Using Independent Component Analysis to Obtain Feature Space for Reliable ECG Arrhythmia Classification

Mohammad Sarfraz

School of Computing, Science and Engineering
University of Salford, Salford, UK
Electrical Engineering, Salman bin Abdul Aziz University
Alkharj, Saudi Arabia
m.sarfraz@sau.edu.sa

Ateeq Ahmed Khan

Electrical Engineering, Salman bin Abdul Aziz University
Alkharj, Saudi Arabia
ateeq@ksu.edu.sa

Francis F. Li

School of Computing, Science and Engineering
University of Salford, Salford, UK
f.f.li@salford.ac.uk

Abstract— **Electrocardiogram (ECG) reflects the activities of the human heart and reveals hidden information on its structure and behaviour. The information is extracted to gain insights that assist explanation and identification of diverse pathological conditions. This was traditionally done by an expert through visual inspection of ECGs. The complexity and tediousness of this onus hinder long-term monitoring and rapid diagnosis, computerised and automated ECG signal processing are therefore sought after. In this paper an algorithm that uses independent component analysis (ICA) to improve the performance of ECG pattern recognition is proposed. The algorithm deploys the basis functions obtained via the ICA of typical ECG to extract ICA features of ECG signals for further pattern recognition, with the hypothesis that components of an ECG signal generated by different parts of the heart during normal and arrhythmic cardiac cycles might be independent. The features obtained via the ICA together with the R-R interval and QRS segment power are jointly used as the input to a machine learning classifier, an artificial neural network in this case. Results from training and validation of the MIT-BIH Arrhythmia database shows significantly improved performance in terms of recognition accuracy. This new method also allows for the reduction of the number of inputs to the classifier, simplifying the system and increasing the real-time performance. The paper presents the algorithm, discusses the principle algorithm and presents the validation results.**

Keywords— Electrocardiogram, independent component analysis, machine learning, pattern recognition, Back propagation neural network.

I. INTRODUCTION

Electrocardiogram (ECG) is a record of electrical activity generated by heart beats and measured from the surface of the body using special electrodes. It can be viewed, in a simpler term, as an electrical signature of heart behaviour. ECG signals are acquired by placing electrodes on the body surface at a number of prescribed locations and connecting the electrodes in different configurations to differential voltage amplifiers and a recorder. Many components of a measured ECG signal originated from different and hypothetically independent sources, the joint effect is a linear combination of them presented at the ECG electrodes. As a common clinical diagnostic tool, the ECG has extended its applications for long-term and on-site monitoring of many cardiac conditions and diseases. Prompt diagnosis, timely treatment and long term monitoring can prevent unexpected heart attack or other forms of heart failure. Automated ECG analysis is a typical signal processing and pattern recognition problem studied by many researchers and the developed systems used broadly in clinical settings over the past few decades.

The applications of Independent Component Analysis (ICA) and Blind Source Separation (BSS) to ECG and other biomedical signal processing are a relatively new but rapidly expanding area of study. The potential of blind signal separation to extract extra information about the heart and the body has been explored [1,2]. Independent component analysis has been found useful in revealing hidden factors of biomedical signals [3-7]. The method of using the ICA and BSS to separate motion induced noise was developed to improve the reliability of automated ECG recognition by the authors [8, 9].

The classification of ECG is a slightly different scenario. It intends to identify diverse cardiovascular conditions and diagnose associated diseases. Some algorithms are designed to classify different types of irregularities in heart beats, e.g. normal and abnormal heartbeats such as atrial fibrillation (AF), Premature Ventricular contraction (PVC) and tachycardia. Others are for the diagnosis of cardiovascular diseases such as myocardial infraction (MI), right and left bundle branch block. It is known that feature selection and suitable feature spaces play a vital role in the success of pattern recognition. This paper explores the use of ICA to acquire key features of

arrhythmia to enable more accurate and computationally efficient pattern recognition.

II. ECG BACKGROUND

The ECG is measured by placing electrodes on the body surface at various prescribed locations and connecting the electrodes in different configurations to voltage amplifiers and a recorder. As a basic configuration, the "three-lead ECG" is the most common among many other possible configurations. It is based on Einthoven triangle, [1] and uses only three leads for measurements to give a general profile of electrical activities of the heart.

