Atrial Fibrillation Identification through ECG Signals

Atrial Fibrillation Identification through ECG Signals

Ng Joe Yee[1], Vikneswaran Vijean[1*], Saidatul Ardeenawatie Awang[1], Chong Yen Fook[1], Lim Chee Chin[1]

[1] Biosignal Processing Research Group (BioSIM), School of Mechatronic Engineering, Universiti Malaysia Perlis, Arau, Perlis, Malaysia

*vikneswaran@unimap.edu.my

Abstract. This paper presents an algorithm formulated to identify the atrial fibrillation complications through electrocardiogram (ECG) signals. The ECG data for the study was retrieved from Physio Net which consists of normal, atrial fibrillation and other rhythms. The Discrete Wavelet Transform (DWT) was used to remove baseline wanders. Pan Tompkins algorithm was utilized to detect the P, Q, R, S and T peak and thus the ECG signals were segmented based on each cycle. The morphological features were extracted directly from the time-series while statistical features were extracted after Stockwell transform (S- transform) was applied to the data. Genetic Algorithm (GA) and reliefF algorithm have been applied separately to select the optimum features for classification purpose. Bagged Tree ensemble algorithm, Decision Tree and k-Nearest Neighbour (KNN) algorithm were used as classifiers to identify atrial fibrillation through ECG signals. The classification results with and without feature selection techniques are presented. Prior to the feature selection, Bagged Tree is the classifier best performing classifier with 86.50% of accuracy, 84.38% of sensitivity and 91.94% of specificity. After feature selection, all the three classifiers have almost the same performance which is nearly 100% of accuracy, sensitivity and specificity. This shows that the proposed combinations of algorithms are reliable and able to improve the identification rate of the normal, atrial fibrillation and other rhythms using lesser number of features.

1. Introduction

The statistical report from the Ministry of Health Malaysia illustrated that the number of Malaysians suffering from cardiovascular disease has incremental trends. The number of patients with arrhythmia recorded as 2.8% in government hospital [1]. The incidences of atrial fibrillation have been reportedly increasing with age. The percentage of diagnosis related to atrial fibrillation disease within age range of 50 to 59 is 0.35%, whereas 2.32% is reported for those in the age range of 60 to 69. The age group of 70 to 79 has the highest percentage for atrial fibrillation disease which is 2.59% [1]. Therefore, the diagnosis of cardiovascular diseases with the use of computer aided tools for complications such as atrial fibrillations is a worthy challenge to be addressed. ECG signals play an important role in this identification process because it can effectively depict the condition of the heart. In this paper, an attempt have been made to classify the ECG signals into normal, atrial fibrillation and other rhythms using information retrieved from morphological features and statistical features of the data. MATLAB platform was used to execute the pre-processing, segmentation, feature extraction, statistical analysis, feature selection and classification steps. Each of the process are detailed in section 2.

2. Method

This section describes and discusses the flow of the research work as follows:

2.1. Data Source

The ECG signals used in this study were retrieved from the 2017 PhysioNet/Computing in Cardiology Challenge. The single short ECG lead recordings have a length in between 30 seconds and 60 seconds. Only the training set containing 8482 single lead ECG recordings labeled as normal, atrial fibrillation and other rhythms were used in the study as the testing data is inaccessible to the public. The ECG signal were sampled at 300Hz and bandpass filtered with a bandwidth of 0.5Hz to 40Hz [3].

2.2. Pre-processing

Pre-processing is an important step to remove the unwanted artefacts and noise in the signals.In this study, DWT was used to correct the baseline wandering which approximately 0.5Hz from the ECG signal [2]. Db5 wavelet was used to decompose the ECG signal into eight levels and the baseline artefact frequency was removed from the signals. Db wavelet family was chosen as it has a similar structure with the QRS complex and the energy spectrum is concentrated at low frequency. Since Figure 1 is the original ECG signal (A00001). Figure 2 is the ECG signal (A00001) after baseline correction.

2.3. Segmentation

Before segmentation is carried out, Pan Tompkins algorithm was utilised to detect P, Q, R, S and T peak. This algorithm will detect the peak based on the analysis of the slope, amplitude and width of ECG signals [4]. After P, Q, R, S and T peak are detected, the ECG signal were segmented based on each cycle. Figure 3 is the first segment of ECG Signal (A00001) with peak labelling.

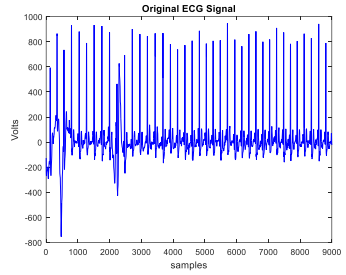

Figure 1. Original ECG Signal (A00001).

