Long-tail Effect on ECG Classification

N. H. Thai, N. T. Nghia, D. V. Binh, N. T. Hai, and N. M. Hung, *Member, IEEE*

***Abstract*— Heart disease affects seriously to human health. ECG signal is critical information to help doctor with heart diagnose prediction. In previous studies on ECG classifier, state-of-art method use MIT dataset to evaluate prediction result and record a high accuracy. However, the dataset has a long tail phenomenon where the number of normal beats is cover 83,6% of all dataset whereas some diagnose beats have a few samples. Therefore, in this paper, the state-of-art method was used to evaluate the system performance where long tail effect is removed. This method was tested in two scenarios, the first scenario is that it considers normal beat as a class for recognition, therefore we could have long-tail effect in the result. The second only consider diagnose beats where long-tail effect is removed. The experiment proves that long-tail phenomenon could affect seriously to prediction result.**

Keywords— ECG, PCA, Neural Network, Model Selection

I. INTRODUCTION

An electrocardiogram (ECG) is a medical test that detects cardiac abnormalities by measuring the electrical activity generated by the heart as it contracts. The ECG can help diagnose a range of conditions including heart arrhythmias, heart enlargement, heart inflammation (pericarditis or myocarditis) and coronary heart disease. The electrical potential generated by electrical activity in cardiac tissue is measured on the surface of the human body. Current flow, in the form of ions, signals contraction of cardiac muscle fibers leading to the heart's pumping action. It is a non-persistent recording produced by an ECG machine. The ECG machine records the electrical activity of the heart muscle and displays this data as a trace on a screen or on paper. The ECG data from normal, healthy hearts have a characteristic shape. Any irregularity in the heart rhythm or damage to the heart muscle can change the electrical activity of the heart so that the shape of the ECG is changed.

The ECG signal is essential for the treatment of patients. Early and accurate detection of the ECG arrhythmia helps doctors to detect various heart diseases. There have been many previous studies on MIT ECG classification dataset with high accuracy [1-3]. The ECG data contained many heart rhythms but also includes a wide variety of noise. These noises can cause ECG analysis difficult. Before ECG data is classified, ECG data should be filtered to remove unwanted noise components. There are many studies eliminated noise components on the ECG signal [4-9].

N. H. Thai and D. V. Binh, Master student, are with the Ho Chi Minh City University of Technology and Education, Ho Chi Minh, Viet Nam. Email: 11141188@student.hcmute.edu.vn (N. H. Thai), 11141013@student.hcmute.edu.vn (D. V. Binh)

N. T. Nghia, PhD student, is with the Ho Chi Minh City University of Technology and Education, Ho Chi Minh, Viet Nam. Email: 1627003@student.hcmute.edu.vn.

N. T. Hai and N. M. Hung, Doctor, are with the Ho Chi Minh City University of Technology and Education, Ho Chi Minh, Viet Nam. Email: nthai@hcmute.edu.vn (N. T. Hai), hungnm@hcmute.edu.vn (N. M. Hung).

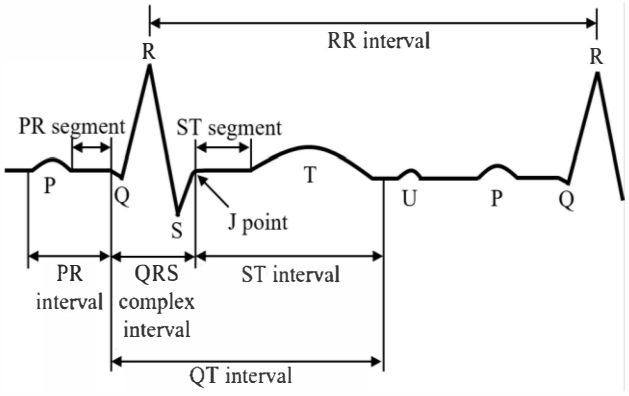

Figure 1. The shape of ECG beat

After a filtering process, ECG signal will be deducted characteristic. Characteristic of ECG signal is shaped on the basis of the waveform P, Q, R, S, T [10] as shown in Fig. 1. Some characteristic extract methods of ECG signal was announced as ST, CWT, DWT, DCT, Pan Tompkins [8, 11-13]. From here, the characteristics of the ECG signal will be dimensionality reduced to take on the training. The dimensional reduced methods ECG including PCI, LDA, ICA, FCM, GA, Symmetric uncertainty [2, 8, 11, 14].