As the electrode in left leg act as ground and the potential difference of other two electrodes are measured in reference with ground electrode hence only two of these three leads are independent. The two-lead ECG configuration is routinely used in the Holter monitoring that enables the recording of the heart activity of a subject continuously over a long time with portable devices [2]
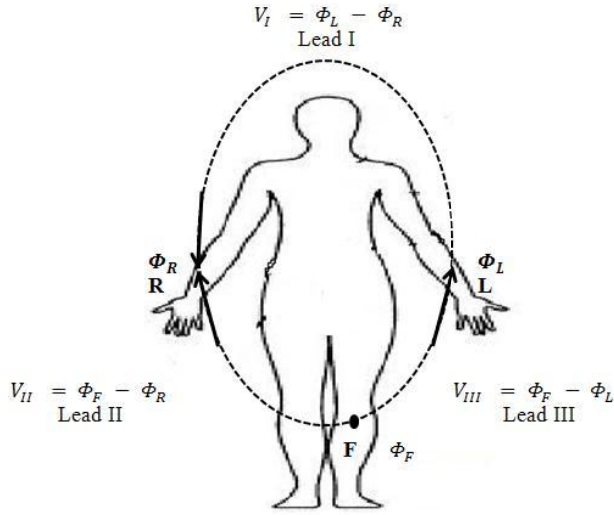

Fig. 1 Einthoven three limb leads method

We used data from lead II obtained from MIT-BIH [12,13]. The source of the ECGs included in the MIT-BIH Arrhythmia Database is a set of over 4000 long-term Holter recordings that were obtained by the Beth Israel Hospital. It is widely used as a benchmark or reference dataset in ECG analysis and pattern recognition work [3, 5, 14-16].

The ECG is described by waves, interval and segments.
- Waves are labelled using the letters P, QRS, T and U.
- Segments are defined as the time duration between different waves, for example S-T segment is the duration between S and T waves, this segment denotes the entire ventricular depolarization.
- The interval on the time durations that includes waves and segments for example Q-T interval

represents the time for ventricular depolarization and repolarization.
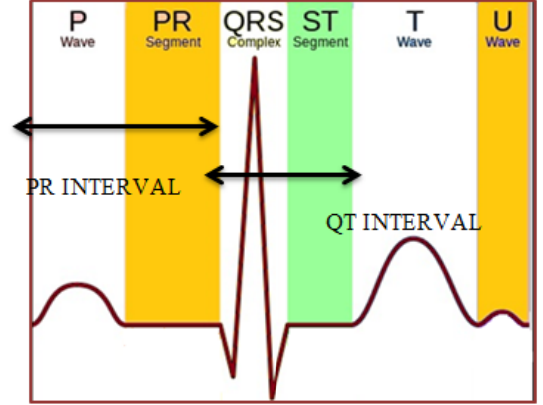

Figure 2 Basic ECG Wave Shape with morphological features

III. INDEPENDENT COMPONENT ANALYSIS

Independent component analysis is a statistical method that is used to identify underlying factors or components that are statistically independent [10].The attractive feature of the ICA is that it is able to statistically separate out individual sources from their mixtures without prior information about the sources and the mixing parameters. Cocktail party problem as illustrated in Fig 2 is a classic example to explain the ICA based BSS. Two talkers are heard by two listeners (or picked up by two microphones), due to distance discrepancy, each listener hears a different linear mixture of the two sources.
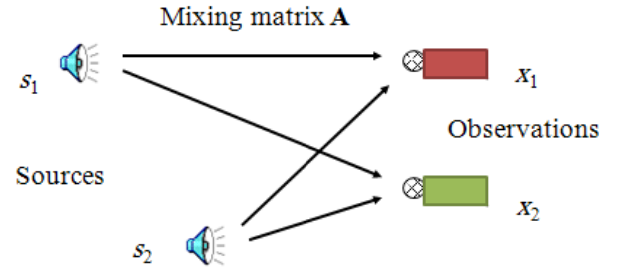

Fig. 3 Cocktail party problem

$$x(t) = As(t) \qquad (1)$$

where vector $x(t)$ represents the signals received, $s(t)$ the original source signals, and A the mixing matrix. The objective is to separate the individual voices from the linear mixture. Given the conditions of same number of sources as there are receivers.

$$x_1(t) = a_{11} s_1(t) + a_{12} s_2(t) \qquad (2)$$

$$x_2(t) = a_{21} s_1(t) + a_{22} s_2(t) \qquad (3)$$

where x_1 and x_2 are the signals received by the microphone, the presence of sources s_1 and s_2 in x_1 and x_2 is determined by the mixing parameters a_{11}, a_{12}, a_{21} and a_{22}. We have to separate individual sources from the mixture, with no information about the sources and mixing parameters available.

