DATA MINING SYSTEM USING FUZZY C-MEANS CLUSTERING FOR CLASSIFICATION OF ECG SIGNALS

Rajesh Elangovan [1], Srinivasan Alavandar [2]

[1]Assistant Professor, Department of Computer Science and Engineering,
Sri Jayaram Institute of Engineering and Technology, Affiliated to Anna University-Chennai, Tamilnadu, INDIA.
(E-mail: rajresearch16@gmail.com)
[2]Professor and Head (Academics), Department of Electrical and Electronics Engineering,
Agni College of Technology, Affiliated to Anna University-Chennai, Tamilnadu, INDIA.
(E-mail: Seenu.phd@gmail.com)

Abstract--**The Electrocardiogram signals for different diseases comprises many indiscriminate features which makes the classification, in to a big task for diagnosis and treatment. The conventional method of classification has many inconvenience to the physicians that forces the public to look after the experienced cardiologist for their heart related diagnosis and treatment. So the situation made us to develop a new classifier to reduce high mortality rate of heart diseases, early prediction and precise classification of ECG signal. In this research the two important tasks needed for development of classifier which comes under data mining are clustering and classification. The data required for classification are extracted with Wavelet Transform (WT) and pre-classification were done with Fuzzy C-Means Clustering (FCM) technique. Finally the classifiers are developed with Feed Forward Neural Network (FFNN) by using the data, extracted by FCM. The simulated waveforms are used in training and testing and to classify three different arrhythmias. The techniques adapted in the design of classifier performs relatively well in terms of classification results, when compared with other classification technique.**

Key words--**Wavelet Transform, Fuzzy C-Means Clustering, FFNN, Data mining.**

I. INTRODUCTION

The blood circulatory system of a human depends on the electrical activity of the heart. To reduce the high mortality rate and to improve any nations wealth, the performance related with human regular activity is very-very important, So early prediction of any diseases in particular heart related diseases is very much important. These kind of prediction are done by placing sensors at the limbs, to check and record the propagation of the electrical potential through cardiac muscles using ECG. These ECG will give useful information about the functioning and condition of the heart. The conditions of the heart is generally decided by the shape of ECG waveform and rate of the heart beats. In these for analysis, a typical structure of the ECG waveform are taken as a reference signal which contains important data related with the condition of the heart under various situation and circumstances.

The patients who are in the intensive care need immediate, continuous monitoring and treatment which mostly depends on the early, quick prediction and classification of ECG signals. Till now the early prediction and classification were done with computers using various technique. Analysis or designing these kind of packages or software to support computers need integration of suitable data mining and pattern classifier to operate and make an effective and efficient system. So far many algorithms have been proposed and developed in the literature for the prediction and classification of ECG signals. [1] and [2] proposed a technique to extract information related with Rpeaks and RR interval by using Discrete Wavelet Transforms (DWT) up to level 2. For this the arrhythmia were taken from MIT-BIH. The extracted data were used for training and testing the multilayer perceptron neural network, done with error back propagation algorithm. The main intention of the work is to estimate the recognition rate of the network used for the classification. In this Rpeak, RR interval and its feature alone were considered, and that too up to level 2 by DWT. It may be useful for earlier prediction or to see the changes happened in the arrhythmia but it won't give exact recognition rate as it was trained with less informed features which will not discriminate with high rate.

[3] Used feed forward ANN as a classifier to classify various condition happened in the recorded ECG signals. Fast Fourier Transform (FFT) is used to remove low frequencies which are present in the signals while extraction the information about Rpeaks and RR interval. To restore the ECG signals, to have efficient analysis, inverse FFT where used for the segmentation of ECG beats and to give inputs for feed forward ANN for training and testing. Error back propagation is used to estimate the classification accuracy as well as Youden index, which summarizing the analysis done during diagnosis test. It ranges from -1 to 1 and will be zero for same propagation of positive results for the segments with and without the disease.

Combinations of various features were used for the ECG beats classification [4-6].In addition to RR interval and Rpeaks, QRS were also taken for the analysis. The feature extraction not only done with DWT but also with ST to check the sensitivity while designing the classifier. In this DWT is used at level 4 without considering or cross-checking with various other levels. Pan Tompkin's algorithm is been used for the detection of QRS. The author concentrated on zero mean and baseline wander for the removal of noise in support with band pass filter.