Classification of the ECG signal is also an important task to understand the heart condition. Classification and detection of abnormal types can help in identifying the abnormality present in the ECG signal of a patient. Following multiple signal classification used heart rate and high accuracy is obtained as MLPNN Modular neural network, Generalized FFNN, Modular neural network, Feed forward PNN, SVM, SVM classifier with Kernel-Adatron (KA), Cascade forward back propagation neural network [15-18].

A good classification consists of many ingredients in which the number of sample training data set is an essential part of the classification. All the research on MIT ECG dataset published have majority data is regular heartbeat, while we are interested in the abnormal heart rhythm. So examining the accuracy of the classification when removing the normal heart rhythm in the MIT data set is necessary. At the same time, we also design experiments to test the effects of the amount of training samples to the accuracy of the classification.

The paper is organized as follows: section 2 presents the method of implementation and the related theoretical basis, the post part 3 test results, and discussions. The final section presents the conclusions of the article.

II. METHODOLOGY

A. Proposed method

There are many methods of classification ECG, the following method is the simplest classification method based

on the proposed implementation of researches have been done recently. Block Diagram ECG classification includes three core areas: data preparation, extracted characteristic block and typical classification blocks as shown in Fig. 2. ECG after downloading from the available data is taken every heartbeat in the time domain, then the heart rate is converted through DWT domain to easily distinguish minor changes and feature extraction. Then, a PCA process is allied for dimensional reduction and feed to a neuron network for classification.

B. Principal Component Analysis

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables.

For a dataset S is given by,

$$S = \begin{bmatrix} S_1 & S_2 & \dots & S_n \end{bmatrix} = \begin{bmatrix} s_{11} & s_{12} & \dots & s_n \\ s_{21} & s_{22} & \dots & s_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m1} & s_{11} & \dots & s_{mn} \end{bmatrix} \quad (1)$$

The covariance value is defined as follows,

$$\mathrm{cov}(S_i, S_j) = \frac{\sum_{k=1}^{m} (s_{ki} - \overline{S_i})(s_{kj} - \overline{S_j})}{m-1} \quad (2)$$

in which, $i, j = 1, 2, \dots n$

Covariance matrix C is calculated according to the following formula,

$$C = \begin{bmatrix} \mathrm{cov}(S_1, S_1) & \mathrm{cov}(S_1, S_2) & \cdots & \mathrm{cov}(S_1, S_n) \\ \mathrm{cov}(S_2, S_1) & \mathrm{cov}(S_2, S_2) & \cdots & \mathrm{cov}(S_2, S_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(S_n, S_1) & \mathrm{cov}(S_n, S_2) & \cdots & \mathrm{cov}(S_n, S_n) \end{bmatrix} \quad (3)$$

The eigenvalues vector U of matrix C is given by,

$$CU = \lambda U \quad (4)$$

and

$$\lambda = \begin{bmatrix} \lambda_1 & \lambda_2 & \cdots & \lambda_n \end{bmatrix}^T \quad (5)$$

The data set will be restored from the main part S as,

$$PU^T = SUU^T = SUU^{-1} = S \quad (6)$$

in which, P (principal component) is the typical components of a dataset S.

