Myocardial Infarction Classification by Morphological Feature Extraction from Big 12-Lead ECG Data

Julia Tzu-Ya Weng[1(✉)], Jyun-Jie Lin[2], Yi-Cheng Chen[3], and Pei-Chann Chang[2]

[1] Department of Computer Science and Engineering,
Yuan Ze University, Taoyuan, Taiwan
julweng@saturn.yzu.edu.tw
[2] Department of Information Management, Yuan Ze University,
Taoyuan, Taiwan
s969203@mail.yzu.edu.tw, iepchang@saturn.yzu.edu.tw
[3] Department of Computer Science and Information Engineering,
Tamkang University, Tamsui, New Taipei City, Taiwan
ycchen@mail.tku.edu.tw

Abstract. Rapid and accurate diagnosis of patients with acute myocardial infarction is vital. The ST segment in Electrocardiography (ECG) represents the change of electric potential during the period from the end of ventricular depolarization to the beginning of repolarization and plays an important role in the detection of myocardial infarction. However, ECG monitoring generates big volumes of data and the underlying complexity must be extracted by a combination of methods. This study combines the advantages of polynomial approximation and principal component analysis. The proposed approach is stable for the 12-lead ECG data collected from the PTB database and achieves an accuracy of 98.07 %.

Keywords: 12-lead ECG · Myocardial infarction · Principal component analysis · Polynomial approximation · Support vector machine

1 Introduction

Myocardial Infarction (MI) is the death of heart due to the sudden blockage of a coronary artery by a blood clot. Blockage of a coronary artery deprives the heart muscle of blood and oxygen, causing injury to the heart muscle. Among the diagnostic tests available to detect heart muscle damage, electrocardiogram (ECG) is one of the most widely used non-invasive diagnostic tools for cardiopulmonary diseases.

ECG monitors the patients' heartbeat and gives accurate and important information about the activities of the atrium and ventricle. A human's normal ECG waveform is shown in Fig. 1. The basic components of an ECG complex are P wave, which represents atrial depolarization, QRS complex, which represents ventricular depolarization and T wave, which corresponds to the period of ventricular repolarization. One normal cardiac cycle starts at the sinus node with the depolarization of the right atrium

W.-C. Peng et al. (Eds.): PAKDD 2014 Workshops, LNAI 8643, pp. 689–699, 2014.
DOI: 10.1007/978-3-319-13186-3_61

and spreads toward the entire atria in a well-ordered manner. Next, the depolarization impulse reaches the ventricles and the fast contraction produces the QRS complex of the ECG. Finally, ventricular repolarization generates the T-wave complex and the cardiac cycle of one heart beat is terminated [1].

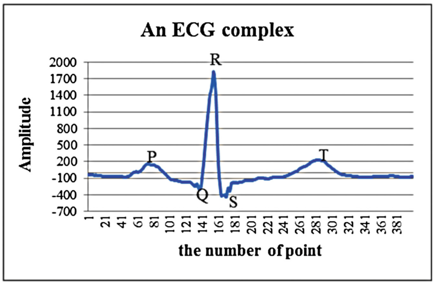

Fig. 1. Basic components of an ECG complex

Clinical 12-lead ECG data are now available in most hospitals and include more detailed information about cardiac disease. The standard 12-lead ECG is composed of six leads: the limb leads, which corresponds to the subject's four extremities, the central terminal, which is the average of the potentials from the limb leads, and six horizontal leads, which are also called chest leads [2]. These leads offer 12 different angles for visualizing the activities of the heart and are named Lead I, II, III, aVL, aVF, aVR, V1, V2, V3, V4, V5 and V6, respectively. Because of the different aspect of recording the polarization and depolarization of a heart-beat cycle, the data volume generated from 12-lead ECG is big and the complexity is high, though more complete information of heart activities can be obtained for cardiac disease classification.

