DISEASE CLASSIFICATION USING ECG SIGNAL BASED ON PCA FEATURE ALONG WITH GA & ANN CLASSIFIER

AMANJYOT KAUR[1], ANITA SUMAN[2]

M.Tech[1] (Scholar, Ece), Bcet, Gurdaspur, Punjab, India,
M.Tech[2] (Assistant Professor, Ece) Bcet, Gurdaspur, Punjab, India,

---***---

Abstract- *Electrocardiogram (ECG), a non-invasive technique is utilized as a primary diagnostic tool for cardiovascular disease. Cleared ECG signals offer essential information about the electrophysiology of the heart diseases and ischemic modifications that may occur. It provides valuable information about the functional aspects of the heart and cardiovascular system. The purpose of this research work is to classify the disease dataset using Genetic Algorithm and train by artificial neural network on the basis of the features extracted and also to test the image on the basis of the features at the database and the features extracted of the waveform, to be tested. The advantage of proposed method is to minimize the error rate of the classification which occurs due to insignificant count of R-peaks. Database from physionet.org has been used for performance analysis. Several experiments are performed on the test dataset and it is observed that Artificial neural network classifies ECG beats better as compared to K-nearest neighbor (K-NN). Precision, Recall, F-measure and accuracy parameters are used for detecting the ECG disease. All the simulation process will be measured in MATLAB environment.*

Keywords- **ECG signals, PCA (Principal component analysis), *(GA) Genetic algorithm*, (ANN) Artificial neural network, MATLAB.**

1. INTRODUCTION

1.1 ECG (Electro-Cardiogram)

Electro-Cardiogram is used to access the electrical activity of a human heart. The diagnosis of the heart ailments by the doctors is done by following a standard changes. In this project our aim is to automate the above procedure so that it leads to correct diagnosis [1].

Heart disease contains any disorder that influences the heart's ability to function normally. Over the last decades, many physicists have been an increasing effort to develop computer-based mechanical diagnostics of the ECG. The ECG signal is created by electrical current of the heart. The shapes of the ECG waveform depend on the anatomic features of the human body and heart, and thus are distinctive from one person to another. Figure 1 illustrates the human heart and the signals responsible for generating a normal ECG signal. Each portion of a heartbeat produces a different deflection on the ECG. ECG signals are recorded as a series of positive and negative waves. The first peak (P wave) of the normal heartbeat is a small upward wave which indicates atrial depolarization. Approximately 160ms after the onset of the P wave, the QRS wave is caused by ventricle depolarization. Finally, one observes the ventricular T wave in the electrocardiogram which represents the stage of repolarization of the ventricles.

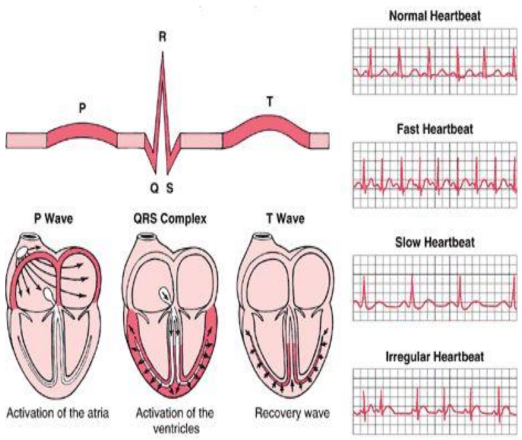


Fig-1: Human heart and the signals

1.2 PCA (Principle component algorithm)

PCA [2] method is just the effective feature extraction method based on face global feature. PCA is considered as one of the most successful linearity analysis algorithm. It can reduce the dimension effectively and hold the primary information at the same time.

1.2.1 TRADITIONAL PCA METHOD

Traditional PCA method was also named Eigen-face. The primary principle of PCA can be simply stated as followed:

Suppose that there are a set of N training face images $\{Y_1, Y_2, Y_3 \dots, Y_N\}$ taking values in a n-dimensional image space. The purpose of the PCA is to find a n×m linear transformation matrix P, transforming the n-dimensional vector YK (1 ≦ K ≦ N) to a m-dimensional, more representative (enlarging the difference between each two images) vector XK, where $X_K = S^U X_K$ (K=1,2…N).

