City Research Online

City, University of London Institutional Repository

This is the published version of the paper.

This version of the publication may differ from the final published version.

sensors

Letter

A Speedy Cardiovascular Diseases Classifier Using Multiple Criteria Decision Analysis

Wah Ching Lee [1], Faan Hei Hung [2], Kim Fung Tsang [2,*], Hoi Ching Tung [3], Wing Hong Lau [2], Veselin Rakocevic [3] and Loi Lei Lai [4]

[1] Department of Electronic and Information Engineering, Hong Kong Polytechnic University, Hong Kong, China; E-Mail: enwclee@polyu.edu.hk

[2] Department of Electronic Engineering, City University of Hong Kong, Hong Kong, China; E-Mails: fhhung4@cityu.edu.hk (F.H.H.); itwhlau@cityu.edu.hk (W.H.L.)

[3] School of Engineering & Mathematical Science, City University London, Northampton Square, London EC1V 0HB, UK; E-Mails: cherry.tung.1@city.ac.uk (H.C.T.); V.Rakocevic@city.ac.uk (V.R.)

[4] Energy Strategy, Planning, Policy Support, R&D Centre, State Grid Energy Research Institute, SGCC Administrative Area, Future Science and Technology Park, Changping, Beijing 102209, China; E-Mail: lailaili@sgeri.sgcc.com.cn

* Author to whom correspondence should be addressed; E-Mail: ee330015@cityu.edu.hk; Tel.: +852-3442-7806; Fax: +852-3442-0562.

Abstract: Each year, some 30 percent of global deaths are caused by cardiovascular diseases. This figure is worsening due to both the increasing elderly population and severe shortages of medical personnel. The development of a cardiovascular diseases classifier (CDC) for auto-diagnosis will help address solve the problem. Former CDCs did not achieve quick evaluation of cardiovascular diseases. In this letter, a new CDC to achieve speedy detection is investigated. This investigation incorporates the analytic hierarchy process (AHP)-based multiple criteria decision analysis (MCDA) to develop feature vectors using a Support Vector Machine. The MCDA facilitates the efficient assignment of appropriate weightings to potential patients, thus scaling down the number of features. Since the new CDC will only adopt the most meaningful features for discrimination between healthy persons *versus* cardiovascular disease patients, a speedy detection of cardiovascular diseases has been successfully implemented.

1. Introduction

Electrocardiogram (ECG) signals, characterized by P waves, Q waves, S waves, QRS complexes and T waves, are important information for cardiovascular disease diagnosis by cardiologists. Such a diagnosis requires the development of a cardiovascular diseases classifier (CDC). Generally, a CDC mainly comprises feature vectors extraction and building a classifier via machine learning algorithms like an Artificial Neural Network or Support Vector Machine. Features can be divided into three categories: non-fiducial features, fiducial features, and hybrid features. Non-fiducial features normally refer to features that do not characterize the ECG signals using P waves, Q waves, S waves, QRS complexes and T waves [1–5], and *vice versa* for fiducial features [6,7]. Hybrid features refer to feature vectors constructed by both non-fiducial and fiducial features [8–10].

In this investigation, a Support Vector Machine (SVM) is utilized to construct the CDC for the four most common types of cardiovascular diseases, namely bundle branch block, myocardial infarction, heart failure, and dysrhythmia. Seven criteria, including overall accuracy (OA), sensitivity (S_e), specificity (S_p), area under the curve (AUC), training time (T_r), testing time (T_e), and number of features (N_f), which are features indicative of the speed and accuracy of detection, are used as the essential parameters to compute the analytic hierarchy process (AHP) score to aid the multiple criteria decision analysis (MCDA) for the evaluation of the optimal CDC. Traditional work usually aims at the highest overall accuracy and/or lowest testing time. In reality, every end user has to specify the weights between criteria. It is not uncommon to find a ratio setting by intuition or simply a direct 1:1 assignment is adopted. It is noted that the practical needs of volunteers are neglected or not targeted. In the new method, assignments of criteria are devised for AHP analysis. The incorporation of AHP analysis in the classifier enables the consideration of the need of volunteers. This letter is organized as follows: the design of an optimal CDC is presented in Section 2. Multiple criteria decision analysis of the optimal CDC is given in Section 3. In Section 4, the AHP is formulated and a performance score is obtained from which the performance is analyzed and compared to traditional schemes. Finally, conclusions are drawn in Section 5.

