Received: 24 August 2018 | Revised: 28 February 2019 | Accepted: 28 April 2019
DOI: 10.1111/exsy.12432

SPECIAL ISSUE PAPER

WILEY Expert Systems

Wearable ECG signal processing for automated cardiac arrhythmia classification using CFASE-based feature selection

Yuwei Zhang[1] | Yuan Zhang[1] | Benny Lo[2] | Wenyao Xu[3]

[1] Shandong Provincial Key Laboratory of Network Based Intelligent Computing, University of Jinan, Jinan, China

[2] The Hamlyn Centre, Imperial College London, London, UK

[3] Department of Computer Science and Engineering, State University of New York at Buffalo, Buffalo, New York

Correspondence
Yuan Zhang, Shandong Provincial Key Laboratory of Network Based Intelligent Computing, University of Jinan. Jinan, China
Email: yzhang@ujn.edu.cn

Funding information
Shandong Provincial Key Research and Development Project, Grant/Award Number: 2017GGX10141; National Natural Science Foundation of China, Grant/Award Number: 61572231

Abstract

Classification of electrocardiogram (ECG) signals is obligatory for the automatic diagnosis of cardiovascular disease. With the recent advancement of low-cost wearable ECG device, it becomes more feasible to utilize ECG for cardiac arrhythmia classification in daily life. In this paper, we propose a lightweight approach to classify five types of cardiac arrhythmia, namely, normal beat (N), atrial premature contraction (A), premature ventricular contraction (V), left bundle branch block beat (L), and right bundle branch block beat (R). The combined method of frequency analysis and Shannon entropy is applied to extract appropriate statistical features. Information gain criterion is employed to select features that the results show that 10 highly effective features can obtain performance measures comparable to those obtained by using the complete features. The selected features are then fed to the input of Random Forest, K-Nearest Neighbour, and J48 for classification. To evaluate classification performance, tenfold cross validation is used to verify the effectiveness of our method. Experimental results show that Random Forest classifier demonstrates significant performance with the highest sensitivity of 98.1%, the specificity of 99.5%, the precision of 98.1%, and the accuracy of 98.08%, outperforming other representative approaches for automated cardiac arrhythmia classification.

KEYWORDS

cardiac arrhythmia, classification, combined frequency analysis and Shannon entropy (CFASE), information gain (IG), wearable electrocardiograms (ECG)

1 | INTRODUCTION

According to World Health Organization statistics, the global death is more than 50 million people and nearly 20 million people die because of cardiovascular diseases (CVDs; World Health Organization, 2017). CVDs are the serious diseases and increasing mortality worldwide especially in developing countries may threaten to human life (Weiwei et al., 2016). Cardiac arrhythmia disease is an imperative group of diseases in CVDs, which can lead to abnormal heartbeats and sudden death such as premature contraction and bundle branch beats (Theis & Meyerbäse, 2002). The circumstances are prompting researchers to investigate the method of classification for cardiac arrhythmias.

Electrocardiogram (ECG) as a vital tool for heart status diagnosis contains a wealth of heart rhythm and physiological information (Bie et al., 2017). Research studies develop in automatic ECG diagnosis have made positive contributions to the timely detection and better management of cardiac arrhythmias. However, traditional clinical ECG signals are long and impossible to be diagnosed in real time. Body sensor networks have opened up a great door for medical application and real-time monitoring (Zhang, Sun, Song, & Cao, 2014). ECG monitoring is a prodigious way of diagnosing CVD. Therefore, it is crucial to develop real-time ECG monitoring system in advance. To realize the real-time medical analysis, the wearable ECG signals are digitalized and transmitted to a smartphone via Bluetooth (Hu, Shao, & Tan, 2012). Meanwhile, to comprehend ubiquitous monitoring with low-cost wearable ECG devices, a lightweight classification algorithm for cardiac arrhythmia is of great importance.

Substantial methods for automated classification of cardiac arrhythmia have been proposed in recent years. These methods for classifying cardiac arrhythmia use a variety of features, including time-domain features, frequency-domain features, time-frequency domain features, and morphological features (Bogovski, 2012; Huang, Liu, Zhu, Wang, & Hu, 2014; Lin & Yang, 2014; Jatmiko, Nulad, Matul, Setiawan, & Mursanto, 2011; Ince, Kiranyaz, & Gabbouj, 2009; Übeyli, 2007; Giri et al., 2013; Jayachandran, Joseph, & Acharya, 2010; Li, Yuan, Ma, Cui, & Cao, 2017; Qin, Li, Zhang, Yue, & Liu, 2017). Bogovski et al. presented a method that used time-domain features and support vector machine (SVM) to classify five types of ECG heartbeats (Bogovski, 2012). Random projection with SVM and RR interval features were proposed by Huang et al. to classify five types of ECG heartbeats (Huang et al., 2014). A method based on linear discriminant classification and RR interval features was applied to classify five types of ECG heartbeats by Lin and Yang (2014). ECG time-domain features is acquired easily for classification; however, it is susceptible to external interference. Wavelet transform (WT) can analyse the signal and makes feature extraction in the time-frequency domain. The ECG signals can be decomposed for obtaining a multiresolution representation. Jatmiko et al. utilized back-propagation and fuzzy-neuro learning vector quantization as the classifier and WT coefficients to classify four types of ECG heartbeats (Jatmiko et al., 2011). Ince et al. proposed a method that combined discrete WT (DWT), principal component analysis (PCA), and multidimensional particle swarm optimization to classify five types of ECG heