI Machine Learning Repository, Heart Disease, Python, Classification Algorithms.**

I. Introduction

It is a condition which occurs when the plaque starts to accumulate in the arteries. Based on the statistics according to the race of ethnic group American Indians or Alaska Natives deaths due to heart disease is 18.4%, Asians or Pacific Islanders 22.2%, Non-Hispanic Blacks and Non-Hispanic Whites 23.8% and all other groups it is 23.5% [1]. Heart disease is one of the major contributors for the death of people in the United States. It is estimated that nearly 600,000 Americans die each year due to heart disease. There are several type of heart conditions. The common ones are coronary artery disease and some other causes include valve in the heart, problem with pumping of the heart which can cause heart attack. This disease can be found from younger child to aged person. Some of the causes of heart disease are unhealthy diet, smoking, improper exercise. Other factors include high cholesterol, diabetes and high blood pressure. Some of the symptoms of heart disease include shortness of breath, pain in the shoulder, nausea, chest pain that usually remains for long period of time. There are several tests which are performed by the doctors to diagnose heart disease which include coronary angiograms, ECG, X-rays [2]. There are lots of data produced by the healthcare industry. But many doctors currently use their own perception on prediction of the heart disease for patients. This system of prediction is not always providing accurate results and it also takes lot of time. With the advancement in the technology there are more efficient methods which have been developed for the prediction of the heart disease. The popular one is the data mining technique. It is a recently developed technology which was introduced in 1994 [3]. The term data mining indicates that data is an important factor because with no data there is no mining. Traditional data mining techniques include discriminant analysis, regression analysis etc. and non-traditional data analysis methods include decision trees, neural networks and association analysis. There is a high potential of data mining in the field of healthcare industry. They can be grouped as management of healthcare; evaluation of effectiveness of treatment; detection of abuse; and customer relationship management [4]. This technique is transforming many industries like Telecommunication, Retail, Finance and Science and Engineering. Data mining techniques can be combined with frameworks of cloud computing including Hadoop, MapReduce etc. which can be used for the big data analytics in applications like adaptive ranking of web page, driving patterns which are helpful in the E-commerce [5]. This paper tends to use the different data mining techniques which are commonly used for the prediction of the heart disease. These techniques include Logistic Regression, Naïve Bayes, Support Vector Machine, Decision Tree, Neural Networks and ensemble techniques like Random forest, Vote, Boosting, Bagging and Stacking. The feature selection method used here is Recursive feature elimination.

II. Literature REVIEW

There have been many researches works which have been taken place using different feature selection and classification approach. Some of them have been discussed below.

Artificial Neural Networks were applied for the Heart Disease Prediction System and classified the heart disease based on 13 attributes. The research was based on data from UCI machine learning repository with 303 instances [6].
Three techniques were used to compare the accuracy using Hungarian Institute of Cardiology. Here 11 attributes were selected for experiment. The software used here for conducting the research was WEKA tool. The accuracy resulted were 85.3% for bagging, 84.35% for J48 decision tree and 82.31 for Naïve Bayes [7].

Arrhythmia classification with the feature dimension reduction technique Linear Discriminant Analysis and Support Vector Machine as the classifier was used. The reduced features with linear discriminant analysis applied to the Support Vector Machine showed better results than Principal Component Analysis to reduce features [8].

Prediction of neonatal disease diagnosis with backpropagation Neural network. It finds a pattern for diagnosis and prediction of neonatal disease. The neural network was trained based on different categories of neonatal disease and the accuracy was 75% [9].

Comparative study of the data mining techniques in cardiovascular disease prediction. The data set is Cleveland cardiovascular taken from UCI repository. The classifiers used were RIPPER, Decision Tree, Artificial Neural Network and Support Vector Machine. Accuracy of classifiers are 81.08% (RIPPER), 79.05 % (Decision Tree), 80.06 % (Artificial Neural Networks), and 84.12% (Support Vector Machine). Support Vector machine has highest accuracy among the classifiers [10].

Ensemble classification techniques boosting, bagging and majority vote applied to Cleveland data from UCI machine learning repository. The classifiers used were Bayes Net, Naïve Bayes, C4.5, Multilayer Perceptron and PART. Feature selection method used Brute force method. The accuracy of weak classifier was increased by 7% [11].

Neural networks ensembles based effective diagnosis of heart disease. The new models are created with combination of posterior probabilities obtained from predecessor models. Software used was SAS base software. The data is of Cleveland heart disease. The accuracy, sensitivity and specificity obtained are 89.01%, 80.95% and 95.91% [12].

Naïve Bayes used for heart disease prediction system. This approach classifies the data into five categories namely no, low, average, high and very high. It can also predict the heart disease when unknown sample is given as input. This prediction system can handle different data sets by changing the name of the data file [13].

Random Forest, Decision Tree and Naïve Bayes are used for prediction of the heart disease. The data is taken from the StatLog dataset from UCI repository. The best result was achieved by Random Forest with a precision of 81%. [14].

BagMOOV, bootstrap aggregation with multi-objective optimized voting. It involves ensemble of 5 classifiers namely Naïve Bayes, Linear Regression, Quadratic Discriminant Analysis, instance-based learner and Support Vector Machine. Data sets used are SPECT, SPECTF, Heart disease and Statlog Data set from UCI machine learning repository and Eric data set from ricco. The accuracy of the ensemble method was highest for the Cleveland data set with accuracy, sensitivity, specificity and f-measure of 84.16%, 93.29%, 96.70% and 82.15% [15].

Nine Voting Equal Frequency Discretization Gain Ratio Decision Tree for diagnosing heart disease patients. Here a range of techniques were applied to different Decision Trees. The data is taken from the Cleveland Clinic Foundation Heart disease data set. The sensitivity, specificity and accuracy based on equal width, equal frequency, chi merge and entropy with Info Gain, Gini Index and Gain ratio decision trees were tested with nine voting decision trees. The model resulted in sensitivity of 77.9%, specificity of 85.2% and accuracy of 84.1% [16].

Prediction of heart disease by identifying significant features and data mining technique. The software used is RapidMiner. Seven different classifiers namely k-NN, Naïve Bayes, Decision Tree, Logistic Regression, Vote (Naïve Bayes and Logistic Regression), Support Vector Machine and Neural Network. The feature selection method used here is Brute force method with a lower bound of 3 features. Its data used are Cleveland dataset from UCI machine learning repository and UCI Statlog Heart disease dataset. Vote with significant 9 features resulted in accuracy of 87.41% [17].

Least Square Twin Support Vector Machine approach for diagnosis of heart disease. The feature selected based on F-score which is used to calculate the weights of each feature. The data set used is heart-Statlog disease dataset. The proposed model resulted in accuracy of 85.59% with 11 significant features selected [18].