2. Design of the optimal CDC

Figure 1 summarizes the block diagram of the new method. After the retrieval of ECG data, feature vectors are extracted. The SVM classifiers are then designed based on the features combinations. Therefore, N configurations can be obtained. The best model is selected among configuration f_1 to configuration f_N based on seven criteria, namely overall accuracy, sensitivity, specificity, area under the curve, training time, testing time, and number of features, with the aid of MCDA *via* AHP. The details of the new method are illustrated in the following.
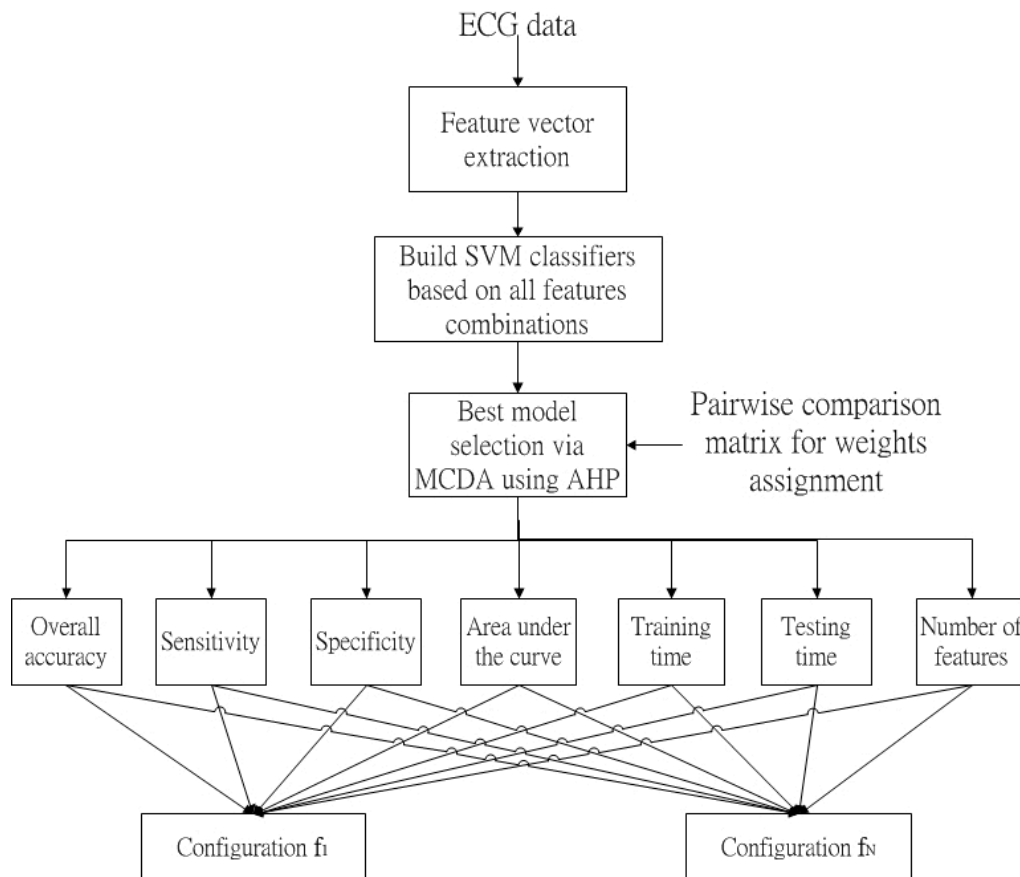
Figure 1. Block diagram of the new method.

