AUTOMATIC ARRHYTHMIA DETECTION ALGORITHM USING STATISTICAL AND AUTOREGRESSIVE MODEL FEATURES

M.F.M Asri[1], N. Yahya[2] and I. Elamvazuthi[2]

[1]PETRONAS Carigali Sdn Bhd
[2]Department of Electrical and Electronic Engineering, Universiti Teknologi PETRONAS, Malaysia

Email : norashikin_yahya@utp.edu.my

ABSTRACT

Human heart healthiness is one of the major components to determine a person's overall healthiness. Automatic arrhythmia detection is important in a remote area where there is a lack of experienced cardiologists. In this work, an automatic arrhythmia detection algorithm is developed using statistical and autoregressive features to assist medical officers in the diagnosis of arrhythmia diseases. Basic statistical components namely mean, energy, standard deviation, mean absolute deviation, fractal, inter-quartile range and min/max value, were calculated. Alongside with statistical features, 10th order auto-regressive model parameters are used as input features to support vector machine (SVM). All features are calculated using an electrocardiogram (ECG) signals windowed into per beat manner. The proposed algorithm is able to classify normal ECG beat and five types of abnormal ECG beat; paced beat, right & left bundle branch block beat, premature ventricular contraction beat and aberrated atrial premature beat. By using SVM with quadratic and cubic kernel function, the proposed algorithm achieved the best accuracy of 95.8%.

Keywords: *ECG, cascade-SVM, AR model, statistical model, heart condition, computer-aided diagnosis*

INTRODUCTION

Arrhythmia is a condition where electrical impulses produce by human heart change from its normal sequence [1], [2]. This phenomenon causes the heartbeat to become too slow, too fast or irregular. As the function of the heart is to pump blood to other organs, it is nominated as the second most important organ in human anatomy after the human brain. Heart health condition is one of the most important health aspects that need to monitor from time to time. An excellent heart condition allows other organs in the human body to function perfectly.

Abnormal heartbeat can be detected in electrocardiography (ECG) signals by considering R-R interval, R-T interval, Rpeak and others. However, there are 17 types of abnormality associated with arrhythmia disease. Manual detection of arrhythmia abnormality required highly experienced cardiologists, but the availability of cardiologists is very limited, especially in a remote area. Therefore, automatic arrhythmia detection is essential in health care services so that patients can be treated immediately, especially in a remote area where scarcity of experienced cardiologists and doctors are the real issues.

In this work, the ECG dataset from MIT-BIH ECG Arrhythmia database [3] will be used. There are more than 15 arrhythmia abnormalities. However, in this work, only six classes to be used in the classification algorithm due to a lack of data for those particular classes. Statistical and autoregressive parameter are the two types of features extracted from windowed ECG signals. Basic statistical features such as energy, mean, standard deviation, min, max, mean absolute deviation, inter-quartile range and fractal. Moreover, the shape or envelope of ECG signals was taken into consideration by generating the 10th auto-regressive model, and its parameters are used as features for classification.

Automatic arrhythmia detection using ECG signals have been developed since the early 1960s. The ECG dataset was recorded using Del Mar Avionics model 445 2-channel reel-to-reel Holter recorders, which later are converted to digital format. In the digital format, the sampling rate of 360 Hz used to accommodate digital notch filter for the removal of 60 Hz interference. The ECG recordings recorded from 47 subjects between 1975 to 1979. The cardiologist labelled the different arrhythmia conditions, and the signals recorded for approximately 30 minutes per patient. ECG was recorded by placing electrode patches to the patient body to detect electrical activity of the heart (ECG). The position of the electrodes differs depending on the patient's anatomy and medical history [2], [4]. Although the device only reads signals from the human body, the patient may experience slight discomfort after removal of the electrode patches [2]. Annotation for MIT-BIH ECG database was done by two cardiologists, labelling heartbeat condition, discarding false signals, and if they cannot have an accord about a certain beat, the beats remain unclassified [4].

There are in total eight statistical features; mean, maximum, minimum, standard deviation, fractal, mean absolute deviation and interquartile range. Statistical features commonly used for signals analysis and image processing algorithm. A study was conducted in Liu et al. [5] for crowd counting, using mean and variance for statistical analysis of in-depth convolutional features. Other work also used statistical features to classify between normal and cardiac arrhythmia ECG signals [6]. Besides, the fractal dimension is commonly used as a feature vector. There are many other methods to calculated fractal. In this project, only the box-counting method is used due to its computational simplicity [7]. In other work, using R-R interval of ECG signals and respiratory signals, the fractal dimension can be calculated and use to classify different sleeping stages [8].

