Prediction and Classification of Cardiac Arrhythmia

Shalini BN[1], Nandini V[2], Sandhya M[3], Bharathi R[4]

[1,2,3]*Student, BE, Computer Science and Engineering, BMS Institute of Technology, Karnataka, India*
[4]*Associate Professor, Dept. of Computer Science and Engineering, BMS Institute of Technology, Karnataka, India*

---***---

Abstract - *Classification of Arrhythmia with high accuracy is an important and challenging task. Arrhythmia which is considered as a life threatening disease must be accurately predicted and multi classified so that the life span can be increased. The dataset is accessed from the UCI database. Pre-processing and normalization steps have been done before prediction and classification of cardiac arrhythmia. Important features are selected from the Extra Trees Classifier method. The data is normalized by using a standard scalar and cleaning of data is carried out by imputing the mean values replacing the missing values. The Ensemble Classifier which is a combination of Logistic Regression, SVM, Random Forest and Gradient Boost is been implemented for prediction and classification of arrhythmia. Ensemble Classifier has outperformed in terms of accuracy performance metric when compared to other machine learning algorithms. The Ensemble Classifier achieves an accuracy of 90% of the prediction of arrhythmia.*

Key Words: **Arrhythmia, UCI, Normalization, Extra Trees Classifier, Ensemble.**

1. INTRODUCTION

Nowadays people are affected by various chronic diseases. Heart diseases are one among that which affects a large population. Stress is also a reason for many people's heart attack. This unwanted heart attack and sudden death can be prevented by early detection and timely treatment of arrhythmia which will reduce the heart attack in people and also prevent the loss of life.

ECG is the most broadly utilized diagnosing gadget or instrument for capacity of heart. Which is being recorded when cathodes set on the body that produces examples of the electrical drive of the heart. ECG signals are made out of P waves, QRS waves, T waves. The connection between these P waves, QRS waves, T waves and RR interims, time term and shape are required for looking at a heart understanding. Arrhythmia is a type of abnormalities in heart beat where heart thumps excessively quick or too moderate which results in heart sicknesses. AI systems can be connected to improve exactness of heart arrhythmia order from ECG signals. Classification of heart arrhythmia relies upon the setting of use, information investigation prerequisite of the predetermined patient for choosing a proper strategy. In this paper we have proposed a productive framework that arrange ECG signal into typical or unhealthy classes which group between the presence and nonattendance of arrhythmia. UCI AI store is the place dataset is been separated from. Multiclass characterization is connected to arrange the records into 16 unique classes where the top of the line is ordinary and rest is sort of cardiovascular arrhythmia. Wrapper strategy is the component determination system that is being completed to decrease the huge arrangement of highlight in the dataset.

At that point as the subsequent stage pre-processing is completed to the chose highlight to acquire consistency the conveyance of the information. SVM strategy, for example, one-against-one, one-against-all, blunder code and other arrangement calculation, for example, Random Forest, Logistic Regression, Gradient Boosting and Ensemble technique is being connected, prepared and tried on the standard dataset to improve the exactness and expectation of cardiovascular arrhythmia. Gathering is perceived to give a remarkable execution in order and beat other grouping calculation which is been contrasted and gives more prominent exactness in forecast and arrangement of heart arrhythmia.

The significant commitments of this paper are:

(1) Wrapper strategy is utilized for choosing most noteworthy and pertinent component.

(2) Ensemble technique which is mix of Random Forest, Logistic relapse, Gradient Boosting and SVM classifier is executed and contrasted and other grouping calculation which beats others and acquire a high multiclass order precision on UCI-arrhythmia dataset.

2. LITERATURE SURVEY

[1] This paper uses SVM and logistic regression method for classification of cardiac arrhythmia. The dataset is taken from the UCI machine repository. Two stage serial fusion classifier systems are used. The SVM's distance outputs are related to the confidence measure uses first level classifier of SVM. The rejection thresholds for positive and negative ECG samples are used. The samples which are rejected are forwarded to the next stage of logistic regression classifier. The final decision is obtained by combining the classifiers performances.