The key in treating ECG complex is using the morphology in time detection [3, 4]. The occlusion of a coronary artery following the rupture of a vulnerable atherosclerotic plaque can represent typical two types of ECG manifestations: ST elevation and ST depression plus T-wave changes [5]. In ECG monitor, ST segment means the change of electric potential during the period which from the end of ventricular depolarization to the origin of repolarization. Hence, ST shape change is a very important parameter for the diagnosis of cardiac disease.

In the early stages of acute MI, the ECG may look normal. Therefore, it is very important to identify a MI from a patient's 12-lead ECG data in the beginning, so that a medical doctor can suggest a patient for expeditious reperfusion therapy and improve prognosis significantly. However, owing to the large volume, great complexity and high dimensionality of 12-lead ECG data, accurately classifying MI and normal data is not a trivial task. Therefore, this study proposes a hybrid approach including polynomial approximation and Principal Component Analysis (PCA) to deal with the challenge in this field. The whole idea is based on studying the effect of feature extraction from ECG data by analyzing the morphological characteristics. In the next section, the related literatures and our analysis workflow will be briefly described.

2 Literature Review

In normal conditions, the ST segment is a horizon line, but in heart diseases, it may show as various waveforms [6]. In MI classification, ST segment change is an important criterion for diagnosis or academic research. Therefore, lots of studies try to use several approaches to extract the features from ST segment in ECG or utilize machine learning techniques to distinguish the difference between normal and MI by using the information in ECG waveforms. We briefly review the related researches below.

2.1 ECG Waveform Analysis in Frequency Domain

Morphological analysis of ECG signals adopts various signal processing strategies over the past two decades. Since ECG complex is a time series data, using Short Time Fourier Transform (STFT) or wavelet can provide a degree of temporal resolution indicating the changes in the frequency spread with time [7]. Wavelet coefficients represent measures of similarity of local shape of the signal with mother and baby wavelets. The multi-resolution properties of WT were effectively utilized for identifying the characteristic points in the ECG waveform and hence for the analysis of PR, RR, ST intervals [8, 9].

2.2 Feature Extraction by Principle Component Analysis

Principle Component Analysis (PCA) is a technique that is generally used for reducing the dimensionality of multivariate datasets [10]. Considering a vector of n random variables x for which the covariance matrix is Σ, the principal components (PCs) can be defined by

$$\mathrm{z} = \mathrm{Ax} \tag{1}$$

where z is the vector of n PCs and A is the n × n orthogonal matrix with rows that are the eigenvectors of Σ. The eigenvalues of Σ are proportional to the fraction of the total variance accounted for by the corresponding eigenvectors, so that the PCs explaining most of the variance in the original variables can be identified.

PCA has been commonly used to analyze ECG features. The reduced dimensions (features) can be used to represent the beat morphology. For certain periods in ECG, PCA can be also applied to ST-T segment analysis for the detection of myocardial ischemia and abnormalities in ventricular repolarization [11]. PCA in ECG signal processing takes its starting point from the samples of a segment located in some suitable part of the heartbeat. PCA utilizes a representation of the data in a statistical domain rather than a time or frequency domain. In ECG signals, the information can also be separated from the noise or baseline shift by PCA analysis [12].

Some researchers also use PCA to analyze multi-lead ECG [13]. A common way to convert the multi-lead ECG into suitable data is to concatenate it. Ge et al. [14] proposed the research to concatenate the 12-lead in the order: Lead I, II, III, aVR, aVL, aVF, V1, V2, V3, V4, V5, and V6. The dimension of the data will be higher than single-lead, but the information of 12-lead can be solved by PCA at one time.

2.3 Modeling ECG Waveforms by Statistical Models

To solve the classification problem in ECG, lots of researchers use statistical models, e.g. Hidden Markov Models (HMMs) [15]. HMMs are the stochastic models used for representing an underlying stochastic process that is not observable, but can be observed through the sequence of observed symbols. Because of the morphology of ECG signals, HMMs are mostly adopted for classification [16] and segmentation or delineation [17, 18]. HMMs can find the suitable segmentations of a heart-beat by calculating the state transition. Because of its ability to model ECG waveforms, HMMs can also be applied as a feature extractor for artificial intelligence-based classifier [19]. In [19], the log-likelihood calculated from HMMs can be regarded as the feature of a single lead of ECG and this approach can be easily extended to multi-lead diagnosis.