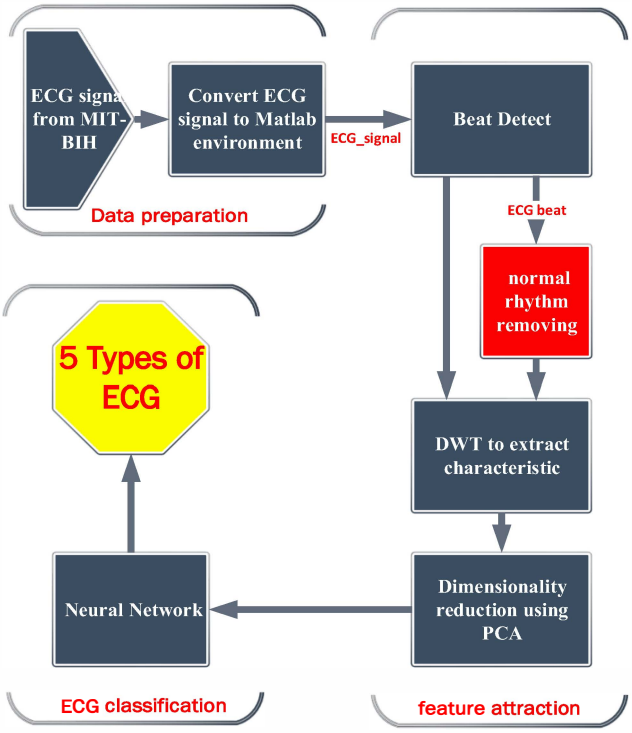

Figure 2. Block diagram of ECG processing

III. EXPERIMENTAL RESULT

To prove the effect of long tail phenomenon, the MIT-BIH arrhythmia dataset is used. The MIT-BIH arrhythmia database is well used in ECG classification researches [19], where the signals were sampled at 360 Hz. The database consists of 48 signals, each of thirty minutes duration of Holter recording. In this analysis, we have used the entire data of MIT-BIH arrhythmia database as recommended by ANSI/AAMI EC57:1998 standard.

Each ECG beat, which consists of 200 samples, is analyzed into four levels using FIR approximation of Mayer's wavelet ('dmey'). The approximate coefficient at the level-4 is included the frequency range from 0 Hz to 11.25 Hz, while detail coefficient at the level-4 is included the frequency range 11.25 Hz to 22.25Hz [20]. After decomposition using wavelet, approximate and detail coefficients were considered for subsequent dimensionality reduction by PCA method. The PCA method was applied on both coefficients of 4th level approximation and coefficients of 4th level detail independently. From each of the approximate and detail coefficients sub band, the first nine principal components were selected based on containment of 99.46% of the original data as shown in Fig. 4. In total eighteenth features, which consist of nine features from the approximate coefficient and nine features from the detail coefficient, were used for subsequent pattern recognition using neural network.

After reducing dimension, heart beat feature is fed into the classification as shown in Fig. 3. As described in [20], the model of classification is feed-forward neural network with the inputs layer consisting of eighteenth input nodes corresponding eighteenth features and one hidden layer including ten hidden nodes. The output layer of neural

network model has five output nodes or six output nodes to represent five or six ECG beat types, respectively. The numbers of output nodes of neural network model are five nodes while removing normal beat in ECG data and the numbers of output nodes of neural network model are six nodes while no removing normal beat in ECG data. The Fig. 3 shown that the neural network model has five output nodes in the output layer. In addition, the neural network weights are updated using the error back-propagation method. To stop neural network training, Mean Square Error (MSE) between the desired response and the actual response of the Neural Network is determined. The neural network weights are updated until the error value of the MSE achieves below 0.0001.

$$TPR = \frac{TP}{TP + FN} \tag{7}$$

$$TNR = \frac{TN}{TN + FP} \tag{8}$$

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{9}$$

To evaluate the system performance, this study uses true positive rate (TPR) and true negative rate (TNR) index as in (7-8). The definition of True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN), is defined as in Table I. A higher TPR and TNR mean that a system has a better performance. Furthermore, to easy identify the system performance, the accuracy (ACC) also calculates by formula as in (9). The accuracy refers to the correspondence between the class labels assigned to a heartbeat type and the true class. The higher accuracy system performance obtains, the better classifier is.

In this study, an experiment is designed which the percent of the train data is changing as Table II and III. In table II, we consider all normal beats and diagnose beats in to our classification. Because the radio of normal beats to diagnose beats are too high, the accuracy is increase slightly where the training sample significantly increase.

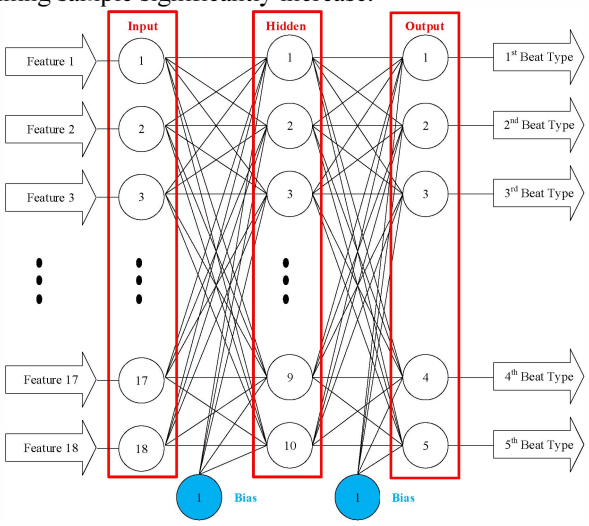

Figure 3. Neural Network Classifier with eighteenth input nodes in input layer, ten hidden nodes in hidden layer and five output nodes in output layer