[2] This paper uses the wavelet energy histogram method and support vector machines classifier for classification of cardiac arrhythmia. The dataset is taken from the MIT-BIH.

The classification of cardiac arrhythmia includes three stages which are ECG signal pre-processing, feature extraction and heartbeats classification. The QRS complex detection is done through discrete wavelet transform using a pre-processing tool for noise removal in signals. The features which are extracted from the complex detection are input to the classifier. The classifier that is used is binary classifier. The classier is used to classify the arrhythmia dataset as normal and abnormal.

[3] This paper uses an anomaly detection scheme for arrhythmia detection. The most important pre-processing phase is well considered before the detection of anomaly. The recursive feature elimination method and random forest Ensemble methods are used for pre-processing of the arrhythmia dataset. Once the pre-processing task is finished the next phase is anomaly detection. Clustering based oversampling PCA method is used for anomaly detection. The method is compared and tested on various datasets. The k-median clustering is used to cluster various similar clusters together. The PCA method has outperformed its detection accuracy when compared to the other methods of anomaly detection.

[4] This paper analyses the ECG signal using SVM method. The ECG signals are extracted from the BIOPAC Acknowledge software. The analysis is done by using parameters like ST interval of ECG signals, ST segment elevation, PR interval, QRS complex. According to the parameters used the arrhythmia diseases are predicted.

[5] This paper uses clustering and regression approach for the prediction of cardiac arrhythmia. Multiclass logistic regression method is used for regression task and DBSCAN is used for clustering. The less instances clusters are subjected to multiclass logistic regression after clustering. The prediction of cardiac arrhythmia is done once the regression is completed. The clustering regression method achieves an accuracy of 80% for the prediction of arrhythmia.

[6] This paper multi classifies the arrhythmia using the SVM invariants approach including one-against-one, error-correction code and one-against-all methods. Using these methods the presence and absence of arrhythmias is detected. Other machine learning algorithms are also employed to for comparison of performances. The comparisons are done using accuracy, root mean square error and kappa statistics. The OAO method of SVM showed better results when compared to other algorithms with an accuracy of 81.11% when the dataset is split with 80/20 and when it is split with 90/10 the accuracy of 92.07% is achieved.

[7] This paper uses deep learning algorithm for multistage classification of arrhythmia. The dataset is taken from the MIT-BIH. According to the standards of AAMI the arrhythmia dataset is classified into five groups. The features are extracted by using various methods namely wavelet packet decomposition, higher order statistics, morphology and discrete Fourier transform techniques. The deep learning method achieved a performance of about 94.15%, 92.64% and 93.38% for accuracy, sensitivity and selectivity.

[8] This paper uses recurrent neural networks in classification of ECG Arrhythmia into normal and abnormal beats. The regular and irregular beats are separated using RNN. The dataset is taken from the MIT-BIH. The ECG data is input to Long Short Term Memory Network and then the data is split to training and testing data. Different RNN models are used for quantitative comparisons and the classification of ECG arrhythmia is done.

3. IMPLEMENTATION

3.1. Arrhythmia Classification Models

Various efficient techniques and intelligent methods have been developed for accurate prediction and classification of cardiac arrhythmia. When large number of datasets is available for prediction of cardiac arrhythmia then deep learning methods are preferred more than the machine learning methods. Different methods have also been developed for automatic prediction of cardiac arrhythmia such as SVM (support vector machine), feed forward neural network (FFN), and regression neural network (RNN). One among them is arrhythmia classification using multilayer perception and convolution neural networks. These two are deep learning methods for classification.