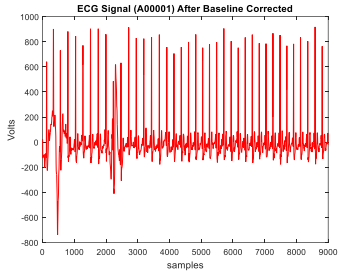

Figure 2. ECG Signal (A00001) after Baseline Correction.

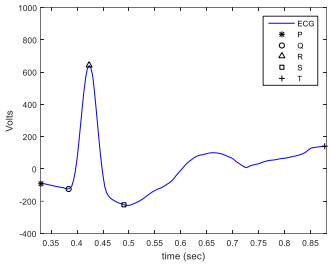

Figure 3. First Segment of ECG Signal (A00001) with Peak Labelling.

2.4. Feature Extraction

Feature extraction is a crucial step where the significant information from the ECG traces are extracted and used as a parameter to classify the ECG signals into normal, atrial fibrillation and other rhythms. The morphological features and statistical features from S-transformed data were used in the study.

The morphological features were extracted directly from the signals after the peak detection process. Features such as heart rate, PR interval, QRS complex interval, RT interval, average RR interval and root mean square of successive RR interval differences (RMSSD) were extracted in the process. Mean of RR interval, RMSSD and heart rate were calculated using Equation (1), (2) and (3).

$$\text{Mean of RR interval, } \bar{I} = \frac{1}{N-1}\sum_{n=2}^{N} I\ (n) \qquad (1)$$

$$\text{RMSSD} = \sqrt{\frac{1}{N-1}\left(\sum_{i-1}^{N-1}((R-R)_{i+1} - (R-R)_i)^2\right.} \quad (2)$$

$$\text{Heart Rate (beats per minute)} = \frac{Number\ of\ RR\ interval}{Totaltime\ (minute)} \quad (3)$$

Stockwell-transform was used to transform the time-domain ECG signals into time-frequency representation [5]. The features that were extracted from Time maximum Amplitude plot (TmA-plot) are the maximum and minimum value of TmA-plot, and energy for the maximum and minimum value of TmA-plot. From Time-Frequency contour (TF-contour), the average maximum frequency was extracted. Energy and power for maximum frequency were therefore calculated. The average skewness and kurtosis values of TF-contour were also measured. Skewness is a measure of the symmetry of the distribution while kurtosis corresponds to the degree of steepness of a probability distribution function. Skewness and kurtosis are defined in Equations 4 and 5.

$$\text{Skewness} = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_i-\bar{x})^3}{SD^3} \quad (4)$$

$$\text{Kurtosis} = \frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_i-\bar{x})^4}{SD^4}\text{-}3 \quad (5)$$

2.5. Feature Selection

Feature selection is an important process in order to obtain a smaller feature set so that a simpler and effective model can perform well in classification. In this study, the Genetic Algorithm (GA) and reliefF algorithm were implemented to select the features that can optimally classify the ECG signals into normal, atrial fibrillation and other rhythms. There were 16 morphological and statistical features extracted for the analysis.

2.5.1 Genetic algorithm

The GA was applied using a tournament-based selection strategy. The size of the chromosome was set as equal to the number of features. The purpose of applying GA was to look for a chromosome that represents a set of features which has the highest accuracy in classification. Initially, a random population is created, representing different points in the search space [6]. Each of the initial population was evaluated according to the fitness function [6]. In GA, the reproduction, crossover and mutation are the important steps to produce the next generation from the present generation. Through the repetition process, GA eliminates the low fitness chromosomes and keeps the high fitness chromosomes to the next generation until a good chromosome is obtained. The goal of genetic feature selection stage was to reduce the dimensionality of the feature before the classification process.

2.5.2 ReliefF algorithm

ReliefF algorithm is a filter-based feature selection method. It is a simple and effective algorithm which is able to evaluate the feature's importance by considering dependence from other features. The primary concept of this algorithm is to provide weight to every feature based on the performance of distinguishing the samples that pertain to the same group. A feature with high rank indicates that it can discriminate two near samples which are in the different groups. On the other hand, the feature with low rank expresses as it only can differentiate two near samples in the same group [7].

2.6. Classification

During this stage, Bagged Tree ensemble classifier, Decision Tree and k-Nearest Neighbour (KNN) were employed to classify the ECG signals as normal, atrial fibrillation and other rhythms.