C4.5 and Fast Decision Tree used for accurately prediction of heart disease with the selection of common features from different data sets. The software used Weka tool. The data set used were Cleveland, Hungarian, Long Beach VA, Statlog project. The classification accuracy of the collected data set is greater than average of the separate data set accuracy [19].

The research used Naïve Bayes, IBK and Random Forest for detection of heart disease. The software used is Waikato Environment for Knowledge Analysis. It has used Statlog Heart Disease dataset taken from University of California, Irvine (UCI) machine learning database. Testing and training are done with 10-fold cross validation technique. Naïve Bayes shows 16.29% are having heart disease and 83.7% are not having any symptoms. IBK shows results of 75% are not affected by heart disease and 24% having heart disease. Random Forest shows 81% of healthy patients and 18% of patients with possible heart attack [20].

III. SOFTWARE

The tools used for prediction of the heart disease is Python. It is one of the most popular programming languages for scientific computing. The library used here are scikit-learn, Mlxtend, Matplotlib, Pandas, Keras. Scikit-learn has a wide range of machine learning algorithms basically for the supervised and unsupervised learning. As it is based on the scientific Python ecosystem it can be applied outside the field of statistical data analysis [21]. Since everything in the data mining rely on data, so there is a need for data preparation library pandas. They have wide range of input and output formats like Excel, csv, NumPy, SQL etc. pandas have good querying possibilities and basic visualizations. The drawback of this library is that the syntax is a bit confusing [22]. Matplotlib used for data visualization. The python implementation of MATLAB plots is Matplotlib. It is written with low level and has lot of possibilities for customization. The syntax may be little confusing but once concepts are mastered then it is possible to make any kind of graph. It has

better flexibility compared to library seaborn which is built on top of the Matplotlib. Seaborn is relatively easier for the beginners to learn compared to Matplotlib [23]. Mlxtend is a new library which contains few algorithms which are basic. It is also a machine learning algorithm. The only library with stacking algorithms and association rule algorithms. Keras is a deep learning library which is built on TensorFlow. Keras is used for constructing Neural Networks in python. Github host the code for the Keras. They support convolutional neural networks and recurrent neural networks in addition to the standard neural network. Google Brain developed TensorFlow. TensorFlow has a very good documentation and a wide range of functionalities other than the basics. The code here is very customable as it is written in the low-level library but it is a bit harder to master. Keras is basically a high-level code. The customization of the code is harder here. It general the customization of the code is easier in the low level [24], [25]. Keras also allows to produce deep model on the web and smartphones of ios and Android [26].

IV. METHODOLOGY

A. Model Building Phase

1) The data set used is the Cleveland data set.
2) The data is first pre-processed.
3) Stratified 10-fold cross validation is used for splitting of data into testing and training.
4) Dimension reduction technique is applied to Neural Network.
5) Feature selection method is applied to Logistic Regression, Support Vector Machine, Naïve Bayes, Decision Tree and Radom forest.
6) For each classifier, based on the best features selected hyperparameter tuning is applied. This gives the best parameters for each classifier.
7) Based on the best feature and best parameters for each of classifier, the accuracy, precision, f1, sensitivity and selectivity are obtained.
8) Ensemble methods is applied to selected best features with best parameters.
9) Best two models are selected.

B. Model Testing Phase

1) Data set used is Statlog data set
2) Best two models are applied.
3) Proposed Model

V. MODEL BUILDING PHASE

A. Data Set

For this research two data sets have been used. The first one is Cleveland Heart disease data set taken from the UCI Machine Learning Repository [42]. This is shown in the Table 1. There are 13 attributes. The number of instances is 303. The attribute Num of value 0 represents the absence of heart disease and value of 1, 2, 3, and 4 represents presence of heart disease

TABLE I

ATTRIBUTE INFORMATION OF CLEVELAND DATA SET

S. No	Attribute	Description	Value
1	Age	Age of persons in years	29 to 79
2.	Sex	Gender of patient 1=Male, 0=Female	0,1
3.	Cp	Chest pain Type 1= Typical Angina 2=Atypical Angina 3=Non-angina 4=Asymptomatic	1, 2, 3, 4
4.	Trestbps	Resting Blood Pressure in mmHg	94 to 200
5	Chol	Serum Cholesterol in mg/dl	126 to 564
6	Fbs	Fasting blood sugar im mg/dl	0, 1
7	Restecg	Resting Electrocardiographic Results	0,1,2
8	Thalach	Maximum Heart Rate achieved	71 to 202
9	Exang	Exercise Induced Angina	0,1
10	Oldpeak	ST depression induced by relative to rest	1 to 3
11	Slope	Slope of the Peak Exercise ST segment	1, 2, 3
12	Ca	Number of major vessels colored by fluoroscopy	0 to 3
13	Thal	3=Normal, 6=Fixed Defect, 7= Reversible Defect	3, 6, 7
14	Num	Class	0, 1, 2, 3, 4

B. Pre-processing

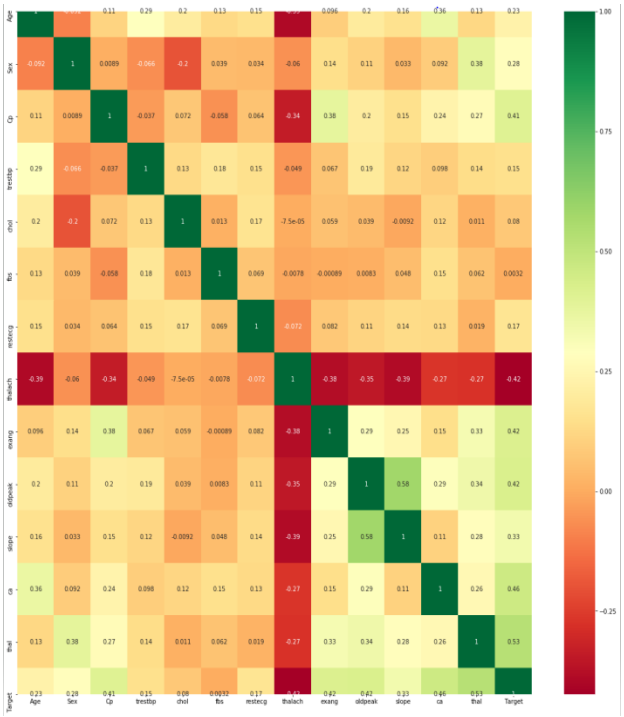
This is an important step before applying our data set to the different classification algorithms for the purpose of classification. In the Cleveland data set the there are a total of 303 instances. It can be seen that there are 6 missing values in the data set. Since it is a data set related to the patient's heart disease it should be handled with care. Here the 6 missing data sets are removed. This leave with 297 instances after the removal of the data. The dependent variable Num which is in the range of 0 to 4 is converted 0 and 1. The values greater than 0 are converted to 1 which indicates the presence of the heart disease. The data which is fed to the classifier's algorithms Neural Networks, Logistic Regression, Support Vector Machine, Decision Tree and Radom forest are standardized. The function used for the purpose of standardization in python is Standard scalar with range 0 and 1. For the Naïve Bayes classifier, the numerical values are transformed into categorical values in range of 1 to 4. The features which were transformed were Age, trestbp, chol, thalach, oldpeak.

