**ABSTRACT**

The use of multiple machine learning algorithms for arrhythmia analysis is explored. We present deferent models built by multi-class supported vector machines (SVM), multi-class Nave Bayes (NB), decision tree and random forest. The performance of the various models in predicting the presence of cardiac arrhythmia and further classifying the instances into 16 pre-denned groups is tested and presented. The random forest classier outperforms other algorithms with a test accuracy of 76%. We provide a discussion on the results of deferent models, together with some insight about of data set.

The health care industry produces a huge amount of data. This data is not always made use to the full extent and is often underutilized. Using this huge amount of data, a disease can be detected, predicted or even cured. A huge threat to human kind is caused by diseases like heart disease, cancer, tumor and Alzheimer's disease. In this paper, we try to concentrate on heart disease prediction. Using machine learning techniques, the heart disease can be predicted. The medical data such as Blood pressure, hypertension, diabetes, cigarette smoked per day and so on is taken as input and then these features are modelled for prediction. This model can then be used to predict future medical data. The algorithms like K-nearest neighbor, Naïve Bayes, support vector machine and decision tree are used. The accuracy of the model using each of the algorithms is calculated. Then the one with a good accuracy is taken as the model for predicting the heart disease.

1. **LITERATURE SURVEY**

According to Ordonez the heart disease can be predicted with some basic attributes taken from the patient and in their work have introduced a system that includes the characteristics of an individual human being based on totally 13 basic attributes like sex, blood pressure, cholesterol and others to predict the likelihood of a patient getting affected by heart disease. They have added two more attributes i.e. fat and smoking behavior and extended the research dataset. The data mining classification algorithms such as Decision Tree, Naive Bayes, and Neural Network are utilized to make predictions and the results are analyzed on Heart disease database.

Yilmaz, have proposed a method that uses least squares support vector machine (LS-SVM) utilizing a binary decision tree for classification of cardiotocogram to find out the patient condition.

Duff, et al. have done a research work involving five hundred and thirty-three patients who had suffered from cardiac arrest and they were integrated in the analysis of heart disease probabilities. They performed classical statistical analysis and data mining analysis using mostly Bayesian networks.

Frawley, et al. have performed a work on prediction of survival of Coronary heart disease (CHD) which is a challenging research problem for medical society. They also used 10-fold cross-validation methods to determine the impartial estimate of the three prediction models for performance comparison purposes.

Lee, et al. proposed a novel methodology to expand and study the multi-parametric feature along with linear and nonlinear features of Heart Rate Variability diagnosing cardiovascular disease. They have carried out various experiments on linear and non-linear features to estimate several classifiers, e.g., Bayesian classifiers, CMAR, C4.5 and SVM. Based on their experiments, SVM outperformed the other classifiers.

Noh, et al. suggested a classification method which is an associative classifier that is constructed based on the efficient FP-growth method. Because the volume of patterns can be diverse and huge, they offered a rule to measure the cohesion and in turn allow a tough choice of pruning patterns in the pattern-generating process.

Parthiban, et al. have proposed a new work in which the heart disease is identified and predicted using the proposed Coactive Neuro-Fuzzy Inference System (CANFIS). Their model works based on the collective nature of neural network adaptive capabilities and based on the genetic algorithm along with fuzzy logic in order to diagnose the occurrence of the disease. The performance of the proposed CANFIS model was evaluated in terms of training performances and classification accuracies. Finally, their results show that the proposed CANFIS model has great prospective in predicting the heart disease.

Singh, et al. have done a work using, one partition clustering algorithm (K-Means) and one hierarchical clustering algorithm (agglomerative). K-means algorithm has higher effectiveness and scalability and converges fast when production with large data sets. Hierarchical clustering constructs a hierarchy of clusters by either frequently merging two smaller clusters into a larger one or splitting a larger cluster into smaller ones. Using WEKA data mining tool, they have calculated the performance of k-means and hierarchical clustering algorithm on the basis of accuracy and running time.

Guru, et all have proposed the computational model based on a multilayer perceptron with three layers is employed to enlarge a decision support system for the finding of five major heart diseases. The proposed decision support system is trained using a back propagation algorithm amplified with the momentum term, the adaptive learning rate and the forgetting mechanics.