2.1. Data Preprocessing and Features Construction

The data is obtained from an online and open access database [11,12]. A group of healthy candidates as well as candidates with the four most common types of cardiovascular diseases are selected: 52 candidates from health control, 15 bundle branch block candidates, 148 myocardial infarction candidates, 18 heart failure candidates and 14 dysrhythmia candidates. The unequal sample size in each class will lead to a bias of the SVM classifier [13]. The Lead I ECG signal is further partitioned into 30 s sub-signals to obtain 500 samples of healthy candidates and 125 samples of unhealthy candidates (of each type of cardiovascular disease). This process aims at equalizing the number of samples in each class (healthy and unhealthy). Before the introduction of these four diseases, the notations are briefed. Denote RR-interval to be the consecutive R points between consecutive ECG signals, QRS complex is the time between Q wave and S wave where point R is between Q wave and S wave. Similarly, QT interval refers to the time between point Q wave and T wave. The background of these four diseases is presented as follows:

(i)   Myocardial Infarction: Irregular heartbeat and thus irregular RR-interval may occur in the ECG signal of the patients [14];

(ii) Bundle Branch Block: Patients have QRS complex with value exceeding 0.12 ms [15];

(iii) Dysrhythmia: The heartbeat can be more than 100 beats per minute or less than 60 beats per minute. Thus, RR-interval is different from the normal ECG signal. Also, the QT interval may increase if the type of cardiovascular disease is ventricular arrhythmias [16];

(iv) Heart Failure: A finding of prolonged QT interval in the ECG signals of the patients [17].

As a result, Q wave, R wave and S wave, QRS complex, and RR-interval are representative features to identify between healthy persons *versus* cardiovascular patients. The feature vector consists of 10 features using the average and standard deviation of these five parameters. Before detecting and computing the features, the ECG signals will undergo data preprocessing [18]. The maximum frequency of an ECG signal is typically less than 60 Hz, thus a bandpass filter with cutoff frequencies at 1 Hz and 60 Hz is implemented. A derivative filter is then applied to sharpen the Q, R, and S wave. Finally, signal squaring and sliding window integration are utilized for the location of Q, R, and S wave.

2.2. Cardiovascular Diseases Classifier Construction

The CDC is constructed by employing SVM with a 10-dimensional feature vector. This algorithm uses a Lagrange Multiplier with a set of support vectors, a set of weighting and an offset bias [19,20]. This report focuses on the design of CDC.

The performance of CDC is dictated by OA, S_e, S_p, AUC, T_r, T_e, and N_f. It directly classifies the ECG signal into healthy (negative response) candidates and unhealthy (positive response) candidates. OA, S_e, S_p, and AUC are related to the accuracy of CDC. T_r is the time required to train the CDC and T_e is the time needed to detect the ECG signal. In this investigation, CDC will be trained up and validated with the ECG datasets. For the analysis of positive response—Class 0, 500 healthy patients are used. For the analysis of positive response—Class 1, 125 bundle branch block patients, 125 myocardial infarction patients, 125 heart failure patients and 125 dysrhthmia patients are retrieved from the database. Table 1 lists the datasets for CDC with binary classifier.

Table 1. Database specification of ECG data for CDC.

Class 0 (Healthy/Negative Response)	Number of Samples	Class 1 (Unhealthy/Positive Response)	Number of Samples
PTB diagnostic (Healthy)	500	Bundle Branch Block	125
		Myocardial Infarction	125
		Heart Failure	125
		Dysrhthmia	125

The CDC utilizes a 10-fold cross validation for performance evaluation [21] and the polynomial kernel function (third order) is utilized for SVM analysis. There is a total of 1023 combinations ($\sum_{n=1}^{10} {}_{10}C_n$), thus 1023 configurations can be formulated from a selection (from 1 to 10) of the 10 features. For the jth configuration where j = 1,…,1023, namely f_j, its corresponding criteria, OA, S_e, S_p, AUC, T_r, T_e, and N_f are recorded. The main settings of SVM are summarized as follows, in general, the default setting is utilized in the MATLAB toolbox:

(i) Number of classes: Two;

(ii) Class 0: 500 Healthy candidates;
   Class 1: 125 bundle branch block candidates, 125 myocardial infarction candidates, 125 heart failure candidates, and 125 dysrhthmia candidates;

(iii) Feature vector: The maximum dimensionality is 10, which consists of: {Q wave average, Q wave standard deviation, R wave average, R wave standard deviation, S wave average, S wave standard deviation, QRS complex average, QRS complex standard deviation, RR-interval mean, RR-interval standard deviation};

(iv) Kernel function: 3rd order polynomial;

(v) Fold of cross validation: Ten-fold 1023 classifiers are constructed in 1023 configurations; the results are tabulated in Table 2.