C. Heatmap

This is used to view the correlation between the different features visually. The seaborn heatmap function is used to build the heat map in python. The figure size is given as 20 by 20. What can be seen from the heatmap is that a dark green color is used to indicate the strong positive correlation and the strong negative correlation is shown with the color of dark red.

The value ranges between +1 and -0.25. The strength within the features can be inferred from the heatmap. It can be seen that the features slope and oldpeak has the highest positive correlation of value 0.58, followed by feature thal and sex of value 0.38 and the feature exang and cp with a value of 0.38. The higher the value of the correlation between two features then higher the chances of redundancy. In, general sense the value of correlation lesser than 0.5 is better, which means lesser chances of redundancy. Since the highest correlation in this data set is close to 0.5 it is no bad, but should have a note

Fig. 1. Heatmap



of the features slope and oldpeak. This map is very much useful in making a sense of which of the features contribute significantly in the prediction of heart disease. The top features which has positive correlation with the Target variable in order are thal (0.53), ca (0.46), oldpeak (0.42), exang (0.42), cp (0.41), slope (0.33), sex (0.28), Age (0.23), restecg (0.17), trestbp (0.15), chol (0.08), fbs (0.0032). The only feature which has a strong negative co-relation with the target variable is thalach (-0.42).

D. Splitting data

The method used to split the data into training and testing data is Stratified cross-validation. It helps in taking equal proportion of the class labels in each of the folds. The parameter used is n_splits it indicates the number of folds required. The number of folds is taken as 10.

E. Classification Algorithm

Classification algorithms are frequently used by machine learning and data mining researches. It is used to predict the categorical class variables based on the other independent variables. The different classification algorithms used for the purpose of research in this paper are discussed as follows.

1) Logistic Regression

Logistic Regression is one of the special cases of the regression models. It is used when the output variable to be predicted is categorical variable. It is used to explain the relationship between a categorical output variable and a mixture of continuous and categorical predictors. This method gained attention from many researches since the year 1988. In general, the logistic regression model results are represented as logit, odds ratio, relative task, marginal probability etc. [30]. The logistic regression model is represented by the following mathematical representation.

$y = b_0 + b_1 x_1 + b_2 x_2 + ... + b_p x_p$ [31]

y is the estimated output value for true Y. b_0 is constant, $b_1 \ldots b_p$ are the parameters estimated corresponding to predictors $x_1 \ldots x_p$. $b_1 \ldots b_p$ are the slopes or regression weights.

Logistic regression is a special case of generalized linear models and is similar to the linear regression. But the assumption of these two are quite different. The difference is that in the logistic regression is that the conditional distribution used is Bernoulli distribution but it is Gaussian distribution for the linear regression as the dependent variable in the logistic regression is binary. Then the predicted values are in terms of probabilities and are in the form of (0, 1) using the logistic distribution function as the logistic regression predicts the probability of a particular outcome [32].

In this experiment binary logistic regression model is used with value 0 representing absence of the heart disease and 1 representing the presence of the heart disease.

2) Support Vector Machine

Support Vector Machine is a machine learning algorithm used for the purpose of classification. It is one of the powerful methods for building the classifier. It has the ability to process the multi-dimensional data. The data points are first classified and then mapped into –n-dimensional space. Support Vector machine consists of a hyperplane which is used to separate two classes of data points. The hyperplane is crated in such a way that closest data points from the hyperplane is as far as possible. The Support Vector Machine has a kernel function which can be used separate the data which is hard. This is done with the help of the transforming to data. In general, the choice of the kernel function can greatly affect the performance of the Support Vector Machine. Generally, the best parameters are selected based on the trial and error method. The best kernel also changes depending on the type of problem one is trying to solve. The type of method for searching the best parameters are discussed in the hyperparameter tuning section below. [33].

3) Naïve Bayes

Naïve Bayes is another type of prediction technique used for classification. It is a probabilistic classifier. It uses the Bayes theorem for solving the problems and process the results from the given data. For example, let A and B be two independent events. Based on the Bayes theorem the conditional

probability of the event A given by B is P (A|B) = [P (B|A) *P (A)]/P (B). The assumption of Naïve Bayes is that the occurrence of an event is independent of the occurrence of the other events [34]. For example, it assumes that X_1, X_2,… X_n are conditionally independent of each other [34]. By using the Bayes theorem, the conditional probability of D given X_1, X_2,… X_n is P (D|X_1, X_2,…, X_n) = [P(X_1, X_2,…, X_n|D)*P(D)] / P(X_1, X_2, …., X_n). Naïve Bayes is widely used for the purpose of classification as it requires small amount of training data decide the parameters for classifications [36]. For this experiment MultinomialNB a type of Naïve Bayes is used for the purpose of the classification. In this all the independent variables should be categorical nature [37].

4) Decision Tree

This is one of the commonly used method in data mining. The decision tree consists of nodes (attributes), branches (decision rules) and leaf (outcome). In this the target value is predicted based on several input variables. So, here each of the node corresponds to one of variables given as input. Decision rules can be established with if-then clause, for instance if condition 1 and 2 are satisfied then the outcome 'X' is result with respect to 'Y'. For the problem considered each leaf node is expressed as the prediction of outcome [27]. The induction of the decision tree algorithms is function recursively. The root node is selected first based on an attribute. Then the efficient smallest tree is created based on the ability of the root node to split the data efficiently. For each of the split happening it tries to pare down a set of instances till each of them have same classification. The one with the most information gained is taken as the best split [28]. Decision tree can handle categorical and numerical data. It provides the results which are easy to interpret and it can also handle the missing data in the attributes by replacing it with the most probable value [29].

5) Random Forest

Random Forest is basically ensemble technique of decision tree. It makes predictions by averaging the predictions of several independent models. Some of the choices made when constructing a random forest method for splitting the leaf's, type of predictor to be used in each leaf and method of injecting randomness into the trees [38]. Some of the elements for random forest are discussed. A random vector Θk, for k^{th} tree is produced. It is independent of the previous random vectors but the distribution is similar. Then a tree is formed with the help of training set and Θk *which* results in the classifier $h(x,\Theta k)$. The input vector is x [39].