Palaniappan, et al. [10] have carried out a research work and have built a model known as Intelligent Heart Disease Prediction System (IHDPS) by using several data mining techniques such as Decision Trees, Naïve Bayes and Neural Network.

Shanta Kumar, et al. [11] have done a research work in which the intelligent and effective heart attack prediction system is developed using Multi-Layer Perceptron with Back-Propagation. Accordingly, the frequency patterns of the heart disease are mined with the MAFIA algorithm based on the data extracted.

Yanwei, et.al [12] have built a classification method based on the origin of multi parametric features by assessing HRV (Heart Rate Variability) from ECG and the data is pre-processed and heart disease prediction model is built that classifies the heart disease of a patient

1. **INTRODUCTION**

In day to day life many factors that affect a human heart. Many problems are occurring at a rapid pace and new heart diseases are rapidly being identified. In today’s world of stress Heart, being an essential organ in a human body which pumps blood through the body for the blood circulation is essential and its health is to be conserved for a healthy living. The health of a human heart is based on the experiences in a person’s life and is completely dependent on professional and personal behaviors of a person. There may also be several genetic factors through which a type of heart disease is passed down from generations. According to the World Health Organization, every year more than 12 million deaths are occurring worldwide due to the various types of heart diseases which is also known by the term cardiovascular disease. The term Heart disease includes many diseases that are diverse and specifically affect the heart and the arteries of a human being. Even young aged people around their 20-30 years of lifespan are getting affected by heart diseases. The increase in the possibility of heart disease among young may be due to the bad eating habits, lack of sleep, restless nature, depression and numerous other factors such as obesity, poor diet, family history, high blood pressure, high blood cholesterol, idle behavior, family history, smoking and hypertension. The diagnosis of the heart diseases is a very important and is itself the most complicated task in the medical field. All the mentioned factors are taken into consideration when analyzing and understanding the patients by the doctor through manual check-ups at regular intervals of time.

The symptoms of heart disease greatly depend upon which of the discomfort felt by an individual. Some symptoms are not usually identified by the common people. However, common symptoms include chest pain, breathlessness, and heart palpitations. The chest pain common to many types of heart disease is known as angina, or angina pectoris, and occurs when a part of the heart does not receive enough oxygen. Angina may be triggered by stressful events or physical exertion and normally lasts under 10 minutes. Heart attacks can also occur as a result of different types of heart disease. The signs of a heart attack are similar to angina except that they can occur during rest and tend to be more severe. The symptoms of a heart attack can sometimes resemble indigestion. Heartburn and a stomach ache can occur, as well as a heavy feeling in the chest. Other symptoms of a heart attack include pain that travels through the body, for example from the chest to the arms, neck, back, abdomen, or jaw, lightheadedness and dizzy sensations, profuse sweating, nausea and vomiting.

Heart failure is also an outcome of heart disease, and breathlessness can occur when the heart becomes too weak to circulate blood. Some heart conditions occur with no symptoms at all, especially in older adults and individuals with diabetes. The term 'congenital heart disease' covers a range of conditions, but the general symptoms include sweating, high levels of fatigue, fast heartbeat and breathing, breathlessness, chest pain. However, these symptoms might not develop until a person is older than 13 years. In these type of cases, the diagnosis becomes an intricate task requiring great experience and high skill. A risk of a heart attack or the possibility of the heart disease if identified early, can help the patients take precautions and take regulatory measures. Recently, the healthcare industry has been generating huge amounts of data about patients and their disease diagnosis reports are being especially taken for the prediction of heart attacks worldwide. When the data about heart disease is huge, the machine learning techniques can be implemented for the analysis.

1. **ARCHITECTURE**

In this algorithm, the input image, *x*, which is the same as the content image, is passed through the Image Transform Net. The output of the the Image Transform Net, *y*ˆis passed throught the VGG-16 [15] net along with the style target and want to use. We can use the Image Transform Net to transfer the style of the style target used to train this network onto any content image in real time. The quality of images obtained using this algorithm is comparable to the algorithm described in sectionPlease refer to for the specific architecture of the Image Transform Net.