Let Y' be the average feature vector before transformation, then X' is the average value of all training samples. The new average vector after transformation is as followed:

$$Y=\frac{1}{M}\sum_{K-1}^{M} Y_K \qquad (1)$$

Employ covariance matrix to represent the scatter degree of all feature vectors relate to the average vector. Subtract all the image samples to the average vector X', then the covariance matrix before transformation is:

$$T_y = \frac{1}{M}\sum_{K-1}^{M}(Y_K - Y)\ (Y_K - Y) \qquad (2)$$

In order to maximize the scatter degree of ZK , the transformation matrix must just be the matrix which compose by the eigen-vectors of S_x. Because every eigenvector has the face-like shape, they are called as eigen-faces. Normalize all eigenvectors can eliminate the correlation of them, and form a set of orthogonal projection coordinates bases. The training images are seemed to be the weighted combination of the set of orthogonal bases. Compute the eigen values of Matrix Sx: let the eigen values queue from large to small, then the m largest eigen values u1, u2…um are obtained. Let the transformation matrix P as: P= [u1, u2…um]T, then it can be used to transform the n-dimensional face images to m-dimensional weighted vectors, but also remain main information of the original images. Through these, it can play well performance in dimension reduction.

1.3 Genetic Algorithm(GA)

In pattern classification problem, the undesirable or redundant features increase the complexity of the feature space or may decrease the classification accuracy [3]. Therefore, GA is used in this work to select the optimum number of features by eliminating the undesirable features which in turn improves the classification performance. Here, binary coding system as shown in Fig. 2 is used to represent the chromosome. If i^{th} bit of the chromosome is 1 then corresponding feature is selected and if bit of the chromosome is 0 then corresponding feature is ignored. The description of the genetic algorithm (GA) based optimum feature selection method is mentioned below [15]:

Step 1: **Initialize the population**: Initialize the number of chromosomeN_p, where each chromosome represents binary bit pattern of N bits. Keep in mind N_p is population size and N is chromosome length.

Step 2: **Feature subset selection**: A feature subset is selected according to Fig. 2.

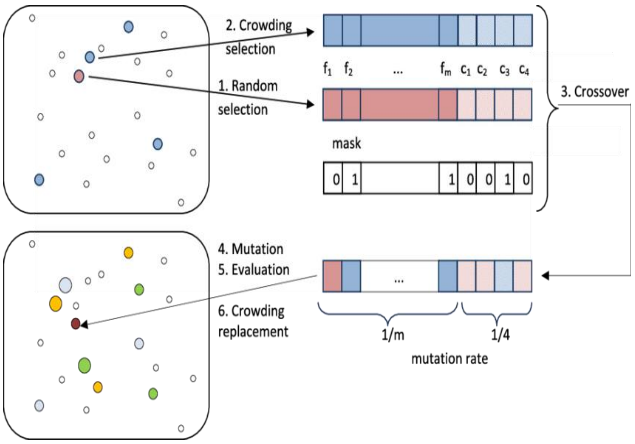


Fig-2: Chromosome bit pattern of the selected feature set.

Step 3: **Fitness calculation**: For each corresponding chromosome, training data set is used to train the N classifier and fitness function is evaluated based on mean square error (MSE) values. Fitness value uniquely separates one chromosome from others. The fitness value is calculated by the following equation:

Fitness=0.8*(nte_{train})-1+0.2*bct(M-M_t)

Where, nte_{train} is a training of nte M is the selected feature and (M−Ms)is the number of reduced features. High fitness value is obtained from the individual chromosome with low training nte and high number of reduced features.

Step 4: **Termination criteria**: When the termination criteria (reaching of maximum generation) is satisfied, the process ends and find the optimized solution which gives the maximum fitness value; otherwise, it proceed with the next generation.

Step 5: **Genetic operation**: In this step, the system searches for better solutions by genetic operations which includes selection, crossover, mutation, and replacement. In this context, Roulette wheel selection is used to select the parent chromosomes in the current generation based on their fitness value. Uniform crossover operator is applied to generate the offspring (child chromosome) and replaces the parent chromosome which has participated in the crossover operation depending on the value of the crossover probabilityp_c. After crossover, mutation operation is applied to flip the bit of the parent chromosomes based on the mutation probability(p_m). Finally, all the chromosomes of previous population (except best) are replaced by the chromosomes of current generation.