2.6.1. Bagged Tree ensemble classifier

The principal of ensemble classifier is that a group of weak learners combined together to form a strong learner. In this study, Bagging (bootstrap aggregation) is the ensemble learning technique adopted in the classification process. For Bagged Tree ensemble classifier, it has combined several decision trees to produce a better predictive performance as compared to a single decision tree [8].

2.6.2. Decision Tree

Decision Tree is a non-parametric supervised learning method used for classification and regression. It is one of the selections for fast classification and which is also easy to analyze. In a Decision Tree, there are present with three types of nodes: root node, internal node and terminal nodes. A root node represents a single input variable and a split on that variable. Each terminal node contains an output variable which is used to make a prediction.

2.6.3. k-Nearest Neighbour (KNN)

KNN classifier is one of the most basic and non-parametric lazy learning algorithm. The classification is performed by finding the shortest distance from a data set which contains the input or training data and dataset which contains the reference values [9-11].

In this study, tenfold cross-validation was applied to the training and testing data. This indicates that the training dataset was divided into ten sets and the nine datasets will be used as training while the remaining one dataset will be used as testing. The process was repeated ten times such that each dataset has the chances to become a validation set [12]. Consequently, the mean accuracy for the three groups (normal, atrial fibrillation and other rhythms) was calculated and the standard deviation was determined. The mean sensitivity and specificity for the three groups were also investigated by exploring the confusion matrix of the classifier results.

3. Results and Discussion

The Bagged Tree ensemble classifier, Decision Tree and KNN algorithms were used to evaluate the classification accuracy of the proposed work using the selected features. The classification results acquired are presented and discussed in Tables 1, 2 and 3. There are two groups of classification results: (a) before feature selection and (b) after feature selection.

Table 1. Results of Bagged Tree before and after Feature Selection.

	No. of learners	Accuracy (%)	Sensitivity (%)	Specificity (%)
Before feature selection	60	86.50±0.44	84.38±8.45	91.94±7.05
After feature selection	30	99.95±0.03	99.95±0.09	99.96±0.06

From Table 1, it can be inferred that before feature selection, 60 learners were required to provide optimal classification accuracy of 86.50% ± 0.44%, sensitivity as 84.38% ± 8.45% and specificity as

91.94% ± 7.05%. After feature selection, 30 learners were able to provide 99.95% ± 0.03% accuracy, 99.95% ± 0.09% sensitivity and 99.96% ± 0.06% specificity. This demonstrates that the use of feature reduction technique was able to improve the classification results of Bagged tree while requiring lesser number of learners. This is because the dimension of the feature matrix has greatly reduced from 37080 x 16 to 37080 x 3 after applying GA and reliefF algorithm. Therefore, the smaller dimension of the feature matrix needed less number of learners to train the classifier. Through feature selection methods, the accuracy, sensitivity and specificity of Bagged Tree ensemble classifier in classifying normal, atrial fibrillation and other rhythms have improved because it has reduced the redundant feature set.

Table 2. Results of Decision Tree before and after Feature Selection.

	No. of splits	Accuracy (%)	Sensitivity (%)	Specificity (%)
Before feature selection	1400	79.75±1.05	77.93±9.13	88.31±8.10
After feature selection	1300	99.93±0.04	99.94±0.10	99.96±0.07

From Table 2, it can be observed that before feature selection, 1400 splits were required to produce 79.75% ± 1.05% accuracy, 77.93% ± 9.13% sensitivity and 88.31% ± 8.10% specificity. After feature selection, 1300 splits provided the accuracy of 99.93% ± 0.04%, sensitivity of 99.94% ± 0.10% and specificity of 99.96% ± 0.07%. The results of the Decision Tree have improved as compared between before and after feature selection. Although the dimension of the feature matrix had been reduced, the number of splits required in Decision Tree does not have much different. This indicates that the classifier still needed almost similar amount of splits for training purpose although improvement can be seen in overall classification results.

Table 3. Results of k-Nearest Neighbour before and after Feature Selection.

	k-value	Accuracy (%)	Sensitivity (%)	Specificity (%)
Before feature selection	13	46.82±0.61	35.75±27.97	68.12±26.92
After feature selection	4	99.95±0.04	99.95±0.10	99.96±0.07

From Table 3, it can be deduced that before feature selection, k-value of 13 provide the accuracy of 46.82% ± 0.61%, sensitivity of 35.75% ± 27.97% and specificity of 68.12% ± 26.92%. After feature selection, the k-value of 4 provided the accuracy of 99.95% ± 0.04%, sensitivity of 99.95% ± 0.10% and specificity of 99.96% ± 0.07%.The results of KNN have greatly improved as compared to before the application of feature selection process. This is because KNN is a non-parametric lazy learning algorithm. When too many inputs applied to it, the solution of KNN becomes too complex to handle and results in false identification. After the application of feature selection, the dimension of feature matrix was reduced. Therefore, the problems become simpler to be handled by KNN's nearest neighbourhood based majority voting method. For the desired k-value after feature selection, it is smaller than the desired k-value before feature selection. This is because when the dimension of the features matrix decreased, the k-value needed to set as small value so that the probability of getting error decreased.