The main drawback of this algorithm is that each network is tied to a single style. If we want to perform style transfer for multiple styles, we have to train a separate network using each style. Not only is this extremely inefficient, it is impractical in many situations. For example, in mobile applications, there is not enough memory in mobile phones to store the parameters of many separate Image Transform Networks. In order to address this drawback, we will next look at algorithms that can perform style transfer for multiple styles using only one network in real time.

**3.1 CLASSIFICATION USING RANDOM FOREST**

Random forests (RF) [13] are combination of tree predictors using decision tree such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them. They are more robust with respect to noise. It is a supervised classification algorithm used for the prediction and it is considered as the superior due to its large number of trees in the forest giving improved accuracy than decision trees. Typically, the trees are trained independently and the predictions of the trees are combined through averaging. Random forest algorithm can use both for classification and the regression based on the problem domain. The algorithm for random forest is given below:

**Step 1:** Randomly select k features from entire m features, where *k* << *m*.

**Step 2:** Surrounded by the *k* features, calculate the node “*d*” using the best split point.

**Step 3:** Split the node into daughter nodes using the best split.

**Step 4:** Repeat 1 to 3 steps until l number of nodes has been reached.

**Step 5:** Construct forest by repeating steps 1 to 4 for n number times to create n number of trees.

Firstly, the *k* features are taken out of total *m* features. In the next stage, in each tree randomly select *k* features in order to find the root node by using the best split approach. The next stage involves calculating the daughter nodes using the same best split approach for the heart disease dataset. Similarly, the tree is formed from the root node and until all the leaf nodes are generated from the attributes. This randomly created tree forms the random forest that is used for making heart disease prediction in patients.

**3.2 CLASSIFICATION USING DECISION TREE.**

Decision Tree (DT) [14] is a simple and easy to implement classifier. The bit through feature to access in depth patients’ profiles is only obtainable in Decision Trees. Decision tree builds classification or regression models in the structure of a tree making it simple to debug and handle. Decision trees can handle both categorical and numerical data. The algorithm works by finding the information gain of the attributes and taking out the attributes for splitting the branches in threes. The information gain for the tree is identified using the below given

***E*(*S*)= -*P*(*P*)log2*P*(*P*)-*P*(*N*)log2*P*(*N*) (1)**

The algorithm for the decision tree is given below:

**Step 1:** Identify the information gain for the attributes in the dataset.

**Step 2:** Sort the information gain for the heart disease datasets in descending order.

**Step 3:** After the identification of the information gain assign the best attribute of the dataset at the root of the tree.

**Step 4:** Then calculate the information gain using the same formula.

**Step 5:** Split the nodes based on the highest information gain value.

**Step 6:** Repeat the process until each attributes are set as leaf nodes in all the branches of the tree.

**3.3 CLASSIFICATION USING NAÏVE BAYES**

Naïve Bayes (NB) is a statistical classifier which assumes no enslavement between attributes. Naive Bayes [15] is based on Bayes rule and it assumes that attributes are independent of each other. The working principle of naïve Bayes classifier is as follows:

• *Training Step*: By assuming predictors to be conditionally independent given for a class, the method estimates the parameters of a probability distribution known as the prior probability from the training data.

• *Prediction Step*: For unknown test data, the method computes the posterior probability of the dataset which is belonging to each class. The method finally classifies the test data based upon the largest posterior probability from the set