2.4 ST Shape Change Classification by Polynomial Approximation

Polynomial approximation can also be called "curve fitting" or "polynomial fitting". A polynomial is a function that can be written in the form $p(x) = c_0 + c_1x + \ldots + c_nx^n$ for some coefficients $c_0, \ldots, c_n$. If, $c_n \neq 0$ then the polynomial is said to be of order n. A first-order (linear) polynomial is just the equation of a straight line, while a second-order (quadratic) polynomial describes a parabola. The purposes of using polynomial approximation are (1) to model a nonlinear relationship between dependent and independent variables and interest on the shape of the fitted curve and the related coefficients; (2) to approximate a difficult function (e.g. the density or the distribution function).

In the medical field, curve fitting can be applied as a feature extractor of morphological characteristics [20, 21]. For the various shapes in ST segment, some have tried to use polynomial approximation method to extract the features in ECG [22, 23]. The analysis only considers the relative shape change of the ST segment, but this approach can be used to describe the variation of ST shape and provide the important features from the coefficient of polynomials.

2.5 Support Vector Machine

This section briefly describes the basic SVM and non-linear SVM concepts for typical two-class classification problems. Assuming there is a training set with N samples $(X_i, y_i | X_i \in \Re^n, y_i \in \{-1, +1\})$, a hyper-plane can be defined by the following linear function

$$f(X) = \omega^T X + b \tag{2}$$

where w is the weight vector $\{w_1, w_2, \ldots, w_n\}$ and n is the number of attributes (dimensions) and b is a bias. In order to obtain the separating hyper-plane with the largest margin for each training example, the function yields $f(X) \geq 0$ for $y = +1$ and $f(X) < 0$ for $y = -1$. The training set from the two different classes are separated by the hyper-plane $f(X) = 0$ and the SVM classifier is based on the hyper-plane that maximized the separating margin.

The main objective of linear SVM is to maximize the margin and the equation can defined as

$$\begin{aligned} M(w,b) &= \min_{x_i:y_i=-1} d(w,b;x_i) + \min_{x_i:y_i=1} d(w,b;x_i) \\ &= \min_{x_i:y_i=-1} \frac{|\langle w, x_i \rangle|}{\|w\|} + \min_{x_i:y_i=1} \frac{|\langle w, x_i \rangle|}{\|w\|} \\ &= \frac{1}{\|w\|}\left(\min_{x_i:y_i=-1} |\langle w, x_i \rangle + b| + \min_{x_i:y_i=1} |\langle w, x_i \rangle + b| \right) = \frac{2}{\|w\|} \end{aligned} \tag{3}$$

Hence, a minimal problem can be given

$$\begin{cases} \text{minimize } L(w) = \frac{1}{2}\|w\|^2 \\ \text{subject to } y_i(\langle w, x_i \rangle + b) \geq 1 \end{cases} \tag{4}$$

After Lagrangian transformation, we can conclude the dual problem in Eq. (5)

$$\begin{cases} \text{Maximize: } L_D = \sum\limits_{i=1} \alpha_i - \frac{1}{2}\sum\limits_{ij} \alpha_i \alpha_j y_i y_j \left(\langle x_i, x_j \rangle \right) \\ \text{Subject to: } \sum\limits_{i} \alpha_i y_i = 0, \alpha_i \geq 0, \forall i \end{cases} \tag{5}$$

SVMs can be extended to classify nonlinear data through nonlinear kernel mapping function $K(X_i, X_j)$ to replace the original dot operation. The modified function is as follows.

$$\begin{cases} \text{Maximize: } L_D = \sum\limits_{i=1} \alpha_i - \frac{1}{2}\sum\limits_{ij} \alpha_i \alpha_j y_i y_j K\left(x_i, x_j \right) \\ \text{Subject to: } \sum\limits_{i} \alpha_i y_i = 0, \alpha_i \geq 0, \forall i \end{cases} \tag{6}$$

SVMs are one of the kernel-based learning algorithm [22], there exist lots of mapping functions [23] and here the most popular kernel functions are listed.