1.4 Artificial Neural Network (ANN)

Our ANN model is a multilayer feed-forward network trained to perform classification from the neural network toolbox in Matlab [4]. The ANN model has 3 distinct layers namely input layer, hidden layer, and output layer. The input layer consists of 800 input neurons (for a single electrode) where each input equals to one data points from EEG data shown in Figure.

The hidden layer consists of 10 neurons where the weights are initialized randomly and tansig was used as the activation function for the neurons. The number of neurons in hidden layer was tested using a range of value ranging from 5 to 15 neurons. The network performed marginally better for 10 neurons in the hidden layer compared to other tested values. There are 2 output neurons which represent normal and depressive state. The network was trained using scaled conjugate gradient back-propagation learning algorithm which is referred to as train's cg function in Matlab .

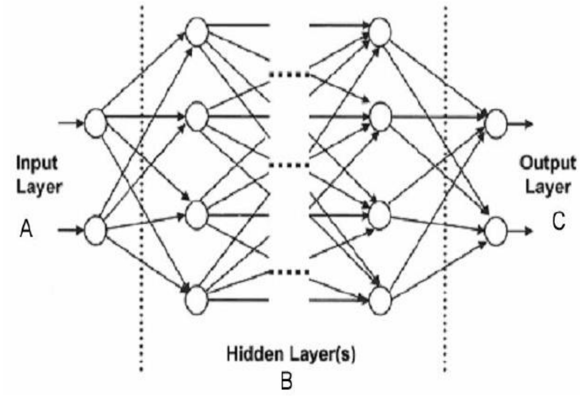


Fig-3: Multilayer feed-forward network

Trains cg learning algorithm was used because of the memory requirements are relatively small and yet much faster than other standard learning algorithms available in the toolbox. The training of the network stops when training exceeds the maximum number of epochs (1000) or exceeds the maximum amount of time (3600s), or when performance meets the goal, or when the performance gradient falls below minimal gradient.

2. LITERATURE REVIEW

C.R et.al, 1977, proposed the work in which noise removal algorithm has been proposed using aseline of ECG signal. In this way low frequency noise has been removed without affection ST segments.

C.S.E.W, 1985, proposed an evaluation method of wave recognition computer programs. In the end an evaluation method for testing of ECG signal has been recommended. And these are based on amplitude and interval details.

Islam, M. K., 2012 deals with learns and examination of ECG signals processing by means of MATLAB tool successfully. Study of ECG signal involve making & simulation of ECG signal, attainment of real time ECG data, ECG signal filtering & giving out, feature extraction, assessment between different ECG signal analysis algorithms & technique, recognition of any abnormalities in ECG and calculating beat rate and so on utilizing the majority familiar and versatile MATLAB software beside with lab view. An utilization of toolbox, MATLAB functions and Simulink can guide us to work with ECG signals for allowance and analysis both in real time and by simulation with great accuracy and ease.

Das Manab Kumar, et.al, 2013 designed a classifier model so as to categorize the beat from ECG signal of the MIT-BIH ECG database. The classifier model comprised of three important stages: feature extraction, selection of qualitative features; and determination of heartbeat classes. In first stage, features were extracted using S-transform where-as second stage utilizes the genetic algorithm to optimize the extracted features which represent the major information of the ECG signal. The final-stage classifies the ECG arrhythmia. In this study, authors have classified six types of arrhythmia such as normal, premature ventricular contraction, atrial premature contraction, right bundle branch block, ventricular fusion and fusion. The experimental results indicate that proposed method gave better result than earlier reported techniques.