The extracted features are used as inputs to MLPNN classifier to assess the accuracy level which have resulted as 69.38 and 97, which is not in acceptable level for the research deals with human health that too of analysis done for diagnosing the heart. By seeing the above issues, for accurate prediction and classification of ECG signals, FCM is been proposed in this paper. Many authors have used the FCM for grouping various heart related diseases such as [7] used FCM for the classification purpose and estimated the accuracy rate as 99.05%. Here RR interval alone is considered as a feature for diagnosis. The extracted features are processed with DWT and denoised with Z score in particular baseline adjustment. The results were almost close to 100%, the thing is author used RR interval alone which will not be enough to come conclusion regarding diseases and treatment. [8] and [9] used FCM as pre-classifier which classifies the signals by seeing the distance between the center itself. In this RR interval and Rpoint location were taken in to account. Daubechies wavelet of order 3 is been used as a mother wavelet under DWT. The features extracted by Db3 is used to estimate center by FCM. The estimated center is been used as a training and testing patterns for MLPNN. Here there is no evidence for speed of execution.

Considering all these issues related with prediction and grouping of ECG signals, FFNN classifier based on Fuzzy c-means clustering with wavelet transform is proposed in the paper. Here, the classifier is trained with only center, which is been extracted from FCM for the classification purpose of all the ECG signals considered in the case study. The feed forward neural network provides accurate results with center alone. Thus the proposed method provide robust and accurate result for ECG classification with less number of data.

To summarize, the paper deals with generation of various ECG signals based on the specification in terms of diseases, then the work handles with wavelet transformation for extracting various data needed for doing analysis with FCM. Next the paper describes the application of center in FFNN for the classification and finally it deals with the results and discussion showing the performance of FFNN with and without using FCM.

II. WAVELET TRANSFORM

Frequency representation are mostly preferred for the analysis that will be done on signals which are characterized by non-stationary parameters. As ECG signals are non-stationary one due to various reasons, which will have to be analyzed with the system which are capable enough to deal with time and frequency. So WT has the ability to process the non-stationary signals which is an ECG signals taken in this research. In any analysis done with WT is mostly represented in different resolution by using high pass and low pass filter to split the signals to extract various coefficients. This WT has the ability to compute and manipulate data even in compressed form. So with the wavelet transform, important data can be extracted from the ECG signal which can be used for recognition and diagnosis. Selection of mother wavelet, its order and level is very important to extract key data , which normally vary problem to problem, here Daubechies 4 is been used [10]. Equation 1 and 2 can be used to compute wavelet coefficients using Daubechies 4 and can be used in subsequent analysis.

$$\phi_{j,n}[t] = 2^{j/2} \sum_n c_{j,} n\phi[2^j \ t - n] \ldots\ldots\ldots\ldots\ldots\ldots \text{(1)}$$

$$\varphi_{j,n}[t] = 2^{j/2} \sum_n d_{j,} n\varphi[2^j \ t - n] \ldots\ldots\ldots\ldots\ldots\ldots \text{(2)}$$

Scaling functions $\phi_{j,n}[t]$ and wavelet function $\varphi_{j,n}[t]$ are the two sets employed by WT, which are associated with low-pass and high-pass filters for decomposition purpose.In equation 1 and 2, $c_{j,}$and $d_{j,}$ are the scaling and wavelet coefficients, which are used to perform both MRA decomposition and reconstruction of the signals.

III. FUZZY C-MEANS CLUSTERING

Standard deviation, Mean and Median absolute deviation is been extracted from the waveforms with the help of WT. After the extraction the most distinguished features that is much suited in the analysis will be extracted with Fuzzy C-means clustering to classify and group the distinct signals. In multi-dimensional space Fuzzy C-means algorithm groups the data points in to a specific number of cluster .The extracted feature by WT were given as a input to Fuzzy c-means clustering to determining center c_i and the membership matrix U which is done based on minimization of the objective function shown in equation (3).