The advantage of machine deep neural networks (DNN) algorithms over machine learning algorithms is that it can recognize more complex features with the help of hidden layers. This feature of DNN makes them capable to handle large datasets with large number of dimensions. The dataset which is used for predicting the disease is taken from kaggle.com. The missing values in the dataset are replaced with the computed mean values. The most important feature of CNN is that it is very easy to train and it does not require the dataset to be pre-processed, as CNN itself performs this task. CNN is self-learned and self-organized which doesn't require supervision which makes it more powerful. It also contains many numbers of hidden layers called the convolution layers which identify the complex features. CNN does even require a separate feature extraction method as the algorithm extracts it on its own. These features of CNN make it more advantageous in the domain of classification problems. The entire dataset is distributed as 70% for training the model and 30% for testing how accurately the model classifies the disease. The CNN algorithm is trained and the hidden layers present in CNN called convolution layers predict the presence of the disease. In the case of MLP minimum of three nodes are used for predicting and the algorithm is trained and tested with the appropriate datasets. The results are MLP can classify the disease with an accuracy rate of 88.7% and CNN can classify the disease with an accuracy rate of 83.5%. The conclusion of this methodology is that MLP deep learning algorithm outperforms the CNN algorithm. A new technique for extracting most relevant features is correlation based feature selection features which are recently used in many of the algorithms. Along with this the incremental backpropagation neural network is used for accurate prediction and classification of cardiac arrhythmia.

Decision trees can also be used in appropriate prediction of the arrhythmia and classify into 16 different classes. An accurate model can be developed for prediction of the disease by using the most efficient machine learning classification algorithms and by using the most appropriate methods for selecting the most relevant features in predicting arrhythmia as the accuracy of prediction depends on the number of important features extracted. The most efficient feature selection technique which is available can be used only with the unsupervised data.

In all the above proposed methods even though the accuracy of prediction is quite descent it still has the room for improvement. Feature selection is the major step in improving the accuracy as it only selects the most relevant features that contributes the most in accurate prediction of cardiac arrhythmia. So an efficient technique has to be proposed for feature selection to reduce the dimensions of the dataset and an efficient classification algorithm has to be proposed for accurate prediction of cardiac arrhythmia.

3.2. Proposed Model

A model for more accurate prediction of cardiac arrhythmia is proposed in this paper. The required data for prediction is collected from the standard repository for machine learning data called UCI repository. After data collection next step is extraction of most relevant features using the wrapper feature selection algorithm which is built around the random forest algorithm. This paper proposes various machine learning classifiers such as the gradient boosting, Random forest, Logistic regression, support vector machine (SVM) invariants such as one-against-one, one-against-rest and error code and Ensemble classifier for the prediction of cardiac arrhythmia. Then these algorithms are compared on the basis of their accuracy of prediction and classification of cardiac arrhythmia. So that the most accurate machine learning classifier can be found for arrhythmia prediction.

3.2.1. Wrapper Feature Selection method

The datasets normally contain large number of dimensions which can yield a classification which is not accurate especially in the multiclassification of the features. Feature selection is mainly employed for extracting the most important features and is also useful in eliminating the redundant attributes present in the dataset. It is the first important pre-processing step in any of the machine learning algorithms and plays an important role in accurate prediction of the disease.

The dataset which is taken from UCI repository has a huge number of attributes which has to be eliminated for improve the accuracy of prediction and it filters out the attributes which are irrelevant. In this paper wrapper technique, which is used for extracting important attributes from the dataset is proposed. In this method the feature selection is built around the random forest like a wrapper. RF is basically an ensemble method which consists of many weak classifiers. The decisions of all these weak classifiers is taken into consideration for the classification through a process called majority voting technique in which the voting of all the classifiers is as decisions is used to arrive at a conclusion. A score is computed by the wrapper technique called the z-score which is used to measure the relevancy of the attributes.