Table 2. CDC of each configuration.

f_j	OA	S_e	S_p	AUC	T_r (s)	T_e (s)	N_f
f_1	0.324	0.350	0.298	0.321	3.5	2.3	1
f_2	0.310	0.324	0.296	0.303	3.4	2.5	1
f_3	0.298	0.288	0.308	0.287	3.6	2.4	1
…	…	…	…	…	…	…	…
f_{1021}	0.986	0.988	0.984	0.972	4.9	3.4	10
f_{1022}	0.964	0.970	0.958	0.946	5.1	3.4	10
f_{1023}	0.970	0.974	0.966	0.949	4.3	3.5	10

3. Multiple Criteria Decision Analysis of the Optimal CDC

In Table 2, seven criteria, namely OA, S_e, S_p, AUC, T_r, T_e, and N_f, are employed for performance evaluation of the 1023 scenarios. Multiple criteria decision making (MCDM) has been utilized in many areas since the 1990s [22]. It entails using the particular characteristics of cardiovascular diseases. By allocating appropriate weightings, the analytic hierarchy process (AHP) is adopted to evaluate and analyze the best scenarios among the 1023 scenarios investigated. The allocation of weightings confronts the feedback from an AHP analysis of 200 volunteers from which a pairwise comparison 7×7 matrix A_m (m = 1, …, 200) is formulated. It is intuitively understood that T_e should be as low as possible and that the accuracy should be kept to an acceptable level. Since the speed of detection is the prime factor of importance, the analysis on MCDA reveals that high weightings should be assigned to OA, S_e, S_p, AUC, T_e. These five parameters are referred as primary parameters. While N_f is typically preferred to be small for speedy detection, it is noted that T_r will not affect the detection time. Hence N_f and T_r are classified as the secondary parameters.

The volunteers are required to fill in the $a_{m,ij}$, where i and j are between 1 and 7, in Table 3. The AHP based MCDA CDC is referred as the new classifier (NC). Traditional classifiers (TC) in [3,7,8] are also evaluated. Both the NC and the TC are applied to the three feature groups (non-fiducially features, fiducially features and hybrid features in [3,7,8]. The performance comparison between the NC and the TC is tabulated in Table 4. Based on the discussion for AHP formulation, the assignment of values of $a_{m,ij}$ are based on the following guidelines:

(i)   Write 1 if equal importance of i and j;
(ii) Write 3 if i is slightly more important than j;
(iii) Write 5 if i is more important than j;
(iv) Write 7 if i is strongly more important than j;
(iv) Write 9 if i is absolutely more important than j.

Table 3. Pairwise comparison 7×7 matrix A_m.

	OA	S_e	S_p	AUC	T_r	T_e	N_f
OA	1	$a_{m,12}$	$a_{m,13}$	$a_{m,14}$	$a_{m,15}$	$a_{m,16}$	$a_{m,17}$
S_e	$a_{m,21}$	1	$a_{m,23}$	$a_{m,24}$	$a_{m,25}$	$a_{m,26}$	$a_{m,27}$
S_p	$a_{m,31}$	$a_{m,32}$	1	$a_{m,34}$	$a_{m,35}$	$a_{m,36}$	$a_{m,37}$
AUC	$a_{m,41}$	$a_{m,42}$	$a_{m,43}$	1	$a_{m,45}$	$a_{m,46}$	$a_{m,47}$
T_r	$a_{m,51}$	$a_{m,52}$	$a_{m,53}$	$a_{m,54}$	1	$a_{m,56}$	$a_{m,57}$
T_e	$a_{m,61}$	$a_{m,62}$	$a_{m,63}$	$a_{m,64}$	$a_{m,65}$	1	$a_{m,67}$
N_f	$a_{m,71}$	$a_{m,72}$	$a_{m,73}$	$a_{m,74}$	$a_{m,75}$	$a_{m,76}$	1

Table 4. Performance of NC *versus* TC.