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} {u_{ij}}^m \parallel x_i - c_i \parallel^2 \ldots\ldots\ldots\ldots \text{(3)}$$

$$C_i = \frac{\sum_{i=1}^{N} {u_{ij}}^m x_i}{\sum_{i=1}^{N} {u_{ij}}^m} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(4)}$$

$$u_{ij} = \sum_{\mathrm{k}=1}^{\mathrm{c}} \left[\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right]^{\frac{-2}{\mathrm{m}-1}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(5)}$$

Where $u_{ij}$is the degree of membership of x_i in cluster j, m is the number of clusters, and c_i is the n-dimensional centres of the cluster which is shown in equation (4) and (5) [9].

IV. FEED FORWARD NEURAL NETWORK

A feed forward neural network maps set of input data on to a set of output. It is just a modification of linear perceptron with three or more layers of neuron with nonlinear activation function. FFNN is more powerful than standard perceptron, as it can distinguish data that is nonlinearly separable. In figure1a general multilayer feed forward neural network is depicted. Based on the task the input space can be either an actual or a normal representation. Neuron is the basic processing element with all its inherent properties are used to perform classification task.

The output of the neuron can be estimated by using equation 6 [11]

$$y_j(p) = f\left(\sum_{i=1}^{n_{p-1}} \mathrm{w_{ij}}\, y_j(p-1) + b(p)\right) \ldots\ldots\ldots\ldots \text{(6)}$$

Where $y_j(p)$ is the output of the j^{th} neuron in the p^{th} layer , $\mathrm{w_{ij}}$ is the weight from i^{th} neuron in the $(p-1)th$ layer to the j^{th}

neuron in the p^{th} layer. *b(p)* and *f(.)* is the bias and activation function.

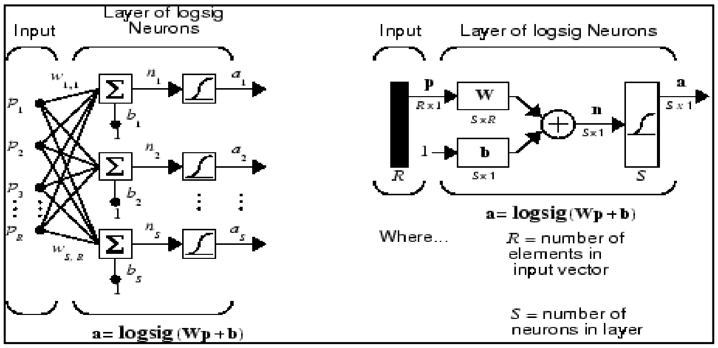


Figure 1. FFNN architecture

V. RESULT AND DISCUSSION WITH CASE STUDY

A. Data generation and Pre-processing

In this work, WT, FCM and FFNN is been used to classify the ECG signals which is described sequentially. To carry out this research three ECG signals were taken in to account such as Longest Episode of Ventricular Tachycordia (LE-VT), Conversion from Atrial Fibrillation to Normal Sinus Rhythm (CAT-NSR) and Complete Heart Block (C-HB). These all are generated with the help of MATLAB coding [12] by using the specifications given by the expert physicians and researchers which all have discussed in the literature. The generated waveforms are shown in figure 2 to 4.

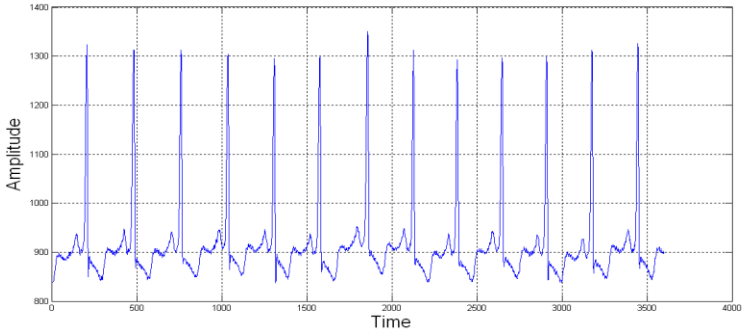


Fig 2. Longest Episode of Ventricular Tachycordia (LE-VT)