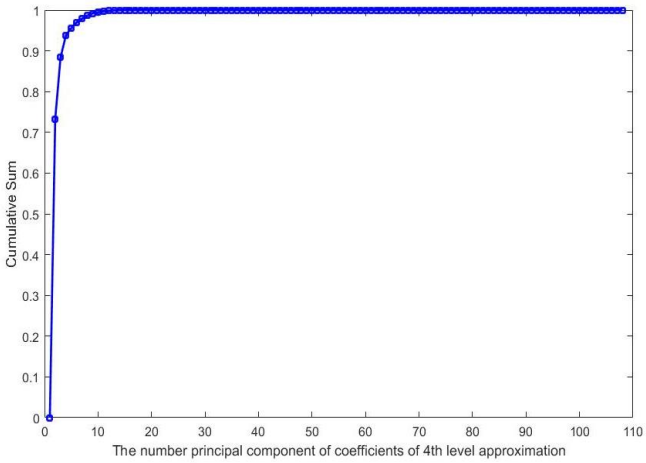

Figure 4. Cumsum of latent factor of the principal component from the coefficients of 4th level approcimation

TABLE I. CONFUSION MATRIX

		Predicted	
		Positives	***Negatives***
Actual	***Positives***	True positives **TP**	False negatives **FN**
	Negatives	False positives **FP**	True negatives **TN**

TABLE II. EXPERIMENTAL RESULT IN LONG TAIL SCENARIO (UNIT IS PERCENT)

Training	10	20	30	40	50	60	70	80	90
Testing	90	80	70	60	50	40	30	20	10
TPR	73,39	78,67	79,25	82,85	80,38	82,87	84,43	83,70	83,19
TNR	58,26	58,91	61,60	64,44	63,48	65,11	68,35	66,41	66,75
ACC	**93,08**	**93,25**	**93,72**	**94,26**	**93,96**	**94,38**	**94,82**	**94,68**	**94,60**

TABLE III. EXPERIMENTAL RESULT IN LONG TAIL SCENARIO (UNIT IS PERCENT)

Training	10	20	30	40	50	60	70	80	90
Testing	90	80	70	60	50	40	30	20	10
TPR	84,70	86,94	87,50	88,24	88,01	89,20	89,93	89,33	89,95
TNR	82,00	83,54	85,68	85,26	85,69	87,00	87,86	86,70	87,29
ACC	**87,97**	**89,67**	**90,52**	**90,79**	**90,96**	**91,79**	**92,34**	**91,77**	**92,18**

As shown in Fig. 4, the accuracy of classifier with normal beat including in ECG data is higher the accuracy with normal beat removing in ECG data. The average accuracy, which do not reject normal beat, is very high of 94,08%, and the average accuracy, which reject normal beat, is only 90,89%. In addition, the accuracy in situation 2 (removing normal beat) is fast increase when the number of training dataset is increase whereas the accuracy in situation 1 (including normal beat) is low increase when the number of training dataset is increase.

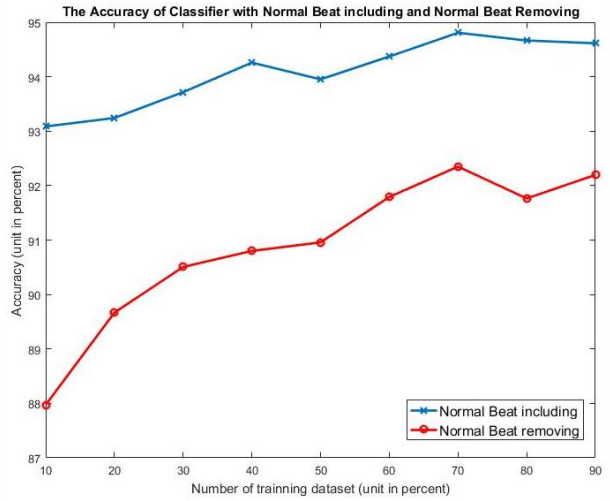

Figure 5. Comparison of the accuracy with long-tail and no long-tail classification