(1) Linear kernel

$$K(X_i, X_j) = X_i \cdot X_j \tag{7}$$

(2) Polynomial kernel of degree h:

$$K(X_i, X_j) = (X_i \cdot X_j + 1)^h \tag{8}$$

(3) Gaussian radial basis function kernel:

$$K(X_i, X_j) = \exp\left[-\|X_i - X_j\|^2 \Big/ 2\sigma^2\right] \tag{9}$$

(4) Sigmoid kernel:

$$K(X_i, X_j) = \tanh(kX_i \cdot X_j - \delta) \tag{10}$$

As clinical data are linearly inseparable, the nonlinear SVMs are applied and Gaussian RBF is selected as the kernel function in this study. With the kernel mapping function, data from two classes can always be separated by a hyper plane found by using support vectors and margins. In this study, Gaussian RBF kernel has the sigma value of 1 and we use SVM in the Bioinformatics Toolbox of Matlab and randomly selection cross-validation to retrieve the average accuracy.

3 Methodology

In this paper, two approaches for extracting features are compared. Each heartbeat is analyzed in order to increase the accuracy for MI classification. One method is concatenating ST-T segments in 12-lead ECG and then, using PCA to extract the features. The other method is applying PCA on coefficients of polynomial approximation. Figure 2 shows the overall workflow in this study.

3.1 Pre-processing

This study adopts a simple way to decompress the effect of noise and baseline drift. First, a low-pass filter is applied. The threshold of frequency is set as 40. Then each lead is separated into several signals by Empirical Mode Decomposition (EMD) [24], especially suited for nonlinear and non-stationary signals [25]. EMD can be formulated as the following equation:

$$X(t) = \sum_{j=1}^{n} IMF_j + r \tag{11}$$

The result of EMD produces n intrinsic mode functions (IMFs) and a residue signal. The residue signal can be regarded as a trend line in the original signal. In this study, the residue signal is regarded as the baseline wander due to the low frequency. After subtracting the assumed baseline wander, a median-filter is used to make the signals more stable and clear.

3.2 Heartbeat Location

QRS-wave location is necessary for ST-T segment identification and heartbeat isolation. To locate QRS wave, ICA is used to process the 12-lead ECG data and the estimated sources are sorted by the kurtosis value calculated from each source. The source with the largest kurtosis value is chosen. Following the approach proposed by [26], heartbeat can be isolated automatically from the complete 12-lead ECG complex. After locating the QRS-wave, the location of R-peak can be defined.

The next is to separate the ST-segment from the whole ECG complex. First, given the R-R interval range between two heartbeats, we assume there is a point J. Second, we calculate the one order difference of the candidate interval and selecting the first minimum value as point J in ECG. As suggested by [6], the threshold for deciding the range of different value is between 0.05 to 0.15. We use the ST segment to diagnosis MI disease, and the end point of ST segment in this section means the peak of T wave. To detect the T peak point in T wave, 12-lead ECG data are superimposed together and calculate the maximum value between point J to the next R peak position. Figure 3 shows the result according to the above steps, and the interval between point J and T wave is picked as the ST segment. Here, we define the ST-T segment as the interval from the beginning of point J, followed by the QRS wave, to the peak of the T-wave.