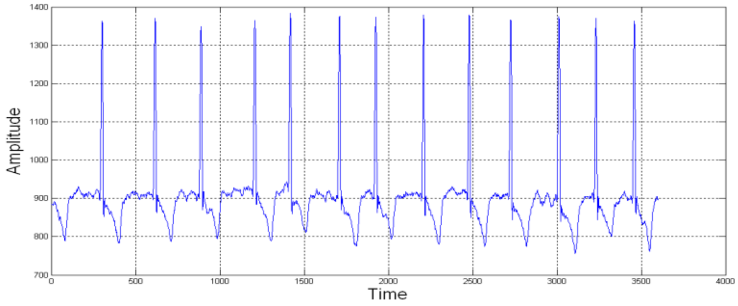


Fig 3. Conversion from Atrial Fibrillation to Normal Sinus Rhythm (CAT-NSR)

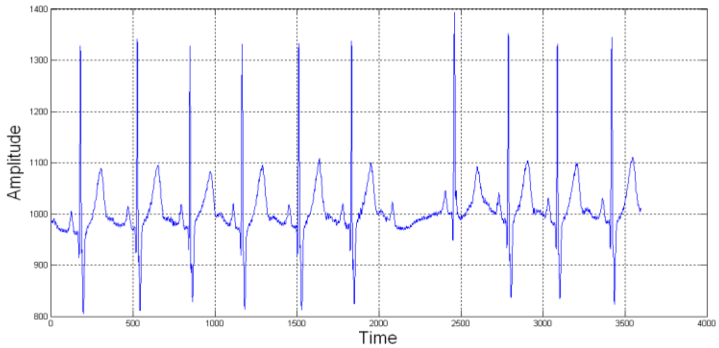


Fig 4. Complete Heart Block (C-HB)

These waveforms are single waveform which will not be enough to carry out the classification issues. So, for each case 100 orientations were generated, so totally for 3 cases 300 orientations were used in the analysis. The generated waveforms are used and processed with WT, to extract key data such as Mean, Median, Standard deviation, Median absolute deviation, L1 norm …etc. Here only Standard deviation, Mean and Median absolute deviation alone is used, which all are depicted in figure 5 to 7.