The ICA is usually performed by formulating an objective function, e.g. mutual information or negentropy, and then minimizing or sometimes maximizing it. This transforms the ICA problem to a numerical optimization problem [11]. There are a number of algorithms for the ICA. In this study, a fixed-point algorithm was adopted to estimate the independent components [12]. FastICA is one of the popular ICA techniques of this type [13]and is used by numerous researchers for various purposes. . FastICA attempts to separate underlying sources from the given measurement set based on their 'non-Gaussianity'. The basic principle behind the FastICA is that the fast fixed-point iterative algorithm undertakes to find projections that maximize the non-Gaussianity of components by their kurtosis (the fourth-order cumulant given to a random variable). In simpler terms, as kurtosis is identically zero for Gaussian distributed signals, the objective is to maximize the magnitude of the kurtosis to make the estimated sources as non-Gaussian (i.e. as independent) as possible. The kurtosis that is used to describe the peakedness of a distribution is given by

$$kurt(x) = E\{x^4\} - 3(E\{x^4\})^2 \qquad (4)$$

For a zero-mean random variable x. Further details about the FastICA algorithm can be found in [4, 20]. In the proposed algorithm the ICA is used in an unusual way. Instead of performing blind source separation, it is used to determine the underlying features or independent components (ICs) of different ECG arrhythmias that are statistically independent to one another. Thereby the ICs are selected according to their significance and mutual independence with each other to form a feature set along with other morphological and time domain features to be used as input to classifier.

IV. PROPOSED METHOD

The basic ICA model is that a vector of unknown sources s is not observed directly, but from a linear combination of them.

$$x = As \qquad (5)$$

Where x is a N x 1 column vector, A is an N x N mixing matrix, and s is a N x 1 column vector of the source signals. The columns of A, $\{a_i\}$ are known as the basis function which generate the observed signals. The inverse of mixing matrix A is W, which transforms the original signals into the unknown source coefficients.

$$s = Wx \qquad (6)$$

.
The objective of the ICA algorithm is to find the basis function by adapting and learning from the ECG data x in this study. The cost function in ICA can be minimization of mutual information of de-mixing model output as used in Infomax algorithm [15] or the maximization of non-Gaussianity used in FastICA algorithms [16]. Using ICA as a feature extraction method, the basis functions a_i, are considered as the basis features and the component of s, s_i is considered to be the coefficient for each basis feature in the feature space. So the model can be written as

$$x = \sum_{i=1}^{N} s_i a_i \qquad (7)$$

As compared to correlation-based transformations such as principal component analysis (PCA), the ICA not only uses second-order statistics to decorrelate the signals, but also uses high-order statistics to reduce high-order dependencies. Thus the output coefficients corresponding to different basis functions are as statistically independent as possible.
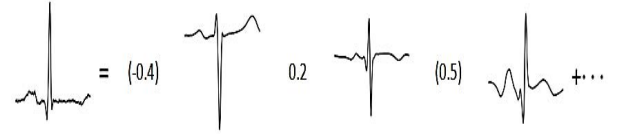

Fig. 4. A normal heartbeat is a linear combination of ICA bases with Coefficients, e.g.(-0.4)·a_i+(0.2)·a_j+0.5·a_k+...

Fig. 4 demonstrates the linear decomposition of ECG signals as in Eq. 8. The segment of a normal ECG signal with single heartbeat can be represented as a linear superposition of the ICA basis functions $\{a_i, a_j, a_k, ...\}$ scaled by the corresponding coefficients $\{s_i, s_j, s_k, ...\}$.The heartbeat segment is represented mostly by$\{a_i, a_j, a_k\}$, as most of the other coefficients are almost zero. The ICA basis functions reveal the statistical structures of the single heartbeat segments of ECG.