Method	Datasets (Number of Samples)	Features	Results (Related Work TC)	Results (New Work NC)
Two-layered Hidden Markov Model [3]	MIT-BIH database (34,799 samples from 16 Arrhythmia candidates)	P-R interval, QRS complex interval and T sub-wave interval	OA = 0.992 S_e = 0.993 S_p = 0.992 AUC = 0.971 T_r = 3.7 s T_e = 2.7 s N_f = 3	OA = 0.987 S_e = 0.99 S_p = 0.984 AUC = 0.966 T_r = 3.4 s T_e = 1.9 s N_f = 2
Cross wavelet transform with a threshold based classifier [7]	The PTB Diagnostic ECG database (18,489 samples from 52 healthy control candidates and 148 myocardial infarction candidates)	Total sum of wavelet cross spectrum value and total sum of wavelet coherence	OA = 0.976 S_e = 0.973 S_p = 0.988 AUC = 0.949 T_r = 6.2 s T_e = 4.1 s N_f = 6	OA = 0.966 S_e = 0.978 S_p = 0.958 AUC = 0.933 T_r = 5.6 s T_e = 2.8 s N_f = 4
SVM [8]	CU database, VF database, and AHA database (40,956 samples from 67 Ventricular fibrillation and rapid ventricular tachycardia candidates)	Leakage, count 1, count 2, count 3, A1, A2, A3, time delay, FSMN, cover bin, frequency bin, kurtosis, and complexity	OA = 0.952 S_e = 0.951 S_p = 0.951 AUC = 0.943 T_r = 4.8 s T_e = 2.7 s N_f = 13	OA = 0.947 S_e = 0.952 S_p = 0.942 AUC = 0.937 T_r = 4.5 s T_e = 1.6 s N_f = 10

The pairwise comparison 7×7 matrix A_m is then normalized, and $Anorm_m$ can be obtained by modifying the matrix entries $a_{m,ij}$ in A_m into matrix entries $anorm_{m,ij}$ in $Anorm_m$:

$$anorm_{m,ij} = \frac{a_{m,ij}}{\sum_{l=1}^{7} a_{m,lj}} \tag{1}$$

By averaging each row of Equation (1), the corresponding 7×1 priority matrix w_m with entries $w_{m,k}$ for $k = 1,\ldots,7$ is given by:

$$w_{m,k} = \frac{1}{7} \sum_{l=1}^{7} anorm_{m,kl} \tag{2}$$

Denote $C_{p,q}$, ($p = 1,…,7$ and $q = 1,…,1023$) be the pth criteria, and q^{th} scenario of CDC. $C_{p,q}$ is normalized to become $C_{p,q,norm}$. The final score for each scenario, AHP_q, is evaluated by:

$$AHP_q = \sum_{l=1}^{7} C_{p,q,norm} \left(\frac{1}{200} \sum_{m=1}^{200} w_{m,l}\right) \tag{3}$$

To avoid inconsistency in the construction of pairwise comparison matrices, the optimal CDC is concluded from the highest value of AHP_q [23]. It is evaluated that the optimal CDC is obtained from scenario f_{652}, with feature vector composes of average of Q, standard deviation of Q, standard deviation of S, average of QRS mean, standard deviation of QRS, average of RR-interval, and standard deviation of RR-interval, with AHP_{652} as follows: OA = 0.988, S_e = 0.992, S_p = 0.985, AUC = 0.982, T_r = 4.5 s, T_e = 2.8 s, N_f = 7.