4. Conclusion

Before feature selection, Bagged Tree with the number of learners equal to 60 has better performance as compare to Decision Tree and KNN. The accuracy, sensitivity and specificity of Bagged Tree are 86.50% ± 0.44%, 84.38% ± 8.45% and 91.94% ± 7.05% respectively prior to feature selection. After feature selection, Bagged Tree, Decision Tree and KNN have similar results which are almost 100% of accuracy, sensitivity and specificity. The results denote that feature selection is an important step in order to improve the performance of classifiers. In this study, GA and reliefF algorithm have been implemented to select the optimum features that able to differentiate between the three classes. This process has greatly reduced the redundant features and improved the performances of classifiers.

References

[1] W. Jin Shyan, R. Bau, F. Ahmad, R. Ismail, and A. F. Yean Yip, 2013, "Prevalence of asymptomatic atrial fibrillation in Malaysian patients with hypertension," *Med. J. Malaysia*, vol. 68, no. 2, pp. 141–143.

[2] K. S. Kumar, B. Yazdanpanah, and P. R. Kumar, 2015, "*Removal of noise from electrocardiogram using digital FIR and IIR filters with various methods*," 2015 Int. Conf. Commun. Signal Process. ICCSP 2015, pp. 157–162.

[3] G. D. Clifford et al. 2017, "AF Classification from a Short Single Lead ECG Recording: the PhysioNet/Computing in Cardiology Challenge 2017," Comput. Cardiol. (2010)., vol. 44, pp. 1–4.

[4] L. Sathyapriya, L. Murali, and T. Manigandan, 2014, "*Analysis and detection R-peak detection using Modified Pan-Tompkins algorithm,*" in 2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies, 2014, pp. 483–487.

[5] V. Vijean, M. Hariharan, and M. N. Mansor, 2016, "Analyzing selected visual anomaly through ST-based multi-resolution VEP decomposition," Proceeding - 2016 IEEE 12th Int. Colloq. Signal Process.its Appl. CSPA 2016, no. March, pp. 64–67.

[6] H. Li, D. Yuan, X. Ma, D. Cui, and L. Cao, 2017, "Genetic algorithm for the optimization of features and neural networks in ECG signals classification," *Sci. Rep*., vol. **7**, no. 1, p. 41011, Dec. 2017.

[7] F. G. Silva Teodoro, S. M. Peres, and C. A. M. Lima, 2017, "*Feature selection for biometric recognition based on electrocardiogram signals*," in 2017 International Joint Conference on Neural Networks (IJCNN), 2017, pp. 2911–2920.

[8] C. Paper, A. Mert, and B. Teknik, "*Ecg Signal Classification Using Ensemble Decicion Tree*," in Trends in the Development of Machinery and Associated Technology, 2016, no. September 2012, pp. 459–462.

[9] S. Jayalalith, D. Susan, S. Kumari, and B. Archana, 2014, "K-nearest Neighbour Method of Analysing the ECG Signal (To Find out the Different Disorders Related to Heart)," *J. Appl. Sci.*, vol. 14, no. 14, pp. 1628–1632, Dec. 2014.

[10] V. Vijean, M. Hariharan, and S. Yaacob, 2011, "Discrimination of vision impairments using single trial VEPs," *Proc. - 2011 IEEE Int. Conf. Control Syst. Comput. Eng. ICCSCE 2011*, pp. 182–186, 2011.

[11] Q. W. Oung, H. Muthusamy, S. N. Basah, H. Lee, and V. Vijean, 2018, "Empirical Wavelet Transform Based Features for Classification of Parkinson's Disease Severity," J. Med. Syst., vol. 42, no. 2.

[12] M. Hariharan, V. Vijean, C. Y. Fook, and S. Yaacob, 2012, "Speech stuttering assessment using sample entropy and Least Square Support Vector Machine," Proc. - 2012 IEEE 8th Int. Colloq. Signal Process. Its Appl. CSPA 2012, pp. 240–245.