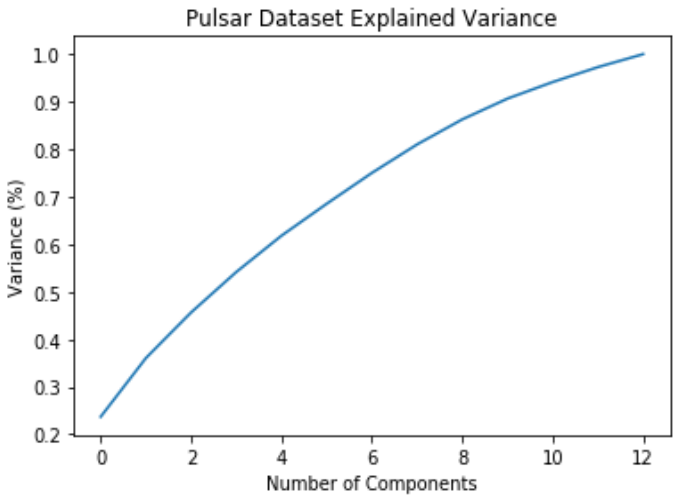
6) Neural Networks

A Neural Network system which are similar to biological neural networks usually of the animal brains. Neural Networks consists of connected nodes. A Neural Network consists of input, hidden and the output layers. The neural network here is built using the sequential model. The input, hidden and the output layer are built using Dense Library. There are 13 input features which can be fed to the neural network. The number of hidden layers is 2 with each hidden layer having 5 nodes. The activation function for the hidden layers used here is ReLu. Since it is a binary classification of heart disease sigmoid is used as an activation function. So, after creation of all the layers we need to compile the neural network. For compiling the optimizer and the loss function need to be set. Adam is used as optimizer. It stands for Adaptive moment estimation. It is combined in the form RMSProp + Momentum. To smooth the gradient descent momentum takes past gradients into consideration. The loss function between the actual output and the predicted output is calculated using binary_crossentropy as it is a binary model prediction. The model which is created is then fitted to the training data. For this the parameters batch size (10) and epochs (50) needs to be set. The batch size indicates the number of samples which are used per gradient update. The iteration over the entire data set is called as epoch [40].

F. Dimension Reduction Technique for Neural Networks

Dimension reduction helps in reduction of the number of features which can be inputted to the classification algorithm. This is done by the reduction of the dimensions in the feature space. Dimensional reduction technique helps in the reduction of problems of overfitting. The dimensional reduction technique used is principle component analysis. In principle component analysis it creates principle components which are linear combination of the original variables. All the 13 features are given as input to the principle component analysis. A scree plot (Fig 2) is generated with the x label as the number of components and the y axis with variance explained. This chart helps us to understand the number of principle components to be retained. Scree plot explains the cumulative variance for each of the principle components. Here we need to look for an elbow. It can be observed from the graph that the maximum number of components for which 100% of the variance of the data is explained is about 12 components. In this the number of components selected is 10 which explains about greater then

Fig. 2.



90% of the variance. The selected 10 components are then fitted to the neural network for training the data. Then the input dimensions of the layers should also be changed into 10 to be compatible.

G. Feature Selection Method

In most of the research papers it is seen that a common features selection method is applied to all the different classification methods. In, this research a separate feature selection method was used for the classification algorithms Logistic Regression, Support Vector Machine, Naïve Bayes, Decision Tree and Radom forest. The feature selection applied here is Recursive Feature elimination method. It is a type of backward feature elimination method. This is how it works, recursive feature elimination starts with the features included in the model, then it drops the least variables useful. The least variables are dropped based on the one with the minimum accuracy. This step of dropping down of the variables is carried on based on the predefined number of best features needed to be produced. The two parameters to be set in the recursive feature elimination are classifier and the number of best features that needs to be produced. There are five classification algorithms which were discussed above will be entered one by one as the classifier. The number of best features can take values from 1 to 13. If the value 1 is given with logistic regression as the classifier then it gives the best one feature for the logistic regression out of the 13 features. Similarly, if a value of 2 is given for logistic regression classifier it produces the best 2 features of the 13 features. For, number of features as 13 with logistic regression it gives all the 13 features of all the input 13 features. This indicates that there are total of 13 iterations for each classifier. The output for Logistic Regression is shown in TABLE 2.

TABLE 2

S.No	Best features	Accuracy	Precision	f1	Sensitivity	Specificity
1	12	76.42	76.22	73.07	79.37	71.92
2	11,12	79.13	77.77	78.27	78.12	80.10
3	12,11,8	82.49	84.43	80.65	86.25	78.02
4	2, 12, 11, 8	84.49	88.22	82.26	90	77.96
5	12,11,10,8,2	83.80	86.21	81.67	88.12	78.57
6	12,11,10,8,1,2	84.49	86.56	83.37	88.75	79.34
7	12,11,10,8,7,1,2	83.83	84.37	81.95	86.87	80.21
8	12,11,10,8,7,3,2,1	84.51	85.39	82.48	88.12	80.21
9	12,11,10,8,7,3,2, 1	83.83	84.60	81.80	87.5	79.5
10	12,11,10,9,8,7,5,3,2,1	85.51	86.33	83.67	88.75	81.64
11	12,11,10,9,8,7,6,5,3,2,1	84.18	85.01	83.47	87.5	80.21
12	12,11,10,9,8,7,6,5,4,3,2,1	83.16	84.28	80.95	86.87	78.68
13	12,11,10,9,8,7,6,5,4,3,2,1,0	82.82	83.64	80.76	86.25	78.68

H. HYPERPARAMETER TUNING

This is a method for tuning the parameters of the classification algorithms Logistic Regression, Support Vector Machine, Decision Tree and Radom forest. This results in the production of the highest possible accuracy, precision, f1, sensitivity and specificity. The hyperparameter tuning method used for this research is GridsearchCV [41]. Since in python there is no need for parameter tuning for the Naïve Bayes so only the four above mentioned methods are applied with tuning. The parameters used for GridsearchCV are param_grid and cv. The param_grid is used to define the list of dictionaries this allows the searching of the best parameters from the defined parameter names. The number of cross-validation is usually defined with the help of cv. The value of cv is 10 for all the four methods. The following are the list of param_grid for each of the classifier over which the search take place to produce the best parameters. Then the GridsearchCV model for each classifier is fitted to the input data. After the searching of the best parameters for all the different combination of parameter defined for 10 cross-validation, the best parameters for the model is generated.

a) Logistic Regression

Parameters defined are C = np.logspace(-3, 3, 7), penalty = [l1, l2] where l1 is lasso and l2 is ridge. The solver is liblinear.

b) Support Vector Machine

Parameters defined are C = (0.1, 1, 10, 100, 1000), gamma= (1, 0.1, 0.01, 0.001, 0.0001), kernel is linear.

c) Decision Tree

Parameters defined are max_depth = [3, None], min_sample_leaf = [1, 2, 3, 4, 5, 6, 7, 8, 9], criterion = [gini, entropy]

d) Random Forest

Parameters are same as used in case of Decision Tree, only an additional parameter n_estimators = 100 is added.

