Classifying Irregular ECG Samples using KNN

Inya Wannawijit and Thaweesak Yingthawornsuk

***Abstract*—This study has attempted to classify the ECG signals of heart diseases using the Hjorth Descriptors modified from previous works as class indicator. The comparative study has been made on several classifiers, which are Maximum-Likelihood (ML), Radial Basis Function Network (RBF), Support Vector Machine (SVM) and K-Nearest Neighbors (KNN). The extracted feature, which is the directional information from Electrocardiogram of subjects, was extracted from the preprocessed ECG waveform using an extended-version derivation of Hjorth descriptors. Three different ECG signal samples consisting of Normal Sinus Rhythm (NSR), Atrial Fibrillation (AF) and Congestive Heart Failure (CHF) were employed in signal processing, feature extraction and between-class classifications. The results show that the modified higher-order Hjorth Descriptors used in classification can provide the best performance based on the overall correct classification score found solely from the KNN classifier and effectively represent the functionality of heart under abnormality.**

***Keywords*— Hjorth Descriptors, ECG, KNN.**

I. Introduction

The heart rate of human adult can vary between 60 beats per minute (bpm) up to 100 bpm under normal condition [1]. In some condition the heart beat can be risen up to 150 bpm as a result from the rapid rhythmic discharge of impulses spreading out all over the heart in all directions. Cardiac Arrhythmia is the irregular functioning condition of heart in which the normal rhythm is disrupted, therefore heart rate can also be irregular

The Atrial Fibrillation (AF), found most often in people with ages over 60-65 years old, develops when a disturbance in the electrical signals causes two upper atrial chambers of the heart to quiver rather than pump the blood correctly. The most common symptoms of AF are palpitations or rapid thumping in chest, feeling tired or light headed, shorter breath and chest pain.

The Cognitive Heart Failure (CHF) is a condition when the heart can not pump to deliver oxygen rich blood to meet the body's need. CHF can be caused by weakening and stiffening of the heart muscles and also by increasing oxygen demand by the body tissue beyond the capacity of the heart can deliver. Most symptoms come with this heart condition are fatigue, swelling of ankles and legs or abdomen, shortness of breath, increased urination, nausea, abdominal pain and less appetite [2, 3].

The ECG signal has popularly been studied for diagnosis of heart diseases. It represents the functionality of heart in terms of bio-electric waveform that includes all electrical activities occurred in the heart originated from the area within our heart called Sinoatrial (SA) node. Therefore, the ECG signal will be studied via the computer algorithm developed based on Matlab description to pre-process, feature extract and classify samples from ECG classes.

This work aimed to determine a way to separate three different heart diseases by using the Hjorth descriptors as class indicating parameters in associated with classifiers for their accurate classification scores. In the past the Hjorth parameters were used to indicate the statistical time characteristics of the interested signal in processing firstly introduced by Bo Hjorth. All original related parameters called Activity, Mobility and Complexity were introduced in previous work of extraction from EEG signal [4] and other two extended-version descriptors were lately derived and proposed. Re-investigating the performance of the higher-order Hjorth Descriptors in classification mainly with KNN has been conducted and the accuracy rates as result from classification were also compared to those of formerly employed classifiers reported in prior works to reveal the improvement of performance evaluation made by KNN. Several different percentage numbers of training sample sets are similarly used as before in this work. The ways to classify among classes of ECG samples can be categorized into one-against-all and one-against-one, which are first a class of NSR against two combined classes of AF and CHF together and second an AF class against a CHF class, respectively.

The following sections are orderly organized: Related works is described in Section II. Section III is Methodology. Section IV shows the result of study and discussion. Finally, a conclusion is made in Section V.

II. Related Study

This research group [5] attempted to classify ECG signals by using Hjorth descriptors to represent the time-domain activities in QRS complexes of ECG signals. Three databases were extracted for Hjorth based samples to classification. The simplicity to estimate the descriptors was presented and result of classification showed that the Hjorth descriptors is very effective to identify all correct classes of the different ECG samples. The information related to emotional states in person extracted from EEG signals in a form of Hjorth descriptor estimation was proposed in [4, 6]. The classification has been made on the mental activity inherently present in EEG records. The highly adequate accuracy from classifying the mental task EEG signals via Hjorth parameters was found in study. From another former work [6], the automatic lung sound recognition used the Hjorth descriptors to characterize sounds recorded

Inya Wannawijit, King Mongkut's University of Technology Thonburi, Bangkok, Thailand.