Auto-regressive (AR) model is a method to recreate or mimic ECG signal envelop or shape [9]. The 10th order AR parameters were extracted as a feature vector to be used in cascade-SVM classification. There are several methods in AR modelling such as least square, Burg and Yule-Walker and the least-square method was used in this work. In other work, vector auto-regressive was used to identify respiratory activity from ECG and PPG signals [10].

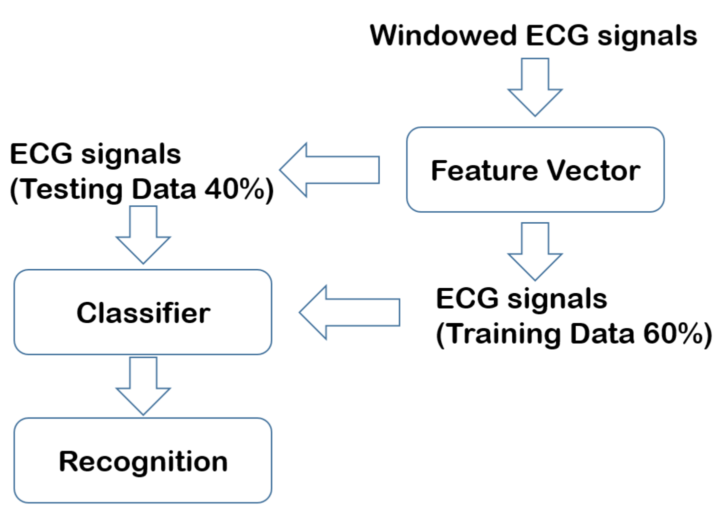

Figure 1 The proposed automatic arrhythmia detection algorithm

Proposed Automatic Arrhythmia Detection Algorithm

The block diagram shown in Figure 1 illustrates the proposed automatic arrhythmia detection algorithm. The ECG signals will first undergo a feature extraction stage calculated in fixed window size. ECG windowing is performed by considering the R-peak as the centre of the signal, and 70% samples from each side of the R-peak are included in the window. Windowed ECG signal is designed to have 20%overlapping with neighbouring signals. This will ensure each window covers ECG data for a single heartbeat.

Feature Extraction

Using windowed ECG signals, features are extracted, and the feature vector is created for every class of abnormalities. Two types of feature are extracted, statistical features and auto-regressive parameters. Statistical features include Mean, standard deviation, energy, min value, max value, inter-quartile range, mean absolute deviation and fractal. Auto-regressive model parameters are created for 10th order parameters, using the least-square method.

The QRS complex is the most crucial region in ECG signal indicating the flow of electrical impulses from atrial to ventricle for movement in the human heart in transporting blood. Normal patients generally have similar maximum R-peak values. Thus, deviation from its normal value may suggest that the patients may experience arrhythmia. On the other hand, Q-wave is the signal related to the minimum value and notably, the direction of the graph produced by Q-wave is opposite to the R-wave. The normal amplitude of Q-wave is typically one-fourth of the R-peak amplitude [11]. Using windowed ECG signals, the minimum and maximum value of the dataset were determined.

The detail on other features are explained as follows:

1. Mean (μ) is the average value of the signal and it is calculated by calculating the ratio of the summation of the sample to the number of samples. The amplitude of the ECG signal is in the order of millivolt (mV) unit and if the amplitude values vary significantly from its normal value, these changes will be reflected in its mean values.

2. Standard deviation (SD) gives an indication of data spread out. Low standard deviation means most of the data have values close to the mean indicating the three major waves P, QRS and T-wave to have the relatively same amplitude.

3. Mean absolute deviation (MAD) of a dataset is a measure of how far, on average, all values are from the mean. It is another way to describe variation in a dataset and similar to standard deviation, it gives a sense of how to spread out the values in a dataset.

4. Energy (En) of discrete signals is obtained by summing the square of the absolute value of samples over the signal duration. It can be used to characterize signal capture under different conditions. Another common statistical value to be extracted is the interquartile range which is the difference of value between 3rd and first quartile.

5. Fractal property of a signal can be determined using the box-counting method giving output known as a fractal dimension (FD). Other available fractal dimension analysis techniques include Higuchi's fractal dimension analysis and multifractal detrended fluctuation analysis. Box counting method was commonly used for fractal dimensions study because of its implementation simplicity [7]. In other work, with R-R interval of ECG signals and respiratory signals, the fractal dimension is used for classification of sleep stages [8].