I. Ensemble

The ensemble methods are used to improve the performance of weak classification algorithms with combining with other strong classifiers.

a) Bagging

The data are randomly selected with replacement. Then the classifier is trained for each sample from bootstrap samples. The output is based on majority with same results. In this method each model is built independently in parallel way.

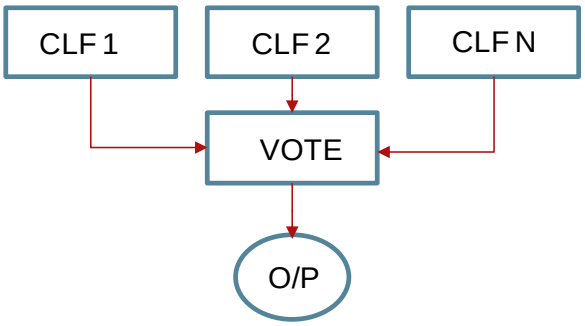
b) Boosting

The data is first divided into subsets then classifier is trained based on the subsets. Then new subsets are created by correcting the previous models. The new learner is built based on sequential pattern.

c) Vote

This technique is used to combine any number of classifiers and the output is based on majority voting. The inputs of each of the classifier are based on the selected best features. The

Fig. 3. Vote

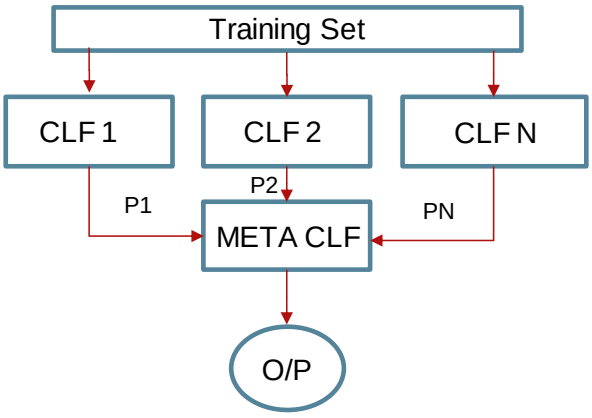


classifiers combined are 1) Random Forest and Support Vector Machine 2 Logistic Regression, Random Forest and Support Vector Machine 3) classifiers Logistic Regression and Support Vector Machine.

d) Stacking

In this multiple base classifier are combined with the metaclassifier. The base classifiers receive inputs from the original data set and the metaclassifier receives output from the base classifiers to predict the outcome. Type 1, Logistic Regression, Random Forest and Support Vector Machine are taken as base classifier and Logistic Regression as meta-

Fig. 4. Stacking



classifier. Type 2, Random Forest and Support Vector Machine as base classifier and Support Vector Machine as meta-classifier. Type 3, Logistic Regression and Support Vector Machine as base classifier and Support Vector Machine as meta-classifier.

J. RESULTS OF CLEVELAND DATA SET

The results are observed with the performance metrics namely Accuracy, Sensitivity, f1, Sensitivity and Specificity.

a) Results with feature selection for 5 classifiers

The best parameters selected are as follows, Logistic Regression parameters are C =0.1, penalty = l2 and solver = liblinear. Support Vector Machine parameters are C =0.1, gamma =1 and kernel = linear. Decision Tree parameters are criteria = gini, max_ depth = 3 and min_sample_leaf = 2. Random Forest parameters are criterion = entropy, max_depth = 3, min_sample_leaf = 4 and n_estimators = 100. The features Age, sex, cp, trestbp, chol, fbs, restecg, thalach, exang, slope, ca and thal are represented as 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, and 12. The transformed features Age, trestbp, chol, thalach, oldpeak, for Naïve Bayes are represented as 15, 16, 17, 18, and 19.

TABLE 3

Technique	Highest Accuracy	Feature
Logistic	85.51	12,11,10,9,8,7,5,3,2,1
SVM	84.82	1,12,11,2,10,9,8,7
Decision Tree	83.42	0,2,4,11,12
Random Forest	84.81	6,1,8,10,0,2,3,4,7,9,11,12
Naïve Bayes	81.49	1,2,5,6,8,10,11,12,15,16,17,18,19

The best accuracy score among 13 iterations for each model are represented in TABLE 3. It can be inferred that Logistic Regression and Support Vector Machine are best classifiers with accuracy of 85.51% and 84.82% with number of features of 10 and 8.

TABLE 4

Technique	Highest Precision	Feature
Logistic	86.33	12,11,10,9,8,7,5,3,2,1
SVM	87.97	1,12,11,2,10,9,8,7
Decision Tree	86.41	0,2,4,11,12
Random Forest	89.05	6,1,8,10,0,2,3,4,7,9,11,12
Naïve Bayes	82.32	1,2,5,6,8,10,11,12,15,16,17,18,19

The best precision score among 13 iterations for each model are represented in TABLE 4. In this case, Random Forest and

Support Vector Machine are having highest precision of 89.05% and 87.97% with the number of best features are 12 and 8.

TABLE 5

Technique	Highest f1	Feature
Logistic	83.67	12,11,10,9,8,7,5,3,2,1
SVM	82.30	1,12,11,2,10,9,8,7
Decision Tree	80.85	0,2,4,11,12
Random Forest	82.28	6,1,8,10,0,2,3,4,7,9,11,12
Naïve Bayes	79.12	1,2,5,6,8,10,11,12,15,16,17,18, 19

The best f1 score among 13 iterations for each model are represented in TABLE 5. It can be inferred that Logistic Regression and Support Vector Machine are best classifiers with F1 of 83.67% and 82.30% with number of features of 10 and 8.