Thaweesak Yingthawornsuk, King Mongkut's University of Technology Thonburi, Bangkok, Thailand.

from lungs in time domain and they are effective in classifying lung sounds. The other study [7] employed Hjorth parameters to identify the Cardiac Arrhythmia with Matlab Classification Toolbox. The contribution from this study provides the knowledge on development of clinical system that can detect the ECG Arrhythmia.

Finally, the lower- order Hjorth descriptors were used as a rapid way of extracting the distinctive characteristics of the ECG signal which contains the most relevant information of cardiac condition. The mean value of accuracy rates from multiple classification was found at 95%.

III. Methodology

A. ECG Database

The ECG database consists of three categorized groups of 30 Normal Sinus Rhythm (NRS) samples, 30 Atrial Fibrillation (AF) samples and 30 Congestive Heart Failure (CHF) samples. All ECG samples are collected from the MIT-BIH online database available at www.PhysioNet.org [8]. Each ECG sample was resampled with the same sampling frequency of 250 Hz and three complete QRS complexes from each ECG signal are segmented and further used in the following steps of preprocessing, feature extraction and performance classification.

B. Higher-Order Hjorth Descriptors

The first-order Hjorth descriptor is defined as a measure of signal variance in a particular order of signal variation. The Hjorth descriptor is favorably chosen for biomedical signal processing because of the simple computation and direct parameter calculation of a time-series as no signal transformation is required.

If x(n) is an ECG signal with n=0,1,2,…,N, therefore x'(n) is defined as the first-order variation derived directly from the signal;

$$x(n)' = x(n) - x(n-1) \tag{1}$$

and $x(n)''$ is defined as the second-order time derivative;

$$x(n)'' = x(n)' - x(n-1)' \tag{2}$$

Activity is the variance of a time-series. It can indicate the surface of power spectrum in the frequency domain, and signal power. If σ means a standard of deviation from $x(n)$, $\sigma_{x'}$ will be defined as the deviation standard of $x(n)'$ so that $\sigma_{x''}$ is deviation standard of $x(n)''$

$$\text{Activity} = \sigma_x{}^2 \tag{3}$$

Mobility is the mean frequency of proportion of standard deviation of the power spectrum. It is defined as the square root of the variance of the first derivative of signal, which is divided by the variance of signal.

$$\text{Mobility} = M_x = \frac{\sigma_{x'}}{\sigma_x} \tag{4}$$

Complexity is used to find change in frequency. It compares the signal's similarity to a pure sine wave, where the value converges to one if the signal is more similar.

$$\text{Complexity} = FF = \frac{M_{x'}}{M_x} = \frac{\sigma_{x''}/\sigma_{x'}}{\sigma_{x'}/\sigma_x} \tag{5}$$

Chaos is similar to complexity, yet it is more complicated. It is the only parameter, which is based on the expansion of Complexity. $x(n)'''$ is defined as the third-order time derivative so that $\sigma_{x'''}$ is standard deviation of $x(n)'''$.

$$x(n)''' = x(n)'' - x(n-1)'' \tag{6}$$

$$\text{Chaos} = KO = \frac{FF'}{FF} = \frac{M_{x''}/M_{x'}}{M_{x'}/M_x} = \frac{\sigma_{x'''}\cdot\sigma_{x'}{}^3}{\sigma_{x''}{}^3\cdot\sigma_x} \tag{7}$$

Hazard is algebraically defined as a ratio of Chaos. It is determined by the following derivative, that is based on the expansion of Chaos. The fourth-order time derivative is defined as $x(n)''''$. The standard deviation of $x(n)''''$ is $\sigma_{x''''}$.

$$x(n)'''' = x(n)''' - x(n-1)''' \tag{8}$$

$$\text{Hazard} = HZ = \frac{KO'}{KO} = \left(\frac{FF''/FF'}{FF'/FF}\right) = \left(\frac{\sigma_{x''''}\cdot\sigma_{x''}{}^3}{\sigma_{x'''}{}^3\cdot\sigma_{x'}}\right)\cdot\left(\frac{\sigma_{x''}{}^3\cdot\sigma_x}{\sigma_{x'''}\cdot\sigma_{x'}{}^3}\right) = \left(\frac{\sigma_{x''''}\cdot\sigma_{x''}{}^6\cdot\sigma_x}{\sigma_{x'''}{}^4\cdot\sigma_{x'}{}^4}\right) \tag{9}$$

By applying the parameter values obtained from the above steps to form the vector of Hjorth descriptors, the Hazard parameter combined with other four descriptors are then used to represent the ECG samples as input to classifier. A whole procedure to extract all Hjorth parameters by following the forementioned equations from (1) to (9) was implemented in Matlab descriptions. Each ECG sample from all three classes was processed step-by-step and all extracted parameters were stored in a form of matrices.

C. K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a supervised machine learning algorithm when target variable is known. It does not make an assumption about the underlying data as non-parametric distribution pattern.

K, in this algorithm, is a number used to identify similar neighbors for the new data point. The algorithm takes K nearest neighbors to decide where the new data point with belong to. This decision is based on feature similarity. Choice of K has a drastic impact on the obtained results from KNN.

There are 4 main steps to show how KNN works. Firstly, a value of K should be chosen as an odd number. Secondly, the distance of the new point to each of the training data must be found. Then, thirdly, it is to find the K nearest neighbors to the new data point. Finally, in case of classification, it is to count the number of data points in each category among the k neighbors.

To explain the distance between the new point and each of the training data, these are 4 main theories for the calculation such as Euclidean, Manhattan, Hamming and Minkowski.

Euclidean distance is the square root of the sum of squared distance between two points. It is also known as L2 norm.

$$\text{Euclidean} = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \quad (10)$$

Manhattan distance is the sum of the absolute values of the differences between two points.

$$\text{Mahhattan} = \sum_{i=1}^{k}|x_i - y_i| \quad (11)$$

Hamming distance is used for categorical variables. In simple terms it tells us if the two categorical variables are same or not.

$$\text{Hamming} = \sum_{i=1}^{k}|x_i - y_i|$$
$$x = y \rightarrow D = 0$$
$$x \neq y \rightarrow D = 1 \quad (12)$$

Minkowski distance is used to find the distance similarity between two points. When p=1, it becomes Manhattan distance and when p=2, it becomes Euclidean distance

$$\text{Minkowski} = \left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{\frac{1}{q}} \quad (13)$$

D. Cross-Validation and Performance Evaluation

All Hjorth parameters extracted from all categorized groups of ECG samples are used to determine the performance of classification. In former work, several classifiers, which are LS, ML, RBF and SVM, are employed to make comparison on accuracy rates found from training and testing each classifier with many scenarios of training sample sets. In this work, the similar way was also applied to KNN classifier with various numbers of sample datasets at 20%, 40%, 60% and 80%.

The K-fold Cross Validation (CV) technique are applied to both phases of training and testing classifier and at least 10 folds of CV are performed to each classification. The mean value of accuracy rates obtained from all folds is calculated and represented for a final accuracy rate of that classification. In each fold of classification, the accurate classification score can be directly determined from the True Positive and True Negative of the confusion matrix, which are the class-1 data correctly identified by classifier as the actual data of class-1 and alternatively the class-2 data correctly identified as the actual data of class-2.

All extracted Hjorth samples are arranged and classified in a pairwise manner between two sample classes against each other. First, we classify the extracted samples of NSR class against two combined classes of AF and CHF together in a form of "one-against-all" and secondly an AF class against a CHF class as "one-against-one", respectively.

IV. Results And Discussion

We estimate the magnitude spectra of the ECG signals from three different categories, which are manually segmented by editing with Matlab program and the first three complete QRS complexes from each ECG file are collected for feature extraction. In figure 1, the spectra estimated from ECG signals of AF and CHF have the higher energy concentrated mostly within a lower frequency range of 0-40 Hz, when compared to that of NSR within the same low frequency range. At the higher frequency range above 40Hz to the maximum frequency response of spectra at 125 Hz, the NSR has less energy than both AG and CHF.

In case of NSR, its spectrum has a very small magnitude when compared to those of AF and CHF. It is hard to make any observation on the quantitative comparison among three ECG classes on the original scale of magnitude, in terms of frequency bandwidth, decaying tendency of energy and location of dominant peaks among different types of ECG signal. Therefore, the normalization of spectral magnitude was then applied to have the estimated spectra of three categorized ECG signals aligned along the frequency axis so that energy distribution over frequency range can be compared.

The CHF has the highest rapidly declining magnitude while AF has less and slower decline of magnitude to its diminished magnitude level. As one can see from figure 2, more fluctuation of spiky peaks in spectrum of CHF can be notified within a low frequency range below 20Hz as compared to that of NSR which has very less fluctuation. In addition to the different magnitude fluctuations, the frequency bandwidths of AF and CHF spectra also highly differ in which the CHF has the narrowest frequency bandwidth among three ECG signal categories. This type of heart failure is an ongoing condition in which the heart's weaken muscle cannot pump the blood normally and also contract correctly. While that of NSR distributes over a wider frequency bandwidth with respective to its location of fundamental frequency.

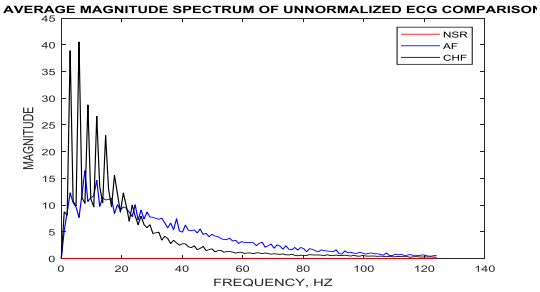

Fig. 1. Magnitude spectra of three categorized ECG signals

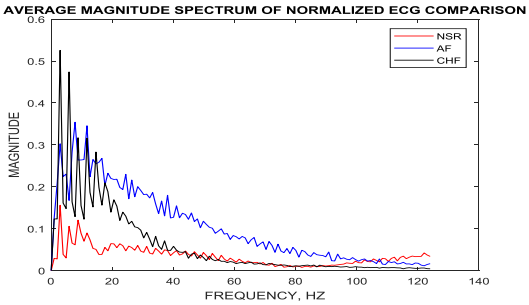

Fig. 2. Normalized magnitude spectra

In figures 3-4, the results of classification from classifying the NSR against both AF and CHF in combination with two different training datasets splitting at various percentages of the original Hjorth parameter dataset show very high accuracy rates when classifying with multiple Hjorth parameters in combination and including all higher-order Hjorth descriptor.

As depicted in plots of the correct classification scores corresponding to classifiers, comprising of LS, ML RBF and SVM, a column of NAB2345 is the mean value of accurate scores obtained from classifying between NSR class against AF class combining with CHF class and by using all higher -order Hjorth descriptors as classified sample input. This mean value indicats the highest satisfied accuracy rate in study almost nearly 100%. And it revealed that sizes of training sample sets do not affect the accuracies much in a case of all-against-one scenario. Overall accuracies of another scenario, one-against-one that is classification made between AF class and CHF class, are much more declined and inconsistent in multiple parameter-combined classifications when compared to the first scenario. The deterioration of accuracy rate in second scenario exists in that both AF and CHF classes similarly share the time-series amplitude fluctuation which may not be distinguishable each other because of a few QRS complexes selected in an early state of this study. The signal characteristics in terms of stationary property of ECG signal may in turn make the Hjorth descriptors not that highly effective in classification for the second scenario, but very effective in the first scenario.

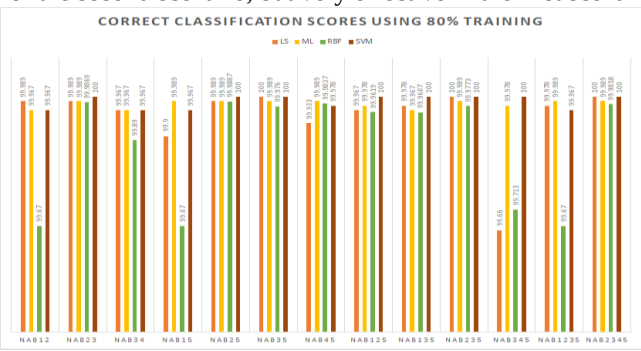

Fig. 3. Accuracy scores from classifying NSR from AF and CHF in combination with 80% training dataset of multiple Hjorths

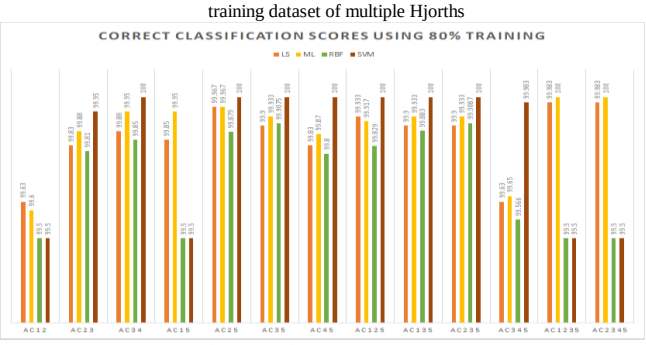

Fig. 4. Accuracy scores from classifying AF against CHF with 80% training dataset

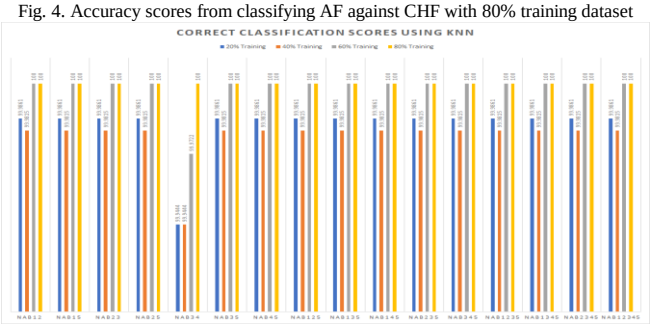

Fig. 5. Accuracy Scores from Classifying NSR from AF and CHF in Combination with 20%, 40%, 60% and 80% Training Dataset using KNN Classifier.

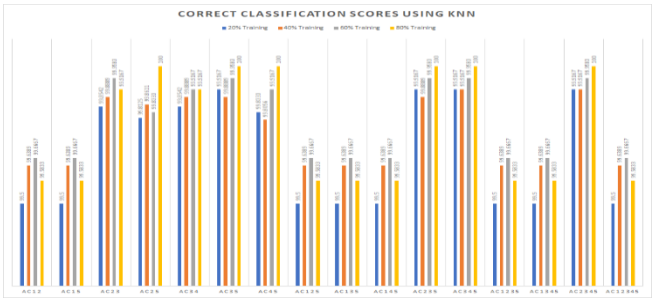

Fig. 6. Accuracy Scores from Classifying AF against CHF with 20%, 40%, 60% and 80% Training Dataset using KNN Classifier.

In figures 5-6, the high accuracy rates can be found when classifying AF against CHF with KNN using any combination of multiple Hjorth descriptors, except the combinations including the Activity parameter. With using the same KNN classifier, the result of classifying NSR from both AF and CHF shows that all combination of parameters provided high accuracy rates mostly associated with larger sizes of 60% and 80% sample sets used in classifications.

V. Conclusion

In this paper, as seen from the results of classifications with using multiple Hjorth parameters as sample input to KNN classifier, the highest accuracy of classification can reach up to 100% in most case of parameter combinations. Employing the Hjorth descriptors and the extended higher-order parameters in classification makes it successful due to the simple and direct derivation of feature and more convenience to implement the algorithm. It can be concluded that the higher order Hjorth descriptors are very effective in classification. In a future direction many steps in the continuing research study need re-consideration to make it more achieved in practical use and clinical validation.

Acknowledgment

This study is supported by Amritsar Pharmacy College of Engineering and Technology; for medical knowledge about Cardiology and Pathophysiology of Disease. The study, and Amritsar College of Engineering and Technology; about facility of accommodation and working environment, Amritsar, Punjab, India. In addition, the research work is financially supported by National Research Council of Thailand (NRCT) under research grant no. 621119 and also approved for conducting the research study under the Code of Human Research Ethics, KMUTT-IRB-2018/0605/231 shown in a certified document dated on August 7, 2018.

References

[1] C. Guyton, "Textbook of Medical Physiology" 8th Edition, Harcourt College Pub, October 1990.

[2] G. Singh Sugga, "Pathophysiology of Common Diseases", Jalandhar: Thakur Publishers, 2014.

[3] WebMD, "Congestive Heart Failure and Heart Disease" [Online]. Available at: https://www.webmd.com/heart-disease/guide-heartfailure#1.

[4] Hjorth, Bo; Elema-Schönander, AB, 1970. "EEG analysis based on time domain properties". Electroencephalography and Clinical Neurophysiology. 29: 306–310.

https://doi.org/10.1016/0013-4694(70)90143-4

[5] Rizal, and S. Hadiyoso, "ECG Signal Classification Using Hjorth Descriptor", 2015 International Conference on Automation, Cognitive Science, Optics, Micro Electro-Mechanical System and Information Technology (ICACOMIT), 2015, pp. 87 - 90. https://doi.org/10.1109/ICACOMIT.2015.7440181

[6] A. Patil, C. Deshmukh, and A.R. Panat, "Feature Extraction of EEG for Emotion Recognition using Hjorth Features and Higher Order Crossings", International Conference on Advances in Signal Processing (CASP), 2016, pp. 429 - 434. https://doi.org/10.1109/CASP.2016.7746209

[7] Ö. Tomak and T. Kayıkçıoğlu, "Comparison of Different Classification Methods in Arrhythmia Detection using Hjorth Descriptors", 24th Signal Processing and Communication Application Conference (SIU), 2016, pp. 1657 – 1660. https://doi.org/10.1109/SIU.2016.7496075

[8] Physionet.org, "ECG Database" [Online]. Available: http://physionet.org/physiobank/database/#ecg.