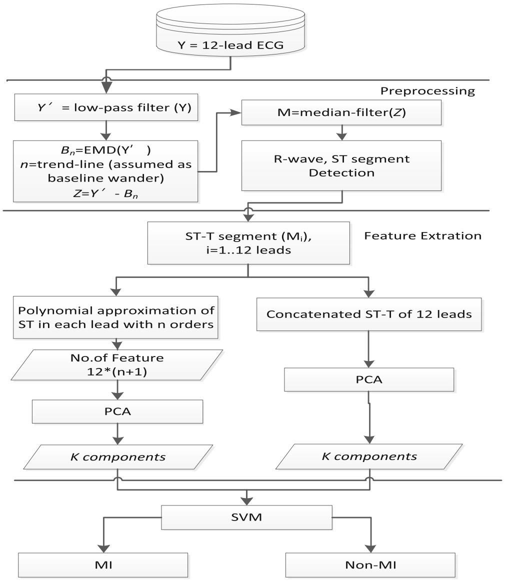
Y = 12-lead ECG
Preprocessing
Y′ = low-pass filter (Y)
M=median-filter(Z)
B_n=EMD(Y′)
n=trend-line (assumed as baseline wander)
Z=Y′ - B_n
R-wave, ST segment Detection
ST-T segment (M_i), i=1..12 leads
Feature Extration
Polynomial approximation of ST in each lead with n orders
Concatenated ST-T of 12 leads
No.of Feature 12*(n+1)
PCA
PCA
K components
K components
SVM
MI
Non-MI

Fig. 2. The proposed framework

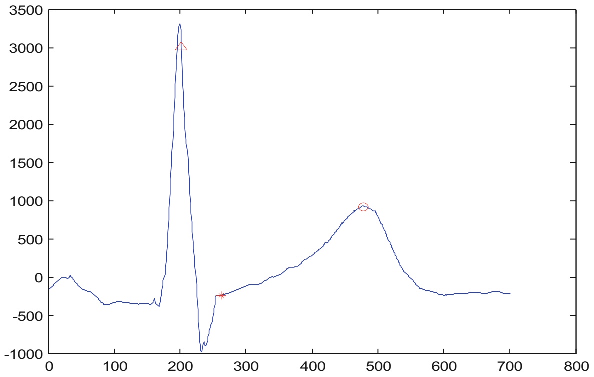

Fig. 3. R peak (Δ), J point (*) and T wave (o) location in an ECG complex

3.3 Feature Extraction by PCA

This study focuses on analyzing whether a single heartbeat belongs to MI in a 12-lead ECG dataset. To combine the information in 12-lead ECG data, the segmented ECG data from the corresponding 12-lead ECG are concatenated in the order of Lead I, II, III, aVR, aVL, aVF, V1, V2, V3, V4, V5, and V6 [21]. PCA is used to reduce and gather significant features.

3.4 Feature Enhanced by PCA

We also adopt another strategy of utilizing PCA to collect coefficients from the polynomial approximation of 12-lead ECG. PCA not only can be used reduce features, but can also be applied to find the most significant features from the original attributes to generate better performance. This study uses polynomial approximation to gather the features of ST segment.

4 Experimental Result

We applied four-fold cross-validation, repeated ten times, to the testing strategy adopted in this research. The performance measurements include accuracy, sensitivity (SE), specificity (SP) and positive predictive (PP). The related equations are listed as below.

$$accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{12}$$

$$SE = \frac{TP}{TP + FN} \tag{13}$$

$$SP = \frac{TN}{TN + FP} \tag{14}$$

$$PP = \frac{TP}{TP + FP} \tag{15}$$

where TP (True Positive) is the number of matched events and FN (False Negative) is the number of events that are not detected by this approach. FP (False Positive) is the number of events detected by this approach. TN (True Negative) indicates the percentage of events identified as truly non-defective, or normal.

The 12-lead ECG data are collected from the PTB-database. There are 549 sets of data, including 148 cases with MI and 301 non-MI cases. The classifier used here is Support Vector Machine and the kernel function selected is RBF kernel with a sigma value of 2. The rule for deciding whether the test case belongs to MI is if more than three-quarters of the number of heartbeats is classified as MI, then the test case is MI; otherwise, the test case will be classified as non-MI case.

Through PCA and polynomial approximation, we try to find the suitable parameters (principle components or PCs) and the appropriate degrees. The number of PCs is initially fixed at 15 and the range of degrees used in polynomial approximation starts from two to ten. We build up a series of testing scenarios, in which the performance of a range of number of PCs is tested with a fixed degree of polynomial approximation. The accuracy increases when the degree of polynomial approximation is four and the number of PCs is larger than seven.

According to our experimental result, we set the PC number at 12 to test a range of degree of polynomial approximation from three to five. Table 1 shows the comparison across different testing models based on their accuracy, sensitivity (SE), specificity (SP) and positive predictivity (PP) measures. These measurements represent the mean values after 30 cross-validations with the corresponding standard deviation. "ST + PCA" indicates the original method of concatenating the ST segments from 12-lead ECG, with features extracted by PCA for SVM. The testing model with the PC number set to 12 and the degree of polynomial approximation fixed at four gives the best overall performance result for MI detection.

Table 1. The performance measurement

		Accuracy	SE	SP	PP
Poly3 + PCA	Mean	96.79 %	**98.74 %**	92.42 %	96.69 %
	std	0.0023	0.0011	0.0077	0.0032
Poly4 + PCA	Mean	**98.07 %**	98.73 %	**96.60 %**	**98.49 %**
	std	0.0029	0.0016	0.0080	0.0035
Poly5 + PCA	Mean	97.96 %	98.71 %	96.26 %	98.34 %
	std	0.0016	0.0013	0.0047	0.0020
ST + PCA	Mean	96.40 %	97.73 %	93.44 %	97.10 %
	std	0.0038	0.0023	0.0115	0.0050

5 Conclusion

PCA and polynomial approximation are considered as two different methods for feature extraction from the ST segment for one heartbeat. We find that these two approaches can be combined to achieve higher performance in MI classification. We further improve the performance of PCA by selecting the proper number of features enhanced by PCA. Since the coefficients of polynomial function can express the variation of ST shape changes, the proposed model indeed increases the performance and reduces the feature space and complexity in a large volume of complex 12-lead ECG data. Through PCA and polynomial approximation, the relationship between ECG and MI disease become more precise.

References

1. Reisne, A.T., Clifford, G.D., Mark, R.G.: The physiological basis of the electrocardiogram. In: Clifford, G.D., Azuaje, F., McSharry, P.E. (eds.) Advanced Methods and Tools for ECG Data Analysis. Artech House Publishing, London (2006)
2. Garcia, T.B., Holtz, N.E.: Introduction to 12-Lead ECG: The Art of Interpretation (2001)
3. de Chazal, P., O'Dwyer, M., Reilly, R.B.: Automatic classification of heartbeats using ECG morphology and heartbeat interval features. IEEE Trans. Biomed. Eng. **51**, 1196–1206 (2004)
4. Reznik, A.G., Ivanov, I.N.: Myocardial morphology in cases of death from acute heart ischemic disease. Arkh. Patol. **69**, 32–35 (2007)
5. Thygesen, K., Alpert, J.S., White, H.D.: Universal definition of myocardial infarction. Circulation **116**, 2634–2653 (2007)
6. Shen, Z., Hu, C., Liao, J., Meng, M.Q.H.: An algorithm of ST segment classification and detection. In: 2010 IEEE International Conference on Automation and Logistics (ICAL'10), pp. 559–564 (2010)
7. Jayachandran, E.S., Joseph, K.P., Acharya, U.R.: Analysis of myocardial infarction using discrete wavelet transform. J. Med. Syst. **34**, 985–992 (2010)
8. Addison, P.S.: Wavelet transforms and the ECG: A review. Physiol. Meas. **26**, R155–R199 (2005)
9. Ghaffari, A., Homaeinezhad, M.R., Akraminia, M., Atarod, M., Daevaeiha, M.: A robust wavelet-based multi-lead electrocardiogram delineation algorithm. Med. Eng. Phys. **31**, 1219–1227 (2009)
10. Jolliffe, I.T.: Principal Component Analysis. Springer, New York (1986)
11. Murugan, S., Radhakrishnan, S.: Automated ischemic beat classification using Genetic Algorithm based Principal Component Analysis. Int. J. Healthc. Technol. Manage. **11**, 151–162 (2010)
12. Chawla, M.P.S.: PCA and ICA processing methods for removal of artifacts and noise in electrocardiograms: A survey and comparison. Appl. Soft Comput. **11**, 2216–2226 (2010)
13. Castells, F., Laguna, P., Sörnmo, L., Bollmann, A., Roig, J.M.: Principal component analysis in ECG signal processing. EURASIP J. Adv. Sig. Process. **2007**, 1–21 (2007). Article ID: 074580
14. Ge, D., Sun, L., Zhou, J., Shao, Y.: Discrimination of myocardial infarction stages by subjective feature extraction. Comput. Methods Programs Biomed. **95**, 270–279 (2009)

15. Chauhan, S., Wang, P., Sing, L.C., Anantharaman, V.: A computer-aided MFCC-based HMM system for automatic auscultation. Comput. Biol. Med. **38**, 221–233 (2008)
16. Andreao, R.V., Dorizzi, B., Boudy, J., Mota, J.C.M.: ST-segment analysis using hidden Markov model beat segmentation: Application to ischemia detection. Comput. Cardiol. **31**, 381–384 (2004)
17. Graja, S., Boucher, J.M.: Hidden Markov tree model applied to ECG delineation. IEEE Trans. Instrum. Meas. **54**, 2163–2168 (2005)
18. Andreao, R.V., Dorizzi, B., Boudy, J., Mota, J.C.M.: ST-segment analysis using hidden Markov model beat segmentation: Application to ischemia detection. Comput. Cardiol. **31**, 381–384 (2004)
19. Chang, P.C., Hsieh, J.C., Lin, J.J., Chou, Y.H., Liu, C.H.: A Hybrid System with Hidden Markov Models and Gaussian Mixture Models for Myocardial Infarction Classification with 12-Lead ECGs. In: Proceedings of the 2009 11th IEEE International Conference on High Performance Computing and Communications (HPCC '09), pp. 110–116 (2009)
20. Georgiou, H., Mavroforakis, M., Dimitropoulos, N., Cavouras, D., Theodoridis, S.: Multi-scaled morphological features for the characterization of mammographic masses using statistical classification schemes. Artif. Intell. Med. **41**, 39–55 (2007)
21. Li, B., Meng, M.Q.H.: Computer-based detection of bleeding and ulcer in wireless capsule endoscopy images by chromaticity moments. Comput. Biol. Med. **39**, 141–147 (2009)
22. Jeong, G.Y., Yu, K.H., Kim, N.G.: A polynomial approximation approach for analyzing ST shape change. In: the 27th Annual International Conference of the IEEE Engineering in Medicine and Biology (EMBS'05), Vol. 7, pp. 4034–4037 (2005)
23. Jeong, G.Y., Yu, K.H., Yoon, M.J., Inooka, E.: ST shape classification in ECG by constructing reference ST set. Med. Eng. Phys. **32**, 1025–1031 (2010)
24. Blanco-Velasco, M., Weng, B., Barner, K.E.: ECG signal denoising and baseline wander correction based on the empirical mode decomposition. Comput. Biol. Med. **38**, 1–13 (2008)
25. Huang, N.E., Shen, Z., Long, S.R., Wu, M.C., Shih, H.H., Zheng, Q., Yen, N.C., Tung, C.C., Liu, H.H.: The empirical mode decomposition and the Hubert spectrum for nonlinear and non-stationary time series analysis. Proc. Royal Soc. A: Math. Phys. Eng. Sci. **454**, 903–995 (1998)
26. Chawla, M.P.S., Verma, H.K., Kumar, V.: A new statistical PCA-ICA algorithm for location of R-peaks in ECG. Int. J. Cardiol. **129**, 146–148 (2008)