6. Auto-regressive (AR) model is a method to recreate or mimic signals to a certain degree of accuracy. Any signals can be represented by higher-order differential equations and this allows the signal shape to be accounted in the feature vector. Using the autoregressive model, better ventricular late potential detection can be achieved [12]. There are several methods in AR modelling such as least square, Burg and Yule-Walker and in this work, the least-square AR model is used giving 80% accuracy to the real ECG data.

Classification of ECG Signals

In this work, the arrhythmia classification method will be tested using three classifiers; Support Vector Machine (SVM) with kernel function, k-nearest neighbours (KNN) and decision tree. For SVM, given a 2-class data, SVM will attempts to find the hyperplane that "best" separates two classes of points. The hyperplane will result in the largest margin between the two classes, and the points that lie on this margin are the support vectors. Not all data are linearly separable; hence, the used of SVM with kernel function will allow the input feature space to be mapped into higher feature space and allow the data to be separated with a nonlinear hyperplane.

On the other hand, the KNN algorithm works by forming a majority vote between the K most similar data to a given new data or observation. Similarity measure in KNN is a distance metric between two data points, and Euclidean distance is the commonly used ones. Other distance metrics include Manhattan, Chebyshev and Hamming distance which can be more suitable for a given setting.

Indecision tree learning, given several input variables, it uses a decision tree as a predictive model to make a conclusion about the target variable. In this work, bagged trees and boosted trees will have experimented since both typically give good classification accuracy.

Performance Metric for Classifier

Outcomes of the classifier can be denoted as true positive (TP), true negative (TN), false positive (FP) or false negative (FN). The accuracy or known as fraction correct measures the proportion of all correctly-categorized instances. It is represented by the ratio of the total correct classifications to the total amount of correct or incorrect classifications given as

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (1)$$

RESULT AND DISCUSSION

ECG Feature Extraction

Figure 2 shows the 3-sec recording of ECG signal from 2 electrodes, MLII and V5. The position of the electrode/patch depends on patient anatomy and medical history [3]. The R-R peaks were automatically detected for windowing the ECG signals into per beat manner.

Figure 3 shows sample values of the statistical features extraction from normal and abnormal ECG signals. The graphs indicate that the feature of specific classes varies between certain values. It is clear that some features show a distinctive pattern of specific classes but for some features, values of inter-class features are closely similar. The separability will be further investigated during the training and testing of classifiers.

To further investigate the significance of the 10 extracted features, SVM classifier with quadratic kernel function is tested with an increasing number of features. Initially selected features, namely SD, MAD and Min, are chosen based on the better separability of these features compared to other features. This deduction is made based on the results presented in Figure 3. Results in Table 1 show better classifier performance in terms of its accuracy as more features were added to the feature vector. In specific, going from 3 to 4 features resulted in a significant increase of 15% in the classifier accuracy. This indicates that the 4 features, SD, MAD, Min and Max are able to separate the classes well. However, the subsequent increase in feature vector had improved the accuracy by a relatively small margin between the range of 0.8% to 7.7%. Notably, the second-highest improvement of 7.7% is achieved when AR model coefficients are added to the feature vector.