The MIT-BIH Arrhythmia Database contains 48 half-hour excerpts of two-channel ambulatory ECG recordings, obtained from 47 subjects studied by the BIH Arrhythmia Laboratory between 1975 and 1979. Twenty-three recordings were chosen at random from a set of 4000 24-hour ambulatory ECG recordings collected from a mixed population of inpatients and outpatients at Boston's Beth Israel Hospital; the remaining 25 recordings were selected from the same set to include less common but clinically significant arrhythmias that would not be well-represented in a small random sample [3].

Since the MIT-BIH data set is completely annotated we know the occurrence of R peak and most of the diagnostic information of the ECG signals. We selected portion on both sides of R peak. Two hundred data points were taken in each sample, which is about 0.556s in terms of the sampling interval. The sampling frequency of this signal is 360 Hz. The extracted sample of 200 data point has all the required information of ECG pulse including P and T wave as well, which give us complete information contained in single pulse including the cases of noise presence along with the original signal.

To estimate the ICs as bases (ICA bases), we randomly selected two samples from each of the records used in the eight ECG beat types, resulting in total 100 samples, each of which has 200 data points which constitute the observed data matrix *[x]*. For 8 types of arrhythmia, randomly selecting 2 ECG samples from each record of the database gives a data matrix of 100x200 dimension.

Records and number of ECG samples used in this study are summarized in Table 1. The ECG segments of the records in the database were divided samples into two sets. One set for training and the other for testing..

TABLE 1 RECORDS AND NUMBER OF ECG SAMPLES USED

Type	MIT-BIH data, file reference	Training (set)	Testing (set)
NORM	100,101,103,105,108 100 100 112,113,114,115,117 100 100 121,122,123,202,205 100 100 219,230,234	4800	4800
LBBB	109,111,207,214	800	800
RBBB	118,124,212,231	800	800
PVC	106,119,200,203,208, 213,221,228,233,116,201,210,215	2600	2600
APB	209,222,232,220,223	1000	1000
PB	102,104,107,217	400	400
VFW	207	200	200
VEB	207	50	50
Total		10650	10650

We randomly selected the ECG samples to obtain the unmixing matrix *W* using FastICA, and then re-order the row of the *W* matrix by the norm of the row from the smallest to the largest. After the calculation of independent components (ICs) .They were employed as bases.

In both training and testing stage, each sample of the training set is projected onto the ICs. The projections constitute the ICA-feature vector. Along with the RR interval, and QRS signal power the feature vector is obtained. Two features were extracted from the RR sequence. The pre-RR-interval was the RR interval between a given heartbeat and the previous heartbeat. The post-RR-interval was the RR-interval between a given heartbeat and the following heartbeat. Both of them were calculated using pan Tompkins algorithm for QRS detection [17].

We used BPNN as classifier in the training stage. The input is feature vectors and the output value can be set at 1, 2, 3, or 8 for NORM, LBBB, RBBB, or PVC (8 types of ECG). The aim of this stage is for adjusting the best parameters in neural networks for the best classification. To study the effect of the Numbers of ICs in ECG beat classification, the numbers of ICs were varied from 5 to 40 and their effects were investigated. We achieved the best performance of classifier with 15 ICs. 10 fold approach was used to select number if

ICs. Further increasing the number of ICs from 15 did not improve classification accuracy regularly. Even though some improvement was recorded but it was neglected due to non-uniformity. After the neural network is trained, we applied the other set of ECG samples for testing the performance of the classifier and calculating the specificity and sensitivity.
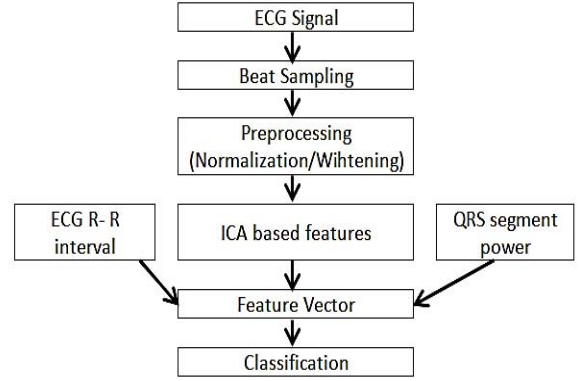

Fig. 5. Block diagram of the proposed feature extraction and classification system

Classification was done using Back Propagation neural network (BPNN) implemented in MATLAB software. ANN are widely used classifier for ECG [7], [9], [18]–[20]. Back-propagation neural network (BPNN) used in this study which is a three-layer feed-forward structure [21] The first layer is the input layer that has the ICA features , pre-RR-interval, post-RR-interval and QRS segment power as inputs.