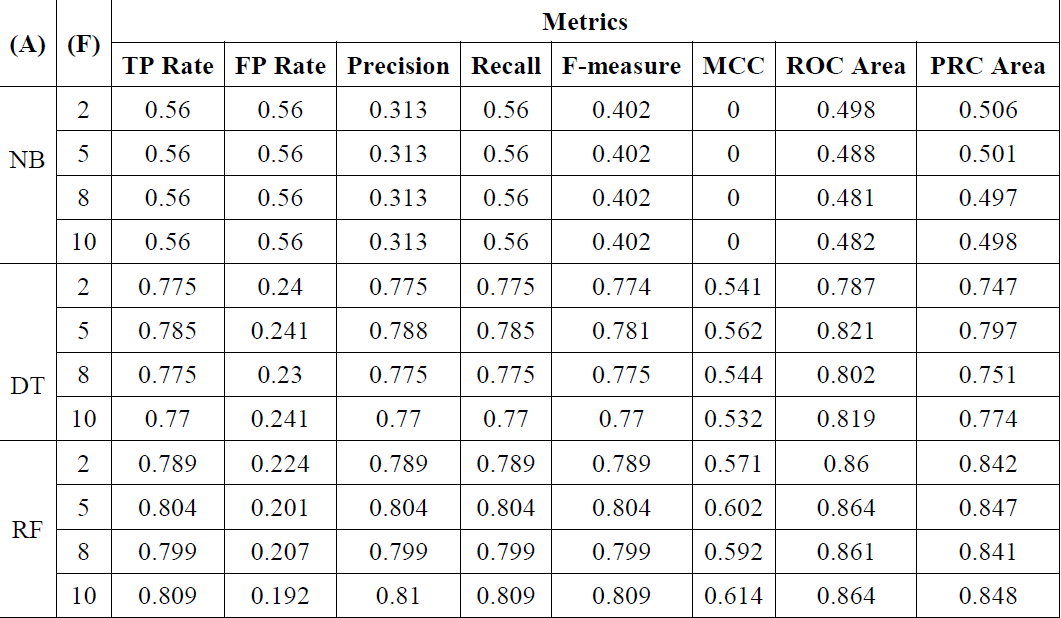
**3.4 CLASSIFICATION USING CROSS VALIDATION**

In *k*-fold cross-validation, the innovative sample is randomly partitioned into *k* subsamples. Then *k* subsamples, a single subsample is retained as the validation data designed for testing the representation, and the remaining *k*-1 subsamples are used as training data. This kind of situation is referred to as use training set generally referring to as entirely utilizing the dataset for training and testing.

For instance, in 10-fold cross validation, the original sample is randomly partitioned into 10 subsamples. From the 10 subsamples, a single subsample is retained as the validation data i.e. testing data which is used for testing the model and the remaining 9 subsamples is treated as training data that is used for training the classification algorithm. The cross validation process is then repeated in the same manner for 10 times (the folds), with each of the 10 subsamples used exactly once as the validation data. By shuffling and swapping the folds of data, the 10 results from the folds then can be averaged (or otherwise combined) to produce a single estimation. The experimental results of the classification of heart disease done using many folds of cross validation are given in Table.2. Here, (*A*) refers to the algorithms and (*F*) refers the number of folds in cross validation.

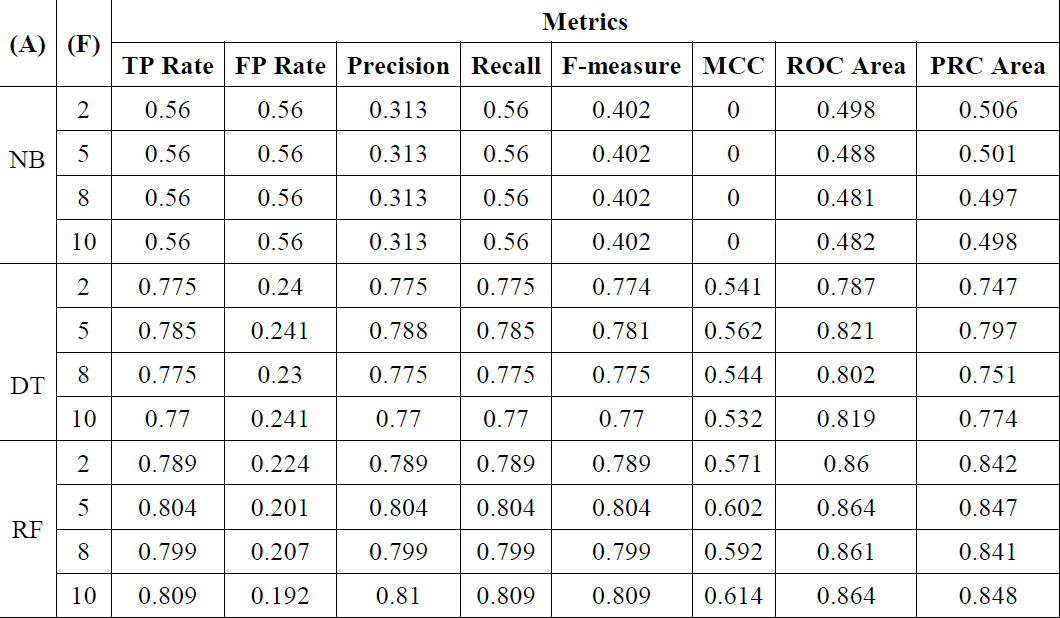
Based on the Table.2, the folds play a vital role in the improvement of the metric values for the precision, recall and f-measure. Although the performance has been improved based on implementation of several folds, the random forest algorithm outperforms the Decision Tree and Naïve Bayes algorithms.