3.2.2. Data Pre-processing

The dataset has features with large numeric values which may directly affect the accuracy of prediction when compared to the features with small numeric values. The steps in the above proposed model are shown below in fig 3 is as follows:

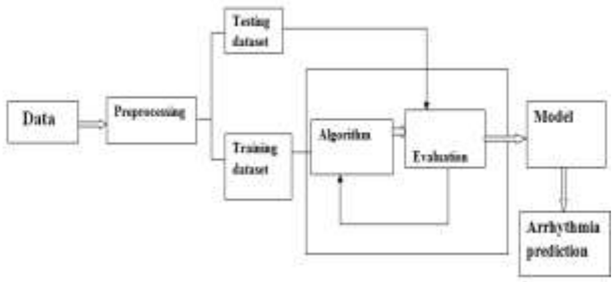

Fig -1: Classification Steps

Therefore, normalization of the dataset has to be carried out after the feature extraction. The dataset is normalized to increase the accuracy of prediction by reducing the number of irrelevant features.

Centralization and the scaling technique are used for data normalization. In centralization transformations the mean values of the attributes in reduced. Then the scaling transformations are performed. And in the pre-processing step the null and the missing values are handled. The nulls and missing values are replaced by the mean values or they can also be filtered out from the dataset.

3.2.3 Arrhythmia Classification

After selecting the relevant features and pre-processing, this normalized data is used for predicting whether the cardiac arrhythmia is present or not and classifying it. The dataset which is normalized is distributed as testing and training datasets. Then the machine learning classifiers such as Gradient boosting, Random forest, Logistic regression, support vector machine (SVM) invariants such as one-against-one, one-against-rest and error code and Ensemble classifier which is a combination of gradient boosting, random forest, logistic regression, support vector machine are trained well with the training datasets. The testing datasets is used to test how accurately the trained model predicts and classifies the disease and the parameter used for evaluating the results is accuracy. In this paper 70% of datasets is used training the model and 30% is used as the testing datasets. The trained model classifies the dataset into 16 groups, the first group represents the absence of the disease which is a normal condition and the rest of the groups represents the presence of cardiac arrhythmia which are the types of cardiac arrhythmia. So this method is called the multi-class classification problem. In this paper all the proposed classifiers are compared and the most accurate one is obtained. After the classification process, it is found that the Ensemble classifier outperforms all other classifiers in terms of accuracy with an accuracy rate of 90%.

4. RESULTS

Performance measures. Accuracy is used as a performance metric for measuring classification algorithm's performance. The proposed Ensemble method combines the algorithms and a voting classifier is used either with the soft voting or the majority rule voting. The table 1 gives a comparison of the accuracies of the various machine learning algorithms implemented. The ensemble method or the voting classifier has shown a better accuracy when compared with other classifiers.

Table -1: Classification accuracy of various algorithms Implemented.

Machine Learning Algorithms	Accuracy
Gradient Boosting	76%
Random Forest	70%
Logistic Regression	74%
SVM	67%
One against One	89%
Output Code	70%
One against Rest	68%
Voting Classifier	90%

Confusion matrix is used to represent the performance of the proposed method. It is a matrix represented form to measure the performance of an algorithm. The rows of the matrix represent the predicted values and the columns of the matrix represent the actual values. The diagonal of the matrix represents the true positives in a multi classification of arrhythmia. The accuracy is calculated by the number of correct predictions of the samples divided by the total of all the samples or observations. The below confusion matrix is plotted for the voting classifier.

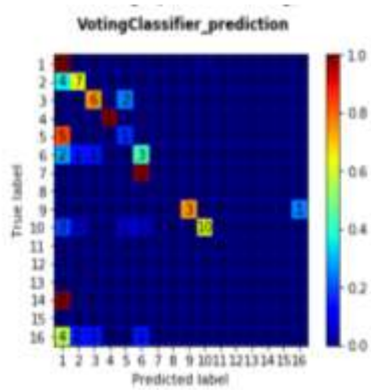

Chart -1: Confusion Matrix of Voting Classifier

The arrhythmia dataset is classified into sixteen classes where the first class is the normal class and the other fifteen classes represent the presence of arrhythmia. The below figure is a bar chart plotted where the x axis represents the different classes of arrhythmia and y axis represents the instances.