18 neuron were used in input layer. In this study 15 ICs were selected to be part of feature vector. The second layer, also called the hidden layer, has 20 neurons and the output layer has 8 neurons for classification, which is equal to the number of ECG beat types to be classified. In this study, the hyperbolic tangent functions are used in the first and second layers, and the identity function is used in the output layer. The weight and bias values in the BPNN are updated by Levenberg-Marquardt optimization method [22] with a learning rate of 0.1. A criterion of 0.01 in mean-square-error is empirically determined to terminate the iterations in the training phase of the classifier. Time taken for the training of classification was about 1.2 seconds in the Matlab computing environment based on the average over 10 times.

V. RESULTS & DISCUSSION

A total of eight ECG beat types were selected from the MIT-BIH arrhythmia database for experiments. The eight beat types employed in the study were normal beat (NORM), left bundle branch block beat (LBBB), right bundle branch block beat (RBBB), artial premature beat (APB), premature ventricular contraction (PVC), paced beat (PB), ventricular flutter wave (VFW), and ventricular escape beat (VEB). The accuracy, specificity and sensitivies of the neural network classifier are summarized in Fig. 6. Classification Performance into eight

classes with proposed method best accuracy of 99.8 % and overall accuracy of 99.65 % were achieved, which is better than any existing methods discussed in the literature. In this study, we used the 10-fold cross validation technique for training and testing of the classifiers as in Fig. 7. Results of Arrhythmia Classification Accuracy with proposed method. The correct classification or misclassification was assessed as True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). Based on these measures the specificity, and accuracy were determined as shown in Fig. 8. Results of specificity with proposed method shows the sensitivity of different classifiers during each fold of the classification using ICs. Matlab software was used for the signal processing, pattern recognition, visualization and user Interface in this study. The proposed methodology in this study can be used in arrhythmia monitoring and identification systems, telemedicine applications, holter device ECG monitoring, and general screening for cardiac health.

It is also interesting to compare the result of our scheme for ECG classification with that of the other methods recently reported. Methods from seven representative studies were chosen for this comparison, including a patient-adaptable classifier using time-domain features and a mixture of experts (MOE) [23]. ICA feature extraction method joining ICA basis function coefficient and wavelet transform coefficient forming an over complete feature vector by [24]. Redundancy and relevance optimization for feature selection criteria and support vector machine for classification with 14 heart beat types proposed by [25]. [26] Used multi-layer perceptron to classify ten arrhythmias with training accuracy of 78% using discrete Fourier transform for feature extraction. ECG classification using discrete wavelet transform along with multilayer perceptron getting 97.98% training accuracy with five classes of arrhythmia by [27]. Llinear discriminant analysis was used on segmented wave shape features and fixed interval wave shape features by [28]. ICA projection along with R-R interval was used with probabilistic neural network (PNN) for ECG classification by [6]. Characterization of ECG based on Fourier-domain BSS (BSS-Fourier) was done by [29]. intelligent diagnosis system using hybrid approach of adaptive neuro-fuzzy inference system (ANFIS) model for classification of Electrocardiogram (ECG) signals proposed by [30]. Table 2 summarizes the comparative results of these methods. Among the seven methods, the proposed method outperforms the other methods with an impressive accuracy of 99.61% to discriminate eight ECG beat types. Even though this comparison may not be completely rational because of the different numbers and types of ECG beats tested, the propose method reveals to be a powerful tool in the computer-aided diagnosis of heart diseases based on ECG.