**3.5 CLASSIFICATION USING PERCENTAGE SPLIT**

In percentage split, the data is split for both the training and testing. It actually splits the data and separates *x*% of the data for wisdom and the rest of it for testing. It is useful when your algorithm is time-consuming.

Classification of heart disease using Cross Validation

For instance, 60%-40% percentage split the classification results will be evaluated on a test set that is a part of the original data. The percentage split is 60%, which means that 60% of the data go for training and 40% for testing. The classification model is built based on this and the experiment is conducted. The experimental results of the classification of heart disease done using many splits of percentage splitting are given in Table.3. Here, (*A*) refers to the algorithms and (*PS*) refers the range of percentage to split i.e. for e.g. 60-40 refers to splitting of the dataset as 60% and 40%. The first is to be utilized for training the classifier and the latter is to be used for testing the classifier.



Classification of heart disease using Percentage Split

The results described in the above Table.3 are achieved by splitting the dataset into percentage for training and testing part that are taken as randomly stratified data points. The metric results show that the values for the precision, recall and f-measure of the Random Forest algorithm are higher. Finally, the results show that the Random Forest is best suited for the prediction of heart disease than the Decision Tree and Naïve Bayes classification algorithms

1. **CONCLUSIONS**

The overall objective of the work is to predict more exactly the occurrence of heart disease using data mining techniques. In this research work, the UCI data repository is used for performing the comparative analysis of three algorithms such as Random Forest, Decision trees and Naive Bayes. From the research work, it has been experimentally proven that Random Forest provides perfect results as compare to Decision tree and Naive Bayes.

The Future work of this research work can be made to produce an impact in the accuracy of the Decision Tree and Bayesian Classification for additional improvement after applying genetic algorithm in order to decrease the actual data for acquiring the optimal subset of attribute that is enough for heart disease prediction. The automation of heart disease prediction using actual real time data from health care organizations and agencies which can be built using big data. They can be fed as a streaming data and by using the data, investigation of the patients in real time can be prepared.

1. **REFERENCES**

[1] Carlos Ordonez, “Improving Heart Disease Prediction using Constrained Association Rules”, Technical Seminar Presentation, University of Tokyo, 2004.

[2] Franck Le Duff, CristianMunteanb, Marc Cuggiaa and Philippe Mabob, “Predicting Survival Causes After Out of Hospital Cardiac Arrest using Data Mining Method”, *Studies in Health Technology and Informatics*, Vol. 107, No. 2, pp. 1256-1259, 2004.

[3] W.J. Frawley and G. Piatetsky-Shapiro, “Knowledge Discovery in Databases: An Overview”, *AI Magazine*, Vol. 13, No. 3, pp. 57-70, 1996.

[4] Heon Gyu Lee, Ki Yong Noh and Keun Ho Ryu, “Mining Bio Signal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV”, *Proceedings of International Conference on Emerging Technologies in Knowledge Discovery and Data Mining*, pp. 56-66, 2007.

[5] Kiyong Noh, HeonGyu Lee, Ho-Sun Shon, Bum Ju Lee and Keun Ho Ryu, “Associative Classification Approach for Diagnosing Cardiovascular Disease”, *Intelligent Computing in Signal Processing and Pattern Recognition*, Vol. 345, pp. 721-727, 2006.