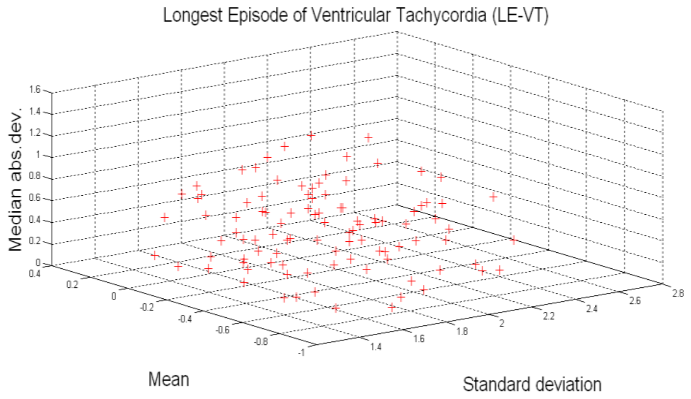


Fig 5. Position of data of (LE-VT) in 3D space

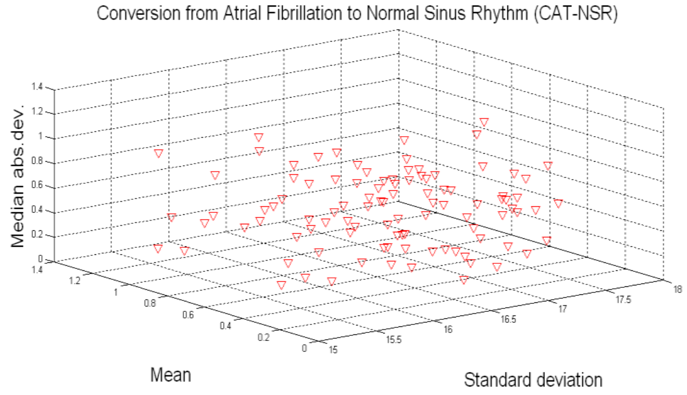


Fig 6. Position of data of (CAT-NSR) in 3D space

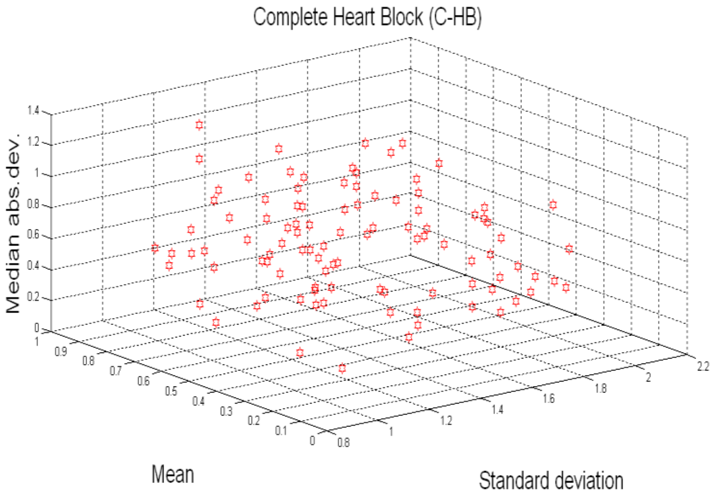


Fig 7. Position of data of ((C-HB)) in 3D space

In these figures, shows how each data with respect to Standard deviation, Mean and Median absolute deviations were placed in the 3 dimensional space that contributes for the severity of the heart diseases. Here data were preprocessed and normalized to have a uniform ranges between the features with respect to amplitude, time duration and various other conditions. These normalized data were arranged as a matrix to depict how these data peak value, lowest values are arranged as an array to make as to select some of key projection in terms of data deviations. Figure 8 to 10. shows the sets of data which is been distributed both in XYZ plan as well as, in a histogram format to represent data in matrix with diagonal and off diagonal data sets of Longest Episode of Ventricular Tachycardia (LE-VT), Conversion from Atrial Fibrillation to Normal Sinus Rhythm (CAT-NSR) and Complete Heart Block (C-HB).

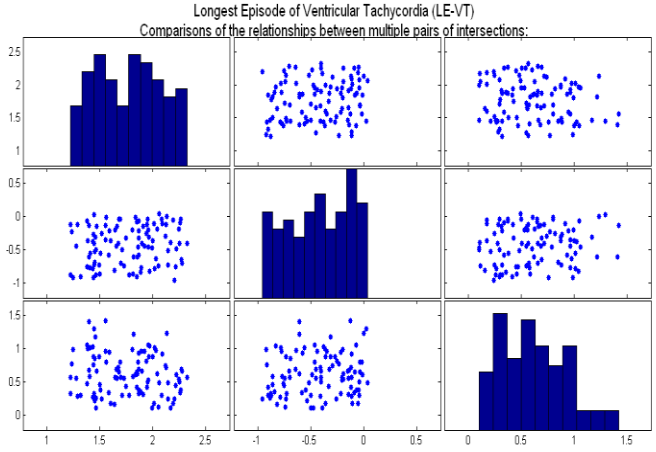


Fig 8.Comparisons of data intersections of (LE-VT)

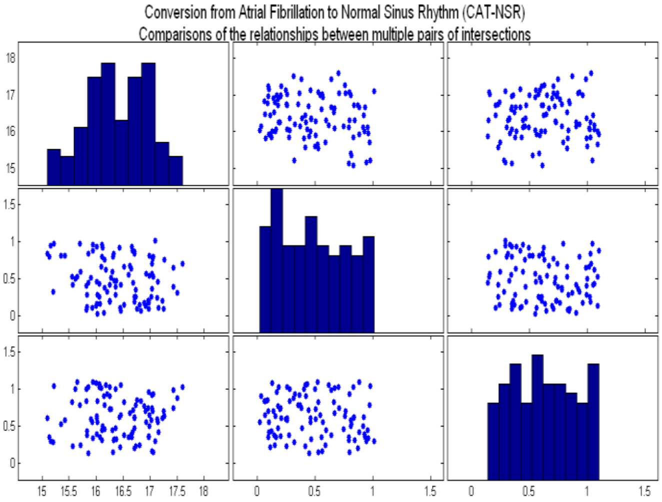


Fig 9 .Comparisons of data intersections of (CAT-NSR)