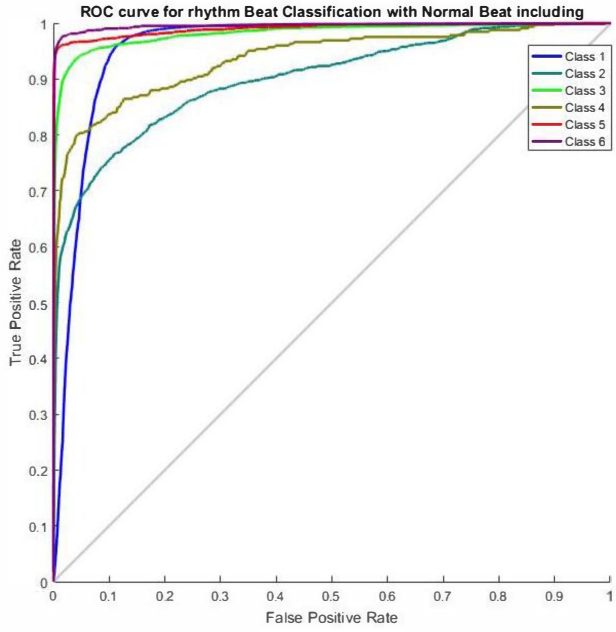

Figure 6. The ROC curves generated by six classes of the neural network classifier while containing of normal beat in ECG data

In case of confusion matrix for multi-class classification evaluating, the ROC curve is plot to show how to operate of classifier system. The curve is created by plotting the true positive rate against the false positive rate. Fig.6 and Fig. 7 present the ROC of neural network classifier in case of five and six output nodes. The Fig. 6 is shown performing of classification with five classes in the output layer. In contract, Fig. 7 is shown performing of classification with six classes in the output layer.

According to the result as shown in Fig .5, Fig. 6 and Fig. 7, the accuracy of classifier with including normal beat is higher the accuracy of classifier with removing normal beat. Furthermore, the ROC curve graphical as Fig. 6 and Fig. 7 illustrates the performance of classifier system in situation 1 to be better the performance of classifier system in situation 2.

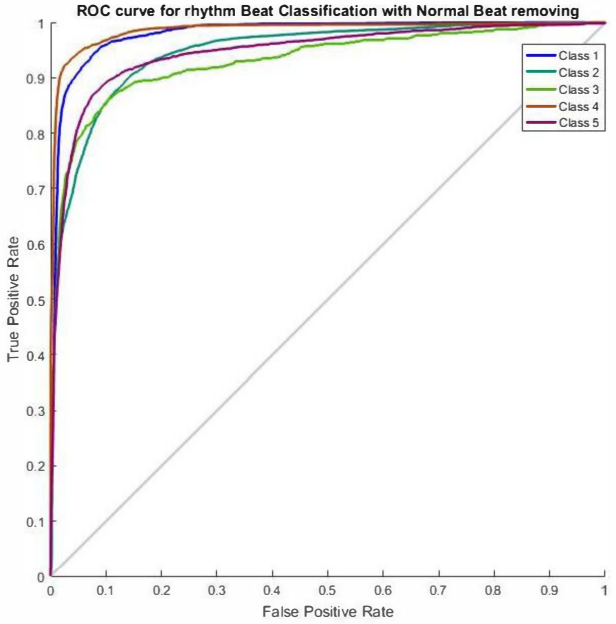

Figure 7. The ROC curves generated by five classes of the neural network classifier while containing of normal beat in ECG data

IV. CONCLUSION

In this paper, we use a state of art method on ECG classification to recognize heart diseases under long-tail effect phenomenon. The method use Wavelet transform to extract features; then PCA is used for dimensional reduction before apply a neuron network for classification task Because the number of normal beats is significantly higher than disease beats, it is difficult to recognize an anomaly beats. Hence, the performances of conventional methods have been degraded as the experimental result points out. Therefore, in particular applications where long-tail effect occurs, a suitable treatment should be applied to improve the performance.

REFERENCES

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

[2] D. Joshi and R. Ghongade, "Performance analysis of feature extraction schemes for ECG signal classification," *Int. J. of Elect., Electron. and Data Commun,* vol. 1, pp. 45-51, 2013.

[3] Z. Zidelmal, A. Amirou, D. O. Abdeslam, and J. Merckle, "ECG beat classification using a cost sensitive classifier," *Comput. methods and programs in biomedicine,* vol. 111, no. 3, pp. 570-577, 2013.