4. AHP Scores and Analysis

The performance scores between the NC and the TC [3,7,8] are evaluated and tabulated in Table 4. In this investigation, the algorithms in related work have been evaluated, with the addition of MCDA using AHP to obtain a best scenario by assigning weights to the seven criteria. As the new work and related works are in the same application area, the classification of cardiovascular diseases, the weight assignment can be reused to facilitate performance comparisons. From Table 4, the percentage changes are evaluated as follows:

1. Percentage change compared with AHP scores from [3]: OA = −0.504%, S_e = −0.302%, S_p = −0.807%, AUC = −0.515%, T_r = −8.109%, T_e = −29.630%, and N_f = −33.333%. It is concluded that there is an improvement of 30% in speed of detection of cardiovascular diseases @~99.5% accuracy.

2. Percentage change compared with AHP scores from [7]: OA = −1.025%, S_e = 0.514%, S_p = −3.036%, AUC = −1.686%, T_r = −9.677%, T_e = −31.707%, and N_f = −33.333%. It is concluded that there is an improvement of 30% in speed of detection of cardiovascular diseases @~99% accuracy.

3. Percentage change compared with AHP scores from [8]: OA = −0.525%, S_e = 0.105%, S_p = −0.946%, AUC = −0.636%, T_r = −6.250%, T_e = −40.741%, and N_f = −23.077%. It is concluded that there is an improvement of 40% in speed of detection of cardiovascular diseases @~99.5% accuracy.

The analysis reveals that in the NC, the speed of detection has been increased by 30%–40% while the accuracy is retained at ~99%–99.5% of the TC. It is seen that there the reduction of OA, S_e, and S_p are less than 1%. Thus the AHP based MCDA CDC is a reliable and speedy detection scheme for cardiovascular diseases.

5. Conclusions

In this letter, an optimal cardiovascular diseases classifier (CDC) has been proposed and implemented by using an analytic hierarchy process (AHP) to facilitate multiple criteria decision analysis (MCDA). The four most common types of cardiovascular diseases, namely bundle branch

block, myocardial infarction, heart failure, and dysrhythmia are considered. Seven criteria, namely OA, S_e, S_p, AUC, T_r, T_e, and N_f are carefully considered and chosen to be the criteria for deriving the AHP score of MCDA to achieve the optimal CDC. The optimal CDC, the new classifier, achieves the following scores: OA = 0.988, S_e = 0.992, S_p = 0.985, AUC = 0.982, T_r = 4.5 s, T_e = 2.8 s, N_f = 7. Analysis and comparison with previous works show that the speed of detection cardiovascular diseases has been increased by 30%–40% while the accuracy is retained at ~99%–99.5% of traditional classifiers. In conclusion, the AHP based MCDA CDC is a reliable and speedy detection scheme for cardiovascular diseases.

Acknowledgments

Author Contributions

All authors are responsible for the writing of the manuscript. Wah Ching Lee is responsible for the original concept of the paper which also includes design, data collection, and pre-processing of data. Faan Hei Hung, Kim Fung Tsang. Hoi Ching Tung, Wing Hong Lau, Veselin Rakocevic, and Loi Lei Lai evaluated the performance of classifier between proposed method and related work.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Lee, Y.S.; Chung, W.Y. Visual Sensor Based Abnormal Event Detection with Moving Shadow Removal in Home Healthcare Applications. *Sensors* **2012**, *12*, 573–584.
2. Noh, Y.H.; Jeong, D.U. Implementation of a Data Packet Generator Using Pattern Matching for Wearable ECG Monitoring Systems. *Sensors* **2014**, *14*, 12623–12639.
3. Liang, W.; Zhang, Y.; Tan, J.; Li, Y. A Novel Approach to ECG Classification Based upon Two-layered HMMs in Body Sensor Networks. *Sensors* **2014**, *14*, 5994–6011.
4. Staa, T.-P.V.; Gulliford, M.; Ng, E.S.-W.; Goldacre, B.; Smeeth, L. Prediction of Cardiovascular Risk Using Framingham, ASSIGN and QRISK2: How Well Do They Predict Individual Rather than Population Risk? *PLoS One* **2014**, *9*, 1–10.
5. Sanz, J.A.; Galar, M.; Jurio, A.; Brugos, A.; Pagola, M.; Bustince, H. Medical diagnosis of cardiovascular diseases using an interval-valued fuzzy rule-based classification system. *Appl. Soft Comput.* **2014**, *20*, 103–111.
6. Tseng, K.-K.; He, X.; Kung, W.-M.; Chen, S.-T.; Liao, M.; Huang, H.-N. Wavelet-Based Watermarking and Compression for ECG Signals with Verification Evaluation. *Sensors* **2014**, *14*, 3721–3736.