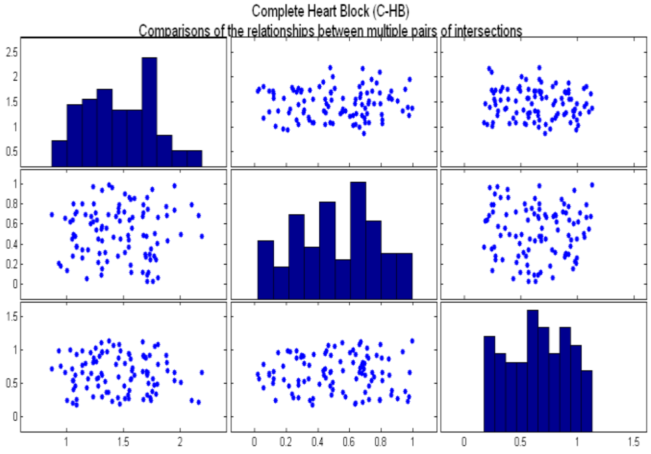
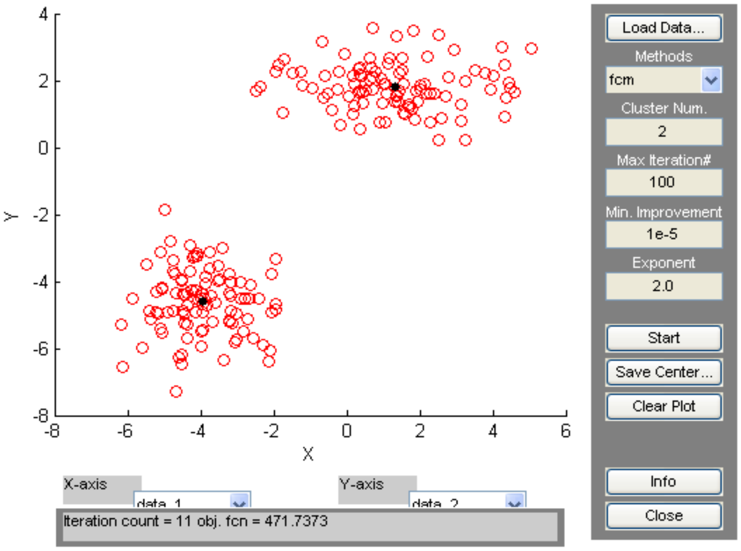


Fig 10 .Comparisons of data intersections of (C-HB)

B. Clustering and Extraction of Data

In this work our main focus is clustering the data, so the clustering is very important step in any data mining problem to extract key data. This technique has the capability to find inherent nature of any data which is the key for discriminating any data sets. The reasons behind the selection of FCM is ,it is a technique which is proved to be more general and useful in case of finding the attributes that contributes to the points being similar to others as well as dissimilar from other. The main objective of FCM is minimization. The first step in the FCM is selection of number of cluster based on the problem and initializing the number of membership matrix according to the problem. The membership value for all points of each cluster is nothing but the element present in the matrix. The initiation and centers are computed randomly at the start of the algorithm.

The centers are calculated such that the center is closer to the value having highest membership value to one cluster.



Methods fcm

Cluster Num. 2

Max Iteration# 100

Min. Improvement 1e-5

Exponent 2.0

X-axis data 1

Y-axis data 2

Iteration count = 11 obj. fcn = 471.7373

The FCM process is shown in the figure 11. for the group namely for the Longest Episode of Ventricular Tachycardia (LE-VT) and Abnormal AV Conduction (A-AVC).The membership matrix is updated based on estimating the cluster centers and location of class centers. To calculate the new membership value for a particular cluster, the distance of that cluster center and other cluster centers are taken into account. The changes in the matrix were monitored, if the changes in the membership matrix is lower than a threshold, the process is being stopped, otherwise new cluster centers are calculated and updated with respect to new cluster centers. This process is continued, till the changes in the membership matrix are minimum.

The FCM finally generated a list of data namely center that contains the information in the order of their relevance. The most relevant data are the start while the least relevant are the end of the process. To perform prediction and classification, the most relevant data are used.