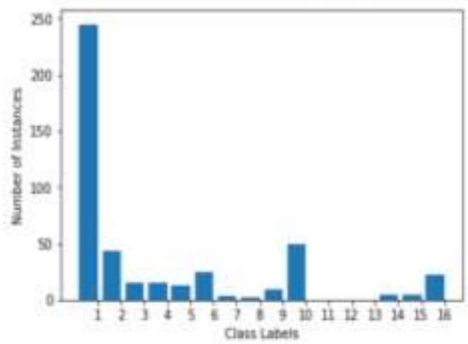

Chart -2: Multi classification of arrhythmia

The accuracy results of the algorithms which showed a better performance than others are plotted in the form of a bar graph. The bar chart shows that the ensemble method outperformed others in terms of accuracy.

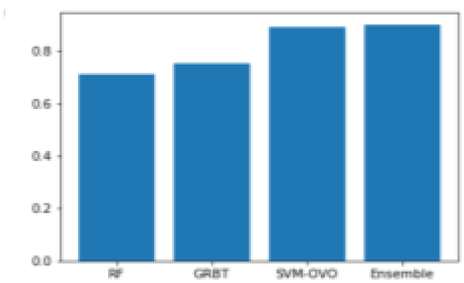

Chart -3: Comparison of accuracies in form of bar chart

5. CONCLUSION

A model is proposed in this paper which is an Ensemble Classifier that gives a better accuracy when compared to other machine learning algorithms. Pre-processing and normalization are most important steps before the prediction and classification of arrhythmia is done. Other machine learning algorithms are also implemented for comparison of performances but Ensemble Classifier has outperformed in accurately classifying arrhythmia. The model can be connected to the ECG device instead of taking the dataset from UCI database. To other biomedical and non-biomedical applications the model can be extended and used.

ACKNOWLEDGEMENT

This work is done, supervised and supported by the students and faculty members of the Department of Computer Science and Engineering, BMS Institute of Technology, Bangalore.

REFERENCES

[1] Asli Uyar and Fikret Gurgen, **Arrhythmia Classification Using Serial Fusion of Support Vector Machines and Logistic Regression**, IEEE International Workshop on Intelligent data Acquisition and Advanced Computing Systems, 2007.

[2] Abhishek S**, QRS Complex Detection and Arrhythmia Classification using SVM**, IEEE 2015 International Conference on Communication, Control and Intelligent Systems.

[3] Asha Ashok, Smitha S, Kavya Krishna MH, **Attribute Reduction based Anomaly Detection Scheme by Clustering Dependent Oversampling PCA**, IEEE 2016 International Conference on Advances in Computing, Communications and Informatics.

[4] Tahmida Tabassum and Monira Islam, **An Approach of Cardiac Disease Prediction by Analyzing ECG Signal**, IEEE 2016.

[5] Prathibhamol, Anjana Suresh and Gopika Suresh**, Prediction of Cardiac Arrhythmia type using Clustering and Regression Approach (P-CA-CRA),** IEEE 2017.

[6] Anam Mustaqeem, Syed Muhammad Anwar and Muahammad Majid**, Multiclass Classification of Cardiac Arrhythmia Using Improved Feature Selection and SVM Invariants,** Computational and Mathematical Methods in Medicine, Hindawi 2018.

[7] Gokhan Altan, Novruz Allahverdi and Yakup Kutlu, **A Multistage Deep Learning Algorithm for Detecting Arrhythmia**, IEEE 2018.

[8] Shraddha Singh, Saroj Kumar Pandey, Urja Pawar and Rekh Ram Janghel, **Classification of ECG Arrhythmia using Recurrent Neural Networks**, International Conference on Computational Intelligence and Data Science, Elsevier 2018.