7. Banerjee, S.; Mitra, M. Application of Cross Wavelet Transform for ECG Pattern Analysis and Classification. *IEEE Trans. Instrum. Meas.* **2014**, *63*, 326–333.

8. Li, Q.; Rajagopalan, C.; Clifford, G.D. Ventricular Fibrillation and Tachycardia Classification Using Machine Learning Approach. *IEEE Trans. Biomed. Eng.* **2014**, *61*, 1607–1613.

9. Sun, L.; Lu, Y.; Yang, K.; Li, S. ECG Analysis Using Multiple Instance Learning for Myocardial Infarction Detection. *IEEE Trans. Biomed. Eng.* **2012**, *59*, 3348–3356.

10. Xie, B.; Minn, H. Real-Time Sleep Apnea Detection by Classifier Combination. *IEEE Trans. Inf. Technol. Biomed.* **2012**, *16*, 469–477.

11. The PTB Diagnostic ECG Database, Physionet. Available online: http://www.physionet.org /physiobank/database/ptbdb/ (accessed on 10 September 2014).

12. Goldberger, A.L.; Amaral, L.A.N.; Class, L.; Hausdorff, J.M.; Ivanov, P.C.H.; Mark, R.G.; Mietus, J.E.; Moody, G.B.; Peng, C.-K.; Stanley, H.E. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation* **2000**, *101*, e215–e220.

13. Vapnik, V.N. *The Nature of Statistical Learning*; Springer: Berlin, Germany, 1995.

14. Kuchar, D.L.; Thorburn, C.W.; Sammel, N.L. Prediction of serious arrhythmic events after myocardial infarction: Signal-averaged electrocardiogram, holter monitoring and radionuclide ventriculography. *J. Am. College Cardiol.* **1987**, *9*, 531–538.

15. Rotman, M.; Triebwasser, J.H. A clinical and follow-up study of right and left bundle branch block. *Circulation* **1975**, *51*, 477–484.

16. Krowka, M.J.; Pairolero, P.C.; Trastek, V.F.; Payne, W.S.; Bernatz, P.E. Cardiac dysrhythmia following pneumonectomy. Clinical correlates and prognostic significance. *Chest* **1987**, *91*, 490–495.

17. Gottdiener, J.S.; Arnold, A.M.; Aurigemma, G.P.; Polak, J.F.; Tracy, R.P.; Kitzman, D.W.; Gardin, J.N.; Rutledge, J.E.; Boineau, R.C. Predictors of congestive heart failure in the elderly: The cardiovascular health study. *J. Am. College Cardiol.* **2000**, *35*, 1628–1637.

18. Tompkins, W.J. *Biomedical Digital Signal Processing C-Language Examples and Laboratory Experiments for the IBM PC*; Prentice Hall: Upper Saddle River, NJ, USA, 2000; pp. 236–264.

19. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: Singapore, 2006; pp. 325–343.

20. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297.

21. McLachlan, G.J.; Do, K.A.; Ambroise, C. Analyzing Microarray Gene Expression Data. *Supervised Classification of Tissue Samples*; John Wiley & Sons: New York, NY, USA, 2004; pp. 221–251.

22. Köksalan, M.; Wallenius, J.; Zionts, S. *Multiple Criteria Decision Making: From Early History to the 21st Century*; World Scientific Publishing: Singapore, 2001; pp. 43–62.

23. Ozdemir, M.S. Validity and inconsistency in the analytic hierarchy process. *Appl. Math. Comput.* **2005**, *161*, 707–720.