Table-1: Comparison of existing techniques

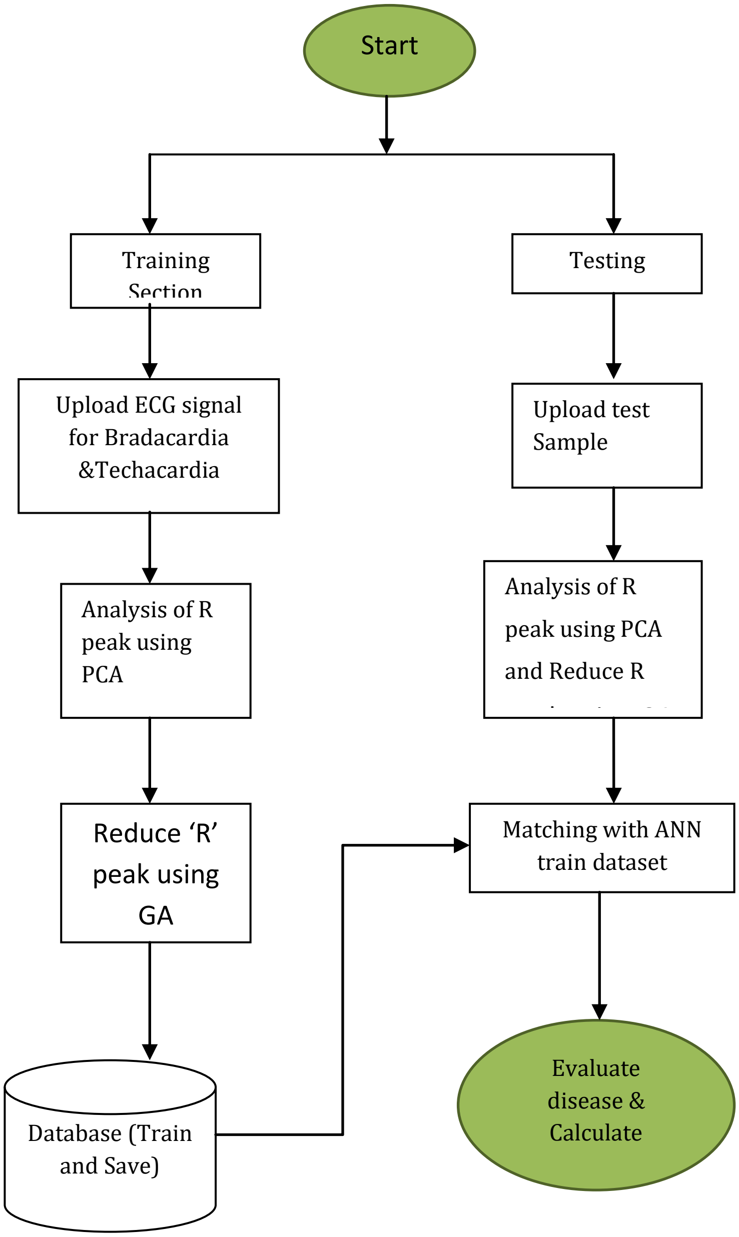
Authors	Simulator	Techniques used	Descriptions	ECG data set
Y. Jewajinda and P. Chongstitvatana (2010)	MATLAB	Genetic algorithm and neural Network	Used for online ECG heart beat recognition using feature extraction and	The MIT-BIH arrhythmia

			classification.	
J. A. Nasiri et al. (2009)	MATLAB	Genetic algorithm and Support vector machine	Genetic algorithm will find the best value by searching and thus optimize th classification fitness function.	The MIT-BIH arrhythmia. The database has 48 records with every record being an ECG signal for the period of 30 minutes.
R. Poli et al. (1995)	MATLAB	Genetic algorithm	GA was used to enhance the QRS complex detection.	The MIT-BIH arrhythmia.
Acharya et al. (2003)	MATLAB	ANN and Fuzzy logic	Both are the classifiers that were used to diagnose ECG signal. Accuracy up to 95% has been obtained.	The MIT-BIH arrhythmia
Ceylan, Rahime, and Yüksel Özbay (2007)	MATLAB	PCA and Wavelet transform were used for feature extraction. Fuzzy and neural network were used.	Authors taken the record of 92 patient in which 40 are male and 52 are females of average age 39.75±19.06. It was concluded that Fuzzy c-mean PCA-NN system perform better than PCA_NN	MIT-BIH ECG database
S. M. Jadhav et al. (2010)	Neuro Solutions (version 5.0)	ANN	Collected data from 452 patients and calculated different parameters like mean squared error (MSE), receiver operating characteristics (ROC) and area under curve (AUC).	UCI ECG arrhythmia data set.
Güler, İnan, and Elif Derya Übeylı(2005)	MATLAB version 6.0 with neural networks toolbox)	Neural network	ECG signal features were extracted using discrete wavelet transform technique. The accuracy of the combined neural network model was higher than the normal neural model.	MIT-BIH ECG database
Rai, Hari Mohan et al. (2013)	MATLAB software package 7.13.	ANN along with Descrete wavelet transform for extracting features.	Authors took total 48 files out of which 25 files were from normal class and 20 file. Back propagation, Feed forward network and multilayred preceptors were used as a classifier	MIT–BIH arrhythmia database