C. *FFNN based classification*

The FFNN is used as a classifier for the classification of ECG signals. Input signals for training are the data which is selected from the data extracted using wavelet transform. Around 70 % of data is been used for training of all the three heart conditions. In this, training was adjusted according to its error. From the data 15% of the data were used for validation to measure network generalization, and to halt training when generalization stops improving. Another 15% percentage is used to test and measure the network performance during and after training. Based on the trial and error method for each learning algorithm the hidden neurons are adjusted, but initially it is started with 20 for the entire network. Mean Squared Error is checked, which is nothing but the average squared difference between outputs and targets. Based on the percentage of error the classification rate were calculated which is shown in the table 1.

ECG signals	**Classification rate %**	
	With Out Fuzzy C-mean Clustering	With Fuzzy C-mean Clustering
LE-VT	96	98
CAT-NSR	95	96
A-AVC	92	99

VI. CONCLUSION

In this paper the two important tasks involved in the classification of Longest Episode of Ventricular Tachycardia (LE-VT), Conversion from Atrial Fibrillation to Normal Sinus Rhythm (CAT-NSR) and Complete Heart Block (C-HB) of ECG signals are data mining and clustering. For data mining initially, it was done with wavelet transforms main to extract basic features and then normalized, clustered using FCM to generate centers. Vectors for all the considered cases are generated by extracting the most relevant data from the particular clusters particularly centers is used for classification. FFNN is been used to perform the classifier role. The extracted data by FCM are used for training, testing and validation. The technique performs well with respect to classification when compared with other conventional approach.

REFERENCES

[1] M. K. Sarkaleh and A. Shahbahrami, "Classification of ECG arrhythmias using Discrete Wavelet Transform and neural networks, "International Journal of Computer Science Engineering and Applications (IJCSEA), vol. 2, no. 1, pp. 1-13, 2012 .

[2] E. Zeraatkar et al., "Arrhythmia detection based on Morphological and time-frequency, Features of t-wave in Electrocardiogram," J. of medical signals and sensors, vol. 1, no. 2, pp. 99-106, 2011.

[3] A. Vishwa, M. K. Lal, S. Dixit, and P. Vardwaj, "Classification of arrhythmic ECG data using machine learning techniques," Int. J. of Interactive Multimedia and Artificial Intelligence., vol. 1, no. 4, pp. 68-71, 2011.

[4] M. K. Das and S. Ari, "ECG Beats Classification Using Mixture of Features" Int. Scholarly Research Notices, 2014.

[5] N. Kannathal, U. R. Acharya, C. M. Lim, P. K. Sadasivan, and S. M. Krishnan, "Classification of cardiac patient states using artificial neural networks," Experimental & Clinical Cardiology, vol. 8, no. 4, pp. 206-211, 2003.

[6] D. Joshi and R. Ghongade, “Performance analysis of feature extraction schemes for ECG signal classification,” International Journal of Elect., Electron and Data Communication, vol. 1, pp. 45-51, 2013.

[7] A. Dallali, A. Kachouri, and M. Samet, "Classification of Cardiac Arrhythmia Using WT, HRV, and Fuzzy C-Means Clustering," Signal Processing: An Int. J. (SPJI), vol. 5, no. 3, pp. 101-109, 2011.

[8] J. S. Wang, W. C. Chiang, Y. T. Yang, and Y. L. Hsu, "An effective ECG arrhythmia classification algorithm," Bio-Inspired Computing and Applications, Springer Berlin Heidelberg, pp. 545-550, 2012.

[9] Y. Ozbay, R. Ceylan, and B. Karlik, "A fuzzy clustering neural network architecture for classification of ECG arrhythmias," Computation in Biology and Medicine, vol. 36, no.4, pp. 376-388, 2006.

[10] Ceylan R. and Ozbay Y, " Comparison of FCM, PCA and WT techniques for classification of ECG arrhythmias using artificial neural network, Expert System with Application 33:286-296, 2007.

[11] S. Haykin, Neural networks: Comprehensive Foundation: Prentice Hall, 1999.

[12] Wavelet, Toolbox, in MATLAB R 15.