TABLE 2 COMPARATIVE RESULTS OF DIFFERENT ECG BEAT CLASSIFICATION METHODS

Work Reference	Number of beat types	Accuracy (%)
Proposed	8	99.6
Sung-Nien Yu	8	98.7
Philip de Chazal	5	90.3
T. M. Nazmy	6	97.1
Xing Jiang	14	98.6
Yu Hen Hu	4	94.0
İ. Güler	5	97.8

Fig. 6. Classification Performance into eight classes with proposed method
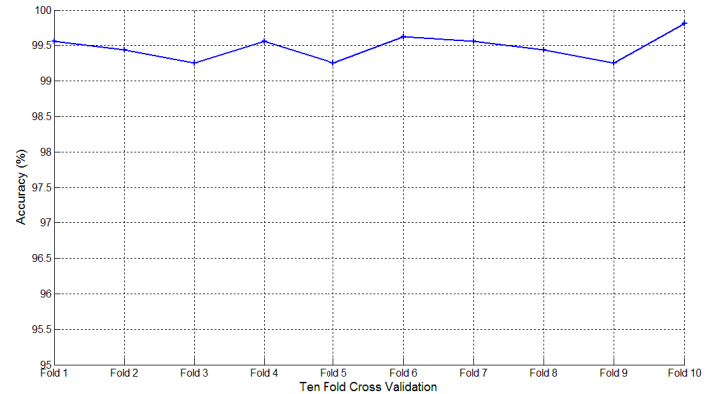

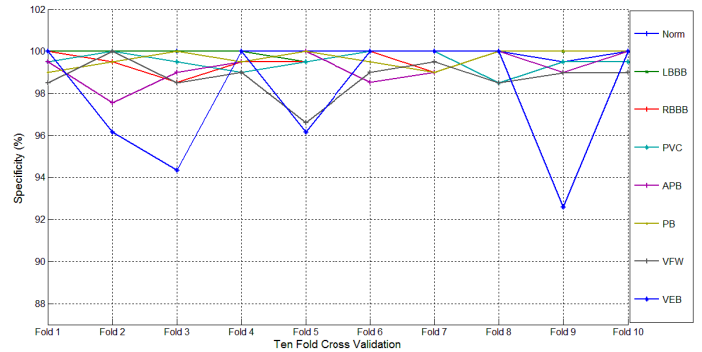Fig. 7. Results of Arrhythmia Classification Accuracy with proposed method

Fig. 8. Results of specificity with proposed method

VI. Conclusion

This paper has proposed and validated a method and algorithms for ECG arrhythmia classification using features extracted with ICA. A new feature set has been developed; it includes morphological ECG features along with ICA extracted features. The classification experiments were performed on the MIT-BIH Arrhythmia Database which includes the normal sinus beat and 8 types of arrhythmia beats. Computer simulations show that ICA based feature extraction method outperforms any other reported methods. It also shows that selecting relevant features of the feature set can improve the recognition performance. This makes a robust feature set for ECG arrhythmia classification. Comparison of classification accuracy, specificity and sensitivity with other state-of-the-art method is done and its efficiency is clearly demonstrated.

References

[1] J. Malmivuo and R. Plonsey, *Bioelectromagnetism: principles and applications of bioelectric and biomagnetic fields*. Oxford University Press, 1995.

[2] C. Ye, M. T. Coimbra, and B. V. Kumar, "Investigation of human identification using two-lead electrocardiogram (ECG) signals," in *Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on*, 2010, pp. 1–8.

[3] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[4] G. B. Moody, W. Muldrow, and R. G. Mark, "A noise stress test for arrhythmia detectors," *Comput. Cardiol.*, vol. 11, no. 3, pp. 381–384, 1984.

[5] N. Belgacem, M. A. Chikh, and F. B. Reguig, "Supervised classification of ECG using neural networks," 2003.

[6] K.-T. Chou and S.-N. Yu, "Categorizing Heartbeats by Independent Component Analysis and Support Vector Machines," in *Intelligent Systems Design and Applications, 2008. ISDA'08. Eighth International Conference on*, 2008, vol. 1, pp. 599–602.

[7] X. Jiang, L. Zhang, Q. Zhao, and S. Albayrak, "ECG arrhythmias recognition system based on independent component analysis feature extraction," in *TENCON 2006. 2006 IEEE Region 10 Conference*, 2006, pp. 1–4.