TABLE 6

Technique	Highest Sensitivity	Feature
Logistic	88.75	12,11,10,9,8,7,5,3,2,1
SVM	90.62	1,12,11,2,10,9,8,7
Decision Tree	89.37	0,2,4,11,12
Random Forest	90.62	6,1,8,10,0,2,3,4,7,9,11,12
Naïve Bayes	85.62	1,2,5,6,8,10,11,12,15,16,17,18,19

The best sensitivity score among 13 iterations for each model are represented in TABLE 6. In this case, both Support Vector Machine and Random Forest are having same highest sensitivity of 90.62% with the number of best features are 8 and 12.

TABLE 7

Technique	Highest Specificity	Feature
Logistic	81.64	12,11,10,9,8,7,5,3,2,1
SVM	77.96	1,12,11,2,10,9,8,7,5
Decision Tree	77.03	0,2,4,11,12
Random Forest	64.61	6,1,8,10,0,2,3,4,7,9,11,12
Naïve Bayes	79.64	1,2,5,6,8,10,11,12,15,16,17,18,19

The best specificity score among 13 iterations for each model are represented in TABLE 7. It can be inferred that Logistic Regression and Support Vector Machine are best classifiers with specificity of 81.64% and 77.96% with number of features of 10 and 8.

TABLE 8

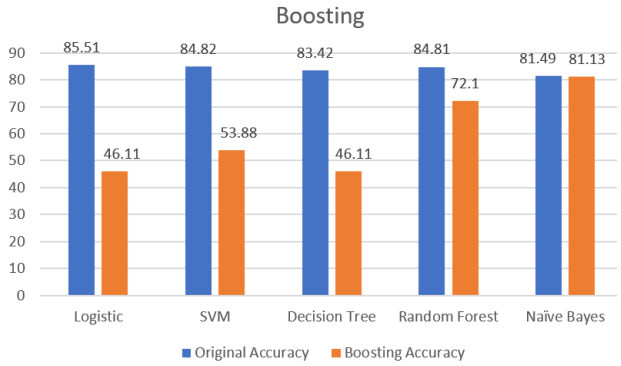
Epoch	Accuracy	Precision	f1	Sensitivity	Specificity
50	82.43	83.97	77.31	87.62	74.38
25	83.79	85.31	81.31	88.12	78.55
10	83.48	84.70	81.36	86.25	79.33

b) Results with Neural Networks

The TABLE 8 shows the results obtained after applying principle components to reduce the dimensions. Three different epochs with values 50, 25 and 10 are carried out and the corresponding performance metrics accuracy, precision and f1, sensitivity and specificity are calculated. It can be observed that performance metrics are the best for epoch 25. The lowest performance is with that epoch 50

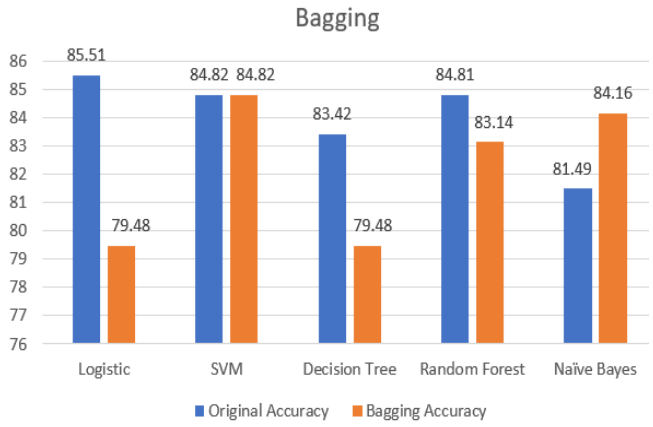
c) Ensemble Results

Fig.5



The results of boosting show that boosting tends to decrease the accuracy of all the classifier as shown in Fig 5. The decrease in percentage of accuracy compared to the original accuracy for Logistic Regression, Support Vector Machine, Decision Tree, Random Forest and Naïve Bayes are 46.07%, 36.47%, 44.72%, 14.98% and 0.4%. But the results of bagging as shown in Fig 6 were something interesting. We can see that the decrease in the accuracy in case of Logistic Regression, Decision Tree and Random Forest by 7.05%, 4.7% and 1.9%. The Support Vector Machine did not experience any change in the accuracy. Naïve Bayes experienced a positive increase of 3.2%.

Fig. 6.



The results of TABLE 9, are of the three different type of voting applied. V1 represents combination of the classifiers Random Forest and Support Vector Machine. V2 represents

TABLE 9

Vote	Accuracy	Precision	f1	Sensitivity	Specificity
V1	84.81	90.37	81.74	93.12	75.05
V2	84.50	86.65	82.14	89.37	78.73
V3	84.81	88.73	82.80	91.25	77.19

the combination of the classifiers Logistic Regression, Random Forest and Support Vector Machine. V3 represents the combination of the classifiers Logistic Regression and Support Vector Machine. It can be inferred that V2 has the lowest overall performance compared to V1 and V2. Both V1 and V2 have similar performance. But for the medical data it is better to have higher sensitivity than specificity. So, in this sense V1 is the better choice of among other two.

TABLE 10

Stack	Accuracy	Precision	f1	Sensitivity	Specificity
S1	85.50	82.96	89.52	91.87	78.02
S2	85.15	82.53	89.46	91.87	77.25
S3	85.17	82.70	88.71	91.25	77.25

The results of TABLE 10, are of the different types of stacking applied with the best selected features for each classifier. S1 represents the combination of the base classifiers Logistic Regression, Random Forest and Support Vector Machine with Logistic Regression as the meta-classifier. S2 represents the combination of the base classifiers Random Forest and Support Vector Machine with Support Vector Machine as the meta-classifier. S3 represents the combination of the base classifiers Logistic Regression and Support Vector Machine with Support Vector Machine as the meta-classifier. It can be inferred that overall performance of S2 and S3 are lower than S1. Also, S1 is the second best next to the Logistic Regression.

d) Best two models

The Table 11 shows the results obtained for the Cleveland

TABLE 11

Type	Accuracy	Precision	f1	Sensitivity	Specificity
Logistic	85.51	86.33	83.67	88.75	81.64
S1	85.50	82.96	89.52	91.87	78.02

data set is compared for classifiers with feature selection, Neural Networks with dimensional reduction, ensemble methods. The top two models obtained were the Logistic Regression and S1 (base classifiers Logistic Regression, Random Forest and Support Vector Machine with Logistic Regression as the meta-classifier).

e) Logistic Regression with multiclass dependent

The Table 12 shows the results of Logistic Regression with 5

TABLE 12

Type	Accuracy	Precision	f1	Sensitivity	Specificity
Logistic 5 class	57.42	57.42	57.42	96.20	26.16
Logistic 2 class	85.51	86.33	83.67	88.75	81.64

classes (0, 1, 2, 3, and 4) as the dependent variable. It could be observed that though the accuracy, precision, f1 and sensitivity are lower compared to Logistic Regression with 2 classes, it is interesting to see that the sensitivity (96.20%) is very high.