[4] C. Francisco, L. Pablo, S. Leif, B. Andreas, and R. J. Millet, "Principal Component Analysis in ECG Signal Processing," *EURASIP Journal on Advances in Signal Processing,* vol. 2007, no. 1, p. 074580, 2007.

[5] K. M. Gaikwad and M. S. Chavan, "Removal of high frequency noise from ECG signal using digital IIR butterworth filter," in *2014 IEEE Global Conference on Wireless Computing & Networking (GCWCN)*, 2014, pp. 121-124.

[6] H. Limaye and V. V. Deshmukh, "ECG Noise Sources and Various Noise Removal Techniques: A Survey," *International Journal of Application or Innovation in Engineering & Management (IJAIEM),* vol. 5, no. 2, pp. 86-92, 2016.

[7] M.Vijayavanan, V.Rathikarani, and D. P. Dhanalakshmi, "Automatic Classification of ECG Signal for Heart Disease Diagnosis using morphological features," *International Journal of Computer Science & Engineering Technology (IJCSET),* vol. 5, no. 4, pp. 449-455, 2014.

[8] D. Patra, M. K. Das, and S. Pradhan, "Integration of FCM, PCA and neural networks for classification of ECG arrhythmias," *IAENG Int. J. of Comput. Sci.,* vol. 36, no. 3, pp. 24-62, 2010.

[9] X. Tang and L. Shu, "Classification of Electrocardiogram Signals with RS and Quantum Neural Networks," *Int. J. of Multimedia and Ubiquitous Eng,* vol. 9, no. 2, pp. 363-372, 2014.

[10] L. Biel, O. Pettersson, L. Philipson, and P. Wide, "ECG analysis: A new approach in human identification," *IEEE Trans. Instrum. Meas.,* vol. 50, no. 3, pp. 808-812, 2001.

[11] V. Kumari and P. R. Kumar, "Cardiac Arrhythmia Prediction Using Improved Multilayer Perceptron Neural Network," *International Journal of Electronics, Communication & Instrumentation Engineering Research and Development (IJECIERD),* vol. 3, no. 4, pp. 73-80, 2013.

[12] V.K.Srivastava and D. D. Prasad, "Dwt - Based Feature Extraction from ecg Signal," *American Journal of Engineering Research (AJER),* vol. 2, no. 3, pp. 44-50, 2013.

[13] M. Korurek and Dogan, "ECG beat classification using particle swarm optimization and radial basis function neural network," *Expert syst. with Applicat.,* vol. 37, no. 12, pp. 7563-7569, 2010.

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

[15] M. Moavenian and H. Khorrami, "A qualitative comparison of artificial neural networks and support vector machines in ECG arrhythmias classification," *Expert Syst. with Applicat.,* vol. 37, no. 4, pp. 3088-3093, 2010.

[16] A. Khazaee, "Heart Beat Classification Using Particle Swarm Optimization," *Int. J. of Intelligent Syst. and Applicat. (IJISA),* vol. 5, no. 6, pp. 25-33, 2013.

[17] S. M. Jadhav, S. L. Nalbalwar, and A. A. Ghatol, "Artificial Neural Network Models based Cardiac Arrhythmia Disease Diagnosis from ECG Signal Data," *Int. J. of Comput. Applicat.,* vol. 44, no. 15, pp. 8-13, 2012.

[18] S. Ayub and J. P. Saini, "ECG classification and abnormality detection using cascade forward neural network," *International Journal of Engineering, Science and Technology,* vol. 3, no. 3, pp. 41-46, 2011.

[19] Physionet. (2014, November 10). *MIT-BIH Arhythmia Database*. Available: http://physionet.org/physiobank/database/mitdb/

[20] R. J. Martis, U. R. Acharya, and L. C. Min, "ECG beat classification using PCA, LDA, ICA and Discrete Wavelet Transform," *Biomedical Signal Processing and Control,* vol. 8, pp. 437-448, 2013.

BÀI BÁO KHOA HỌC

THỰC HIỆN CÔNG BỐ THEO QUY CHẾ ĐÀO TẠO THẠC SỸ

Bài báo khoa học của học viên

có xác nhận và đề xuất cho đăng của Giảng viên hướng dẫn

Thực hiện theo MTCL & KHTHMTCL Năm học 2018-2019 của Thư viện Trường Đại học Sư phạm Kỹ thuật Tp. Hồ Chí Minh.