[8] G. B. Moody, R. G. Mark, and A. L. Goldberger, "PhysioNet: a web-based resource for the study of physiologic signals," *Eng. Med. Biol. Mag. IEEE*, vol. 20, no. 3, pp. 70–75, 2001.

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

[10] J. Karhunen and E. Oja, "Hyvärinen, A., & Independent component analysis.," *John Wiley Sons*, 2001.

[11] T. He, G. Clifford, and L. Tarassenko, "Application of independent component analysis in removing artefacts from the electrocardiogram," *Neural Comput. Appl.*, vol. 15, no. 2, pp. 105–116, 2006.

[12] A. Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis," *Neural Netw. IEEE Trans. On*, vol. 10, no. 3, pp. 626–634, 1999.

[13] A. Hyvärinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural Comput.*, vol. 9, no. 7, pp. 1483–1492, 1997.

[14] C. J. James and C. W. Hesse, "Independent component analysis for biomedical signals," *Physiol. Meas.*, vol. 26, no. 1, p. R15, 2005.

[15] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, no. 6, pp. 1129–1159, 1995.

[16] A. Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis," *Neural Netw. IEEE Trans. On*, vol. 10, no. 3, pp. 626–634, 1999.

[17] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm," *Biomed. Eng. IEEE Trans. On*, no. 3, pp. 230–236, 1985.

[18] A. S. Al-Fahoum and I. Howitt, "Combined wavelet transformation and radial basis neural networks for classifying life-threatening cardiac arrhythmias," *Med. Biol. Eng. Comput.*, vol. 37, no. 5, pp. 566–573, 1999.

[19] N. Belgacem, M. A. Chikh, and F. B. Reguig, "Supervised classification of ECG using neural networks," 2003.

[20] S.-N. Yu and K.-T. Chou, "A switchable scheme for ECG beat classification based on independent component analysis," *Expert Syst. Appl.*, vol. 33, no. 4, pp. 824–829, 2007.

[21] J.-S. R. Jang, C.-T. Sun, and E. Mizutani, "Neuro-fuzzy and soft computing-a computational approach to learning and machine intelligence [Book Review]," *Autom. Control IEEE Trans. On*, vol. 42, no. 10, pp. 1482–1484, 1997.

[22] J.-S. R. Jang, C.-T. Sun, and E. Mizutani, "Neuro-fuzzy and soft computing-a computational approach to learning and machine intelligence [Book Review]," *Autom. Control IEEE Trans. On*, vol. 42, no. 10, pp. 1482–1484, 1997.

[23] Y. H. Hu, S. Palreddy, and W. J. Tompkins, "A patient-adaptable ECG beat classifier using a mixture of experts approach," *Biomed. Eng. IEEE Trans. On*, vol. 44, no. 9, pp. 891–900, 1997.

[24] X. Jiang, L. Zhang, Q. Zhao, and S. Albayrak, "ECG arrhythmias recognition system based on independent component analysis feature extraction," 2006, pp. 1–4.

[25] Y. Wu and L. Zhang, "ECG classification using ICA features and support vector machines," in *Neural Information Processing*, 2011, pp. 146–154.

[26] Z. Dokur and T. Ölmez, "ECG beat classification by a novel hybrid neural network," *Comput. Methods Programs Biomed.*, vol. 66, no. 2, pp. 167–181, 2001.

[27] İ. Güler and E. D. Übeyli, "A modified mixture of experts network structure for ECG beats classification with diverse features," *Eng. Appl. Artif. Intell.*, vol. 18, no. 7, pp. 845–856, 2005.

[28] İ. Güler, "ECG beat classifier designed by combined neural network model," *Pattern Recognit.*, vol. 38, no. 2, pp. 199–208, 2005.

[29] P. de Chazal, "A switching feature extraction system for ECG heartbeat classification," in *Computing in Cardiology Conference (CinC), 2013*, 2013, pp. 955–958.

[30] M. I. Owis, A.-B. Youssef, and Y. M. Kadah, "Characterisation of electrocardiogram signals based on blind source separation," *Med. Biol. Eng. Comput.*, vol. 40, no. 5, pp. 557–564, 2002.

[31] T. M. Nazmy, H. El-Messiry, and B. Al-Bokhity, "Adaptive neuro-fuzzy inference system for classification of ECG signals," 2010, pp. 1–6.