3. Methodology

Step 1: To analyze Bradacardia and Tachycardia heart disease, the whole process is divided into two steps named as:

- Training phase
- Testing Phase


Start
Training Section
Testing
Upload ECG signal for Bradacardia &Techacardia
Upload test Sample
Analysis of R peak using PCA
Analysis of R peak using PCA and Reduce R
Reduce 'R' peak using GA
Matching with ANN train dataset
Database (Train and Save)
Evaluate disease & Calculate

Flow chart-1: Proposed work Flowchart

1. Training Phase

i. Upload ECG signal dataset for Bradycardia and Tachycardia heart disease.
ii. Extract feature from the uploaded ECG signal based on the threshold value according to the QRS peaks.
iii. Develop a code for the Genetic algorithm to optimize the features according to the objective function of GA. s
iv. Store data in Data base.

2. Testing phase

i. Upload test waveform for testing the dieses.
ii. Analyze R peak of the QRS complex and then optimize that by using GA.
iii. Initialize Artificial neural network (ANN) for classification purpose.
iv. Classify the Diseases according to categories which are generated during the training phase. And then calculate the performance parameters.

4. Result and Discussion

This section explains the results obtained after the implementation of the proposed work. In this research work, a system is developed for ECG disease that is Bradycardia and Tachycardia recognition. For optimizing the extracted features Genetic is used whereas, for classification ANN is used.

Above figure shows the main window of the proposed ECG Disease Detection System. The system is mainly categorized into three steps: Feature Extraction & optimization, training and testing. In the given figure there are two panels named as Training and testing panel. After uploading the samples for ECG signal, next step is to extract features. When we click on QRS complex, the waveform gets displayed on the screen. These signals are finding on the basis of threshold technique.

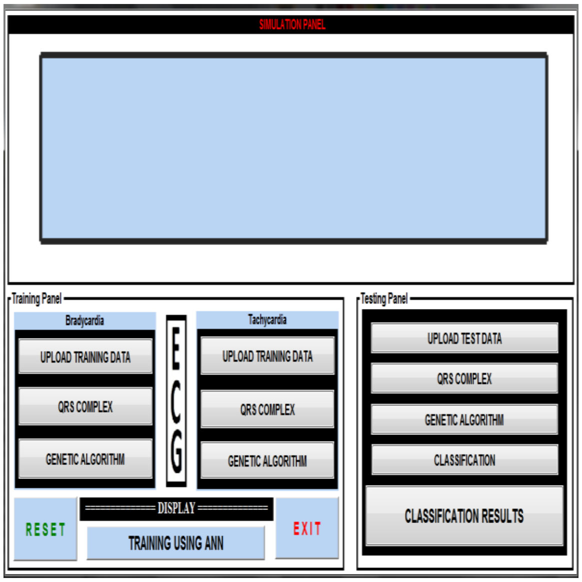
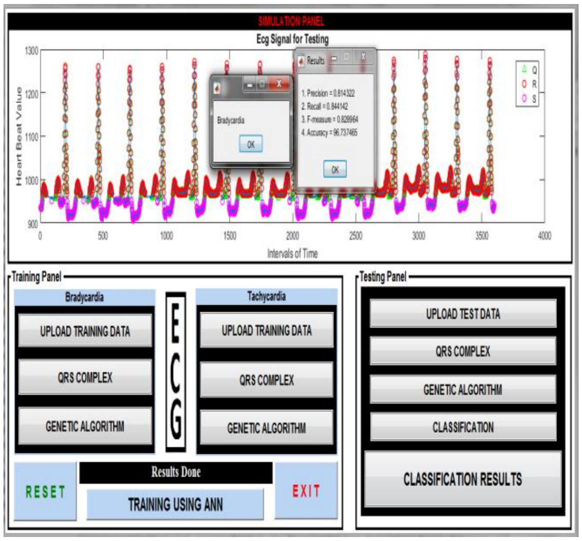


Fig-4: Main window of the proposed ECG Disease Detection System

After extracting the QRS complex from the ECG signal. The extracted features are improved by using optimization algorithm known as Genetic algorithm.



Bradycardia

1. Precision = 0.814322
2. Recall = 0.844142
3. F-measure = 0.828964
4. Accuracy = 96.737465

Fig-5: Classification of test data using ANN

In the proposed work, the cardiac disorder was classified into two parts named as: (i) Bradycardia (ii) Techycardia. For effective training, it is desirable that the training data set be uniformly spread throughout the class domains. The available data can be used iteratively, until the error function is reduced to a minimum.