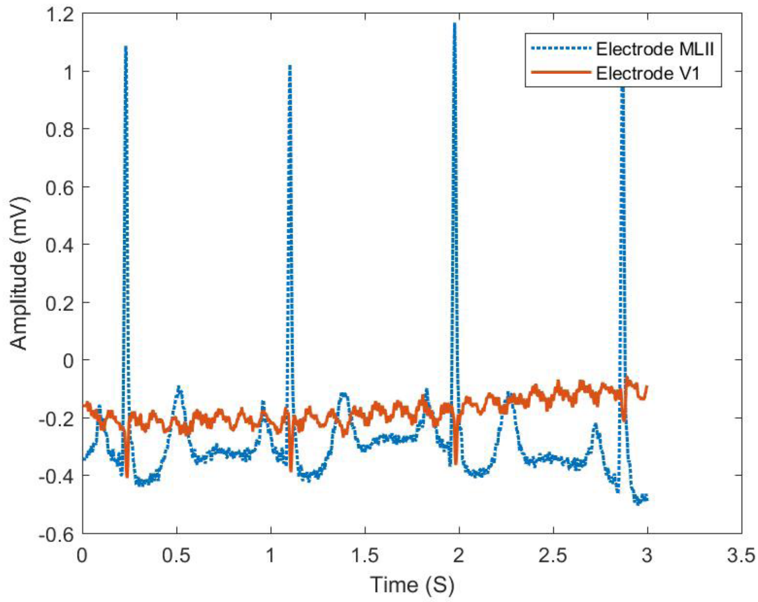

Figure 2 3-sec duration of 4-cycle ECG signal

Table 1 Percentage of Accuracy (Acc) For Increasing Number of Feature

SD	MAD	Min	Max	IQR	En	µ	FD	AR	ACC
✓	✓	✓							66.2
✓	✓	✓	✓						81.3
✓	✓	✓	✓	✓					85.8
✓	✓	✓	✓	✓	✓				86.6
✓	✓	✓	✓	✓	✓	✓			87.4
✓	✓	✓	✓	✓	✓	✓	✓	✓	88.1
✓	✓	✓	✓	✓	✓	✓	✓	✓	95.8

Table 2 Percentage of Accuracy of 6th Top-Ranked Classifiers

Classifier	*Accuracy*
Quadratic SVM	95.8
Cubic SVM	95.8
Boosted Tree	93.8
Bagged Tree	93.7
Weighted KNN	93.5
Cosine KNN	93.4

Arrhythmia Classification of ECG Signal

In this experiment, we input the feature vector to three types of classifiers and compare the performance in terms of classification accuracy. The three selected classifiers are SVM, decision tree and KNN classifier and the accuracy of the 6th top classifiers are tabulated in Table 2. The top classifier is SVM giving an accuracy of 95.8% is achieved with quadratic and cubic kernel functions. For the decision tree, the 2-best performance at 93.8% and 93.7% accuracy are obtained with boosted and bagged tree. Lastly, the KNN classifier resulted in two highest accuracy of 93.5% and 93.4% from weighted and cosine KNN.

classes. The second-best accuracy of 98% is achieved by normal beat with only 1% of normal beat class wrongly classified as the aberrated atrial premature beat. Next, the best performance is by left bundle branch block beat (97%), followed by right bundle branch block beat (95%). The two lowest performance is by premature ventricular contraction beat (93%) and aberrated atrial premature beat (92%). The overall average accuracy of the classifier is 95.8%. Improvement of the classifier can be made by further analysing the data and removing the outliers affecting the classifier performance. Besides, adding a feature vector may also improve the classifier as better class separation can be achieved.

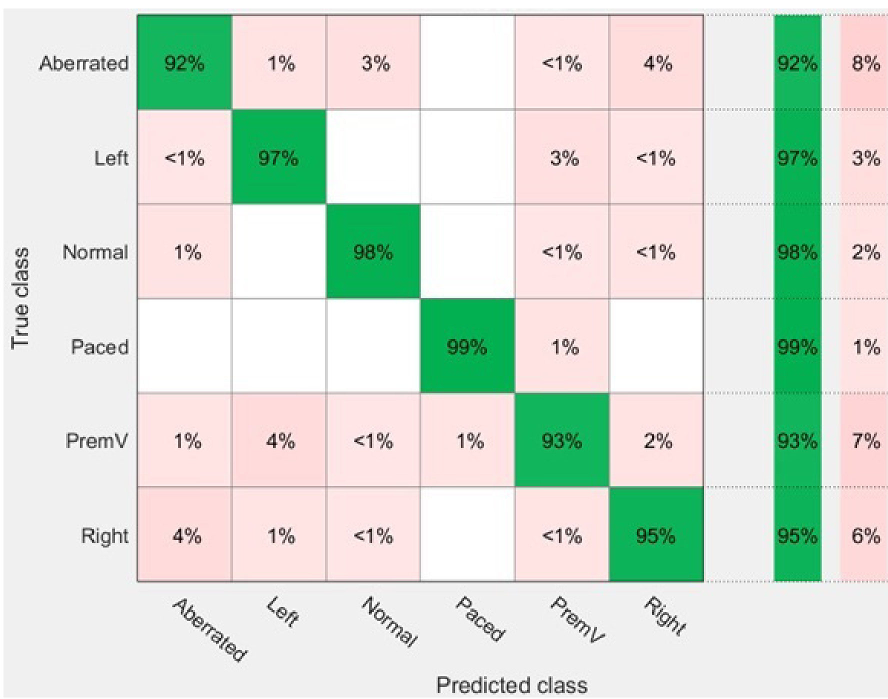

Figure 4 Confusion matrix for the proposed arrhythmia classification technique

We further investigate the performance of the proposed method via analysis of confusion matrix generated using quadratic SVM classifier, as shown in Figure 4. The highest true positive rate or accuracy of 99% is achieved by paced beat with only 1% of paced beat class wrongly classified as premature ventricular contraction beat. The highest accuracy is expected for the paced beat as its feature plot shown in Figure 3 has shown a clear distinctive pattern relative to other

It should be noted that it is highly critical for an automated medical diagnosis system to not wrongly classify abnormal cases as normal. The confusion matrix shown in Figure 4 indicates that the developed arrhythmia detection algorithm has wrongly classified 5% of abnormal cases as normal. This number needs to be further reduced either by removing outliers and adding more features using a technique such as based on wavelet decomposition.

Figure 3 Extracted features (a) maximum value, (b) minimum value, (c) mean, (d) standard deviation, (e) MAD, (f) IQR, (g) energy & (h) fractal of 30 samples normal ECG, (i) average value of 10th order AR coefficients for paced beat, right & left bundle branch block beat, premature ventricular contraction beat and aberrated atrial premature beat

CONCLUSIONS

Automated arrhythmia detection algorithm was developed for classification of ECG signals into six classes, normal ECG beat, paced beat, right & left bundle branch block beat, premature ventricular contraction beat and aberrated atrial premature beat. A total of 9 time-domain features are extracted and tested on three types of classifiers; SVM, decision tree and KNN classifier. The best performance was achieved with SVM with a quadratic or cubic kernel function. Investigation using eight statistical features gave an accuracy of 88.1% and adding feature from 10th order AR model parameters had significantly improved the accuracy to 95.8%. This result indicates that AR parameters carry the significant feature of ECG signals because of the ability of AR features in capturing signals shape. The accuracy of the proposed arrhythmia detection algorithm can be further improved by adding more feature based on wavelet decomposition.

ACKNOWLEDGEMENT

The authors would like to thank the Universiti Teknologi PETRONAS for providing financial support for this research work.

REFERENCES

[1] "Arrhythmia," [Online] https://www.nhlbi.nih.gov/health-topics/arrhythmia. Accessed on 30 May 2018.

[2] G. Walraven, *Basic Arrhythmias*, Seventh Edition, Prentice-Hall, 2010.

[3] R. Mark, P. Schluter, P. Devlin & D. Chernoff, "An annotated ECG database for evaluating arrhythmia detectors", *Frontiers of Engineering in Health Care, Proc. 4th Annu. Conf. IEEE EMBS*, pp. 205-210, 1982.

[4] G.B. Moody & R.G. Mark, "The impact of the mit-bih arrhythmia database," *IEEE Engineering in Medicine and Biology Magazine*, 20, 3, pp. 45–50, 2001.

[5] S. Liu, S. Zhai, C. Li & J. Tang, "An effective approach to crowd counting with CNN-based statistical features", in 2017 International Smart Cities Conference (ISC2), pp. 1-5, 2017. Doi: 10.1109/ISC2.2017.8090827

[6] M.M. Butt, M.U. Akram & S.A. Khan, "Classifying normal sinus rhythm and cardiac arrhythmias in ECG signals using statistical features in the temporal domain," in 2015 9th Asia Modelling Symposium (AMS), pp. 28-31, 2015. Doi:10.1109/AMS.2015.14

[7] R. Hippenstiel, H. El-Kishky & P. Radev, "On time-series analysis and signal classification - Part I: fractal dimensions," in Conference Record of the Thirty-Eighth Asilomar Conference on Signals, Systems and Computers, 2, pp. 2121-125, 2004.

[8] P. Castiglioni, A. Faini, G. Parati & C. Lombardi, "Fractal analysis of cardiorespiratory signals for sleep stage classification," in 2014 8th Conference of the European Study Group on Cardiovascular Oscillations (ESGCO), pp. 83-84, 2014.

[9] K. Juselius, *The Cointegrated VAR Model.* Oxford, UK: Oxford University Press, 2006.

[10] K.V. Madhav, M. Raghuram, E.H. Krishna, N.R. Komalla & K.A. Reddy, "Extraction of respiratory activity from ECG and ppg signals using vector autoregressive model", *2012 IEEE International Symposium on Medical Measurements and Applications Proceedings*, pp. 1-4, 2012.

[11] E. Braunwald, *Heart Disease: A Textbook of Cardiovascular Medicine*, Fifth Edition, Philadelphia: W.B. Saunders Co., 2006.

[12] R. Maniewski, M. Kohutnicki, P. Lewandowski, A. Xiebert & T. Palko, "Frequency analysis of high-resolution ecg using autoregression methods," in 1992 14th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2, Oct 1992, pp. 504-505.