VI. MODEL TESTING PHASE

A. Data Set

The second data set is the Statlog Data set taken from the UCI Machine Learning Repository [43]. It also has the same attributes as that of the Cleveland data set but with different data. The difference is that the attribute Num of value 1

indicates absence and 2 indicates presence of the heart disease. There are 270 instances in this data set.

B. Pre-processing

It is observed that there are no missing values in the data. The dependent variable is converted from 2 to 1 for presence of heart disease and 1 to 0 for absence of heart disease. The data is also standardized.

C. Statlog data Results

TABLE 13

Type	Accuracy	Precision	f1	Sensitivity	Specificity
Logistic	83.63	84.45	80.56	83	84.33
S1	84.37	81.14	86.77	89.33	79.42

The following are the results of the best two models selected from the Cleveland data set. TABLE 13 are the results

TABLE 14

Type	Accuracy	Precision	f1	Sensitivity	Specificity
Logistic	83.60	83.2	80.49	83.83	84.33
S1	**85.48**	**82.48**	**87.57**	**90**	**81.08**

TABLE 15

Source	Technique	Accuracy
Proposed	Stacking with Logistic Regression	85.48 %
Mohammad Shafenoor Amin (2018)	Vote- Naïve Bayes and logistic regression	87.41%
Paul (2016)	Neural Network with Fuzzy	80 %
Verma (2016)	Decision Tree	80.68 %
Ismaeel (2015)	Extreme Learning Machine	86.50 %
El-Bialy (2015)	Decision Tree	78.54 %
Subanya and Rajalaxmi (2014)	SVM	86.76 %
Nahar (2013)	Naïve Bayes	69.11 %
Khemphila and Boonjing (2011)	Neural Network with Genetic Algorithm	80.99 %
Shouman (2011)	Decision Tree with Gain Ratio	84.10 %

obtained without applying feature selection that is all the features are included in the analysis. TABLE 14 are the results of the best selected features. The overall performance is a bit lower for Logistic Regression with feature selection compared to without feature selection. The overall performance of S1 (combination of the base classifiers Logistic Regression, Random Forest and Support Vector Machine with Logistic Regression as the meta-classifier) with feature selection is higher than without feature selection. From TABLE 14, it is observed that the best model is S1 with feature selection. TABLE 15 shows some of the accuracy of different researches works in the past. The proposed accuracy is better than that of the 6 existing models. There are three models which is showing better accuracy with different techniques applied.

VII. FUTURE WORK

The current method used recursive feature elimination for feature selection with classifiers Logistic Regression, Support Vector Machine, Naïve Bayes, Decision Tree and Random Forest. Different feature selection methods can be applied with different classification algorithm. The selected model can be applied to real time data which is larger compared to the existing data to get better results.

VIII. REFERENCE

1. Cdc.gov. (2019). *Heart Disease Facts & Statistics | cdc.gov*. [online] Available at: https://www.cdc.gov/heartdisease/facts.htm [Accessed 10 Dec. 2019].

2. Cdc.gov. (2019). [online] Available at: https://www.cdc.gov/heartdisease/docs/ConsumerEd_HeartDisease.pdf [Accessed 10 Dec. 2019].

3. Springpeople.com. (2019). *The Utilization of Data Mining in Various Industries | SpringPeople Blog*. [online] Available at: https://www.springpeople.com/blog/the-utilization-of-data-mining-in-various-industries/ [Accessed 10 Dec. 2019].

4. H. Koh and G. Tan, "Data Mining Applications in Healthcare", *Journal of Healthcare Information Management*, vol. 19, no. 2, 2019. [Accessed 10 December 2019].

5. 10. Trybula, W.J. (1997). Data mining and knowledge discovery. Annual Review of Information Science and Technology, 32, 197-229.

6. AH Chen, SY Huang, PS Hong, CH Cheng, EJ Lin, "HDPS: Heart Disease Prediction System", IEEE-Computing in Cardiology, September 2011.

7. Vikas Chaurasia and Saurabh Pal, "Data Mining Approach to Detect Heart Dieses", International Journal of Advanced Computer Science and Information Technology Vol. 2, No. 4, 2013.

8. Song.M.H., LeeJ, Cho.S.P., Lee.K.1 and Yoo.S.K (200S), 'Support vector machine Based Arrhythmia classification using reduced Features', International Journal of control, Automation, and Systems, Vol 3, pp.S71-S79

9. Dilip Roy Chowdhury, Mridula Chatterjee R. K. Samanta, An Artificial Neural Network Model for Neonatal Disease Diagnosis, International Journal of Artificial Intelligence and Expert Systems (IJAE), Volume (2): Issue (3), 2011.

10. Milan Kumari, Sunila Godara, Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction, IJCST Vol. (2), Issue (2), June 2011.

11. C. Latha and S. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques", *Informatics in Medicine Unlocked*, vol. 16, p. 100203, 2019. Available: 10.1016/j.imu.2019.100203.

12. R. Das, I. Turkoglu and A. Sengur, "Effective diagnosis of heart disease through neural networks ensembles", *Expert Systems with Applications*, vol. 36, no. 4, pp. 7675-7680, 2009. Available: 10.1016/j.eswa.2008.09.013.

13. D. S. Medhekar, M. P. Bote and S. D. Deshmukh, "Heart Disease Prediction System using Naive Bayes", *INTERNATIONAL JOURNAL OF*

ENHANCED RESEARCH IN SCIENCE TECHNOLOGY & ENGINEERING, vol. 3, no. 3, 2013. [Accessed 10 December 2019].
14. H. David and S. Belcy, "HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES", *ICTACT JOURNAL ON SOFT COMPUTING*, vol. 09, no. 01, 2018.
15. S. Bashir, U. Qamar and F. Khan, "BagMOOV: A novel ensemble for heart disease prediction bootstrap aggregation with multi-objective optimized voting", *Australasian Physical & Engineering Sciences in Medicine*, vol. 38, no. 2, pp. 305-323, 2015. Available: 10.1007/s13246-015-0337-6.
16. M. Shouman, T. Turner and R. Stocker, "Using decision tree for diagnosing heart disease patients", *Proceedings of the 9-th Australasian Data Mining Confere (AusDM'11), Ballarat, Australia*, vol. 121, 2011. [Accessed 10 December 2019].
17. M. Amin, Y. Chiam and K. Varathan, "Identification of significant features and data mining techniques in predicting heart disease", *Telematics and Informatics*, vol. 36, pp. 82-93, 2019. Available: 10.1016/j.tele.2018.11.007.
18. D. Tomar and S. Agarwal, "Feature Selection based Least Square Twin Support Vector Machine for Diagnosis of Heart Disease", *International Journal of Bio-Science and Bio-Technology*, vol. 6, no. 2, pp. 69-82, 2014. Available: 10.14257/ijbsbt.2014.6.2.07.
19. R. El-Bialy, M. Salamay, O. Karam and M. Khalifa, "Feature Analysis of Coronary Artery Heart Disease Data Sets", *Procedia Computer Science*, vol. 65, pp. 459-468, 2015. Available: 10.1016/j.procs.2015.09.132.
20. K. Revathi, "COMPARISON OF CLASSIFICATION TECHNIQUES ON HEART DISEASE DATA SET", *International Journal of Advanced Research in Computer Science*, vol. 8, no. 9, pp. 276-280, 2017. Available: 10.26483/ijarcs.v8i9.4870.
21. F. Pedregosa, G. Varoquaux, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, Oct. 2011.
22. W.Mckinney, "pandas: a Foundational Python Library for Data Analysis and Statistics", Python High Performance Science Computer, 2011.
23. I. Stancin and A. Jovic, "An overview and comparison of free Python libraries for data mining and big data analysis," *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2019.
24. M. Abadi et al., "TensorFlow: A system for large-scale machine learning", Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (0SDI'16), 2016.
25. Keras-Team, "keras-team/keras," *GitHub*, 06-Nov-2019. [Online]. Available: https://github.com/fchollet/keras. [Accessed: 10-Dec-2019].
26. "Why use Keras?," *Why use Keras - Keras Documentation*. [Online]. Available: https://keras.io/why-use-keras/. [Accessed: 10-Dec-2019].
27. H.Y. Chen, T. H. (2010). Applying data mining to explore the risk factors of parenting stress. *Expert Systems with Application*, 598-601.
28. G. Kesavaraj and S. Sukumaran, "A study on classification techniques in data mining," *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, 2013.
29. Gavrilov, "prognoz. business analytics... made simple," *Prognoz blog Benefits of decision trees in solving predictive analytics problems Comments*.[Online].Available:http://www.prognoz.com/blog/platform/benefits-of-decision-trees-in-solving-predictive-analytics-problems/. [Accessed: 10-Dec-2019].
30. C.-Y. J. Peng, T.-S. H. So, F. K. Stage, and E. P. S. John, "THE USE AND INTERPRETATION OF LOGISTIC REGRESSION IN HIGHER EDUCATION," *Research in Higher Education*, vol. 43, no. 3, 2002.
31. Adams, J. L., and Becker, W. E. (1990). Course withdrawals: A probit model and policy recommendations. Research in Higher Education 31(6): 519–538.
32. "Logistic regression," *Wikipedia*, 05-Dec-2019. [Online]. Available: https://en.wikipedia.org/wiki/Logistic_regression#Logistic_regression_vs._other_approaches. [Accessed: 10-Dec-2019].
33. S. H. U. J. U. N. HUANG, N. I. A. N. G. U. A. N. G. CAI, P. E. D. R. O. P. E. N. Z. U. T. I. PACHECO, S. H. A. V. I. R. A. NARANDES, Y. A. N. G. WANG, and W. A. Y. N. E. XU, "Applications of Support Vector Machine (SVM) Learning in Cancer Genomics," *Cancer Genomics & Proteomics*, vol. 15, no. 1, Feb. 2018.
34. HackerEarth, "Naive Bayes Theorem | Introduction to Naive Bayes Theorem | Machine Learning Classification," *YouTube*, 19-Apr-2017. [Online]. Available: https://www.youtube.com/watch?v=sjUDlJfdnKM. [Accessed: 10-Dec-2019].
35. A.A. Mohammed, R. M.-A. (2011). Evaluation of face recognition technique using PCA, wavelets and NAIVE BAYES CLASSIFIER Pattern Recognition. *Elsevier*,Volume 44, Issues 10–11.
36. Devesh Kumar, R. S. (2015). An Adaptive Method of PCA for Minimization of Classification Error Using Naïve Bayes Classifier. *Elsevier*,40 (9-15).
37. "sklearn.naive_bayes.MultinomialNB¶," *scikit*. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html. [Accessed: 10-Dec-2019].
38. Denil, M., Matheson, D., and De Freitas, N. (2014), "Narrowing the Gap: Random Forests In Theory and In Practice," in *Proceedings of The 31st International Conference on Machine Learning*, pp. 665–673. [Google Scholar]
39. L. Breiman, *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
40. R. Khandelwal, "Building Neural Network using Keras for Classification," *Medium*, 10-Jan-2019. [Online]. Available: https://medium.com/datadriveninvestor/building-neural-network-using-keras-for-classification-3a3656c726c1. [Accessed: 11-Dec-2019].
41. "sklearn.model_selection.GridSearchCV¶," *scikit*. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html. [Accessed: 11-Dec-2019].
42. *UCI Machine Learning Repository: Heart Disease Data Set*. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/heart Disease. [Accessed: 10-Dec-2019].
43. *UCI Machine Learning Repository: Statlog (Heart) Data Set*. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/statlog (heart). [Accessed: 10-Dec-2019].
44. A. K. Paul, P. C. Shill, M. R. I. Rabin, and M. A. H. Akhand, "Genetic algorithm based fuzzy decision support system for the diagnosis of heart disease," *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)*, 2016.
45. L. Verma, S. Srivastava, and P. C. Negi, "A Hybrid Data Mining Model to Predict Coronary Artery Disease Cases Using Non-Invasive Clinical Data," *Journal of Medical Systems*, vol. 40, no. 7, Nov. 2016.
46. S. Ismaeel, A. Miri, A. Sadeghian, and D. Chourishi, "An Extreme Learning Machine (ELM) Predictor for Electric Arc Furnaces v-i Characteristics," *2015 IEEE 2nd International Conference on Cyber Security and Cloud Computing*, 2015.
47. R. El-Bialy, M. A. Salamay, O. H. Karam, and M. E. Khalifa, "Feature Analysis of Coronary Artery Heart Disease Data Sets," *Procedia Computer Science*, vol. 65, pp. 459–468, 2015.
48. B. Subanya and R. R. Rajalaxmi, "Feature selection using Artificial Bee Colony for cardiovascular disease classification," *2014 International Conference on Electronics and Communication Systems (ICECS)*, 2014.
49. 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, pp. 96–104, 2013.
50. A. Khemphila and V. Boonjing, "Heart Disease Classification Using Neural Network and Feature Selection," *2011 21st International Conference on Systems Engineering*, 2011.