Table-2: Performance parameters for cardiac disorder

S No.	Precision	Recall	F-measure	Accuracy
1	0.814	0.844	0.828	96.73
2	0.736	0.836	0.782	95.81
3	0.937	0.735	0.823	98.79
4	0.983	0.839	0.905	97.84
5	0.749	0.913	0.823	96.78
6	0.983	0.833	0.902	97.88
7	0.749	0.737	0.743	97.74
8	0.846	0.830	0.837	98.36
9	0.887	0.874	0.881	99.63
10	0.846	0.739	0.788	97.69

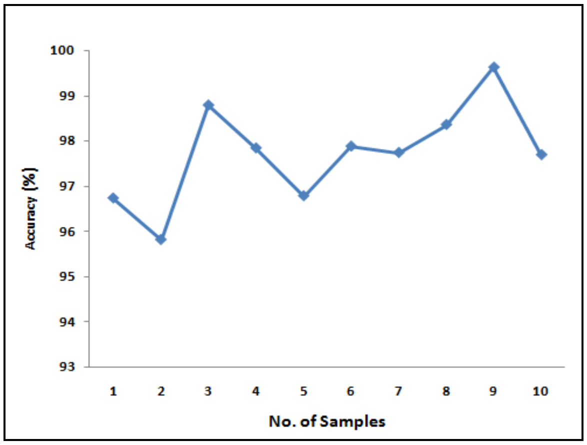


Chart-1: Accuracy of the ECG signal

The graph obtained for accuracy has been displayed above. The graph has number of sample along x-axis and accuracy value along y-axis. The average value of accuracy obtained for the proposed work is 97.72.

5. CONCLUSION

ECG signals are used for detecting the cardiac diseases and the improvement in ECG feature extraction has become important for diagnosing the long recording. An ECG signal is a graphical representation of the cardiac movement for computing the cardiac diseases and to ensure the abnormalities in the heart. The objective of this research work is to categorize the disease dataset using Genetic algorithm and to train the Neural Network on the source of the features extracted and moreover to test the image on the origin of the features at the database and the features extract of the image to be tested. This research was based on studying the executed approaches in the ECG diseases and then to recommend a novel practice /algorithm for classification of two cardiac disorders named as Bradycardia, and Tachycardia reliant on Artificial neural network and the Genetic algorithm. Clinical databases have accrued large quantities of knowledge regarding patients and their medical condition. This dissertation presents the ECG Disease Detection System based on GA, and ANN, in which detection is based on performance parameters like precision, Recall, F-measure and Accuracy. Simulation results have shown that the obtained value of precision, recall, f-measure and accuracy in favor of proposed tested waveform are 0.853, 0.818, 0.8312 and 97.72 % for accuracy.

6. Future Scope

Future scope lies in the use of former classifiers like SVM with the aim of having multidimensional data and making use of feature reduction algorithms, so that accuracy rate can be enhanced. SVMs bring a unique solution, since the optimality problem is rounded. This is an advantage to Artificial neural network (ANN) which has several solutions related with local minima and for this reason may not be tough over different samples. For optimization algorithms similar to artificial bee colony (ABC) and (PSO) Particle swarm optimization would be used.

ACKNOWLEDGEMENT

I would like to get this opportunity to convey my thoughtful gratitude and regard to my Guide Mrs. ANITA SUMAN, for her commendable guidance, precious feedback and constant support throughout the period of the project. Her valuable ideas were of immense helpful throughout my development work. Her perceptive condemnation kept me working to make this task in a much better way. Working under her was an tremendously knowledgeable experience for me.

References

[1] Karimian, Nima, et al. "Highly reliable key generation from electrocardiogram (ECG)." IEEE Transactions on Biomedical Engineering 64.6 (2017): 1400-1411.

[2] Khandelwal, Chhaya Sunil, et al. "Review Paper on Applications of Principal Component Analysis in Multimodal Biometrics System." Procedia Computer Science 92 (2016): 481-486.

[3] Desale, Ketan, et al.. "Comparative Study of Pre-processing Techniques for Classifying Streaming Data." (2015).

[4] Saravanan, K., et al. "Review on classification based on artificial neural networks." International Journal of Ambient Systems and Applications (IJASA) Vol 2.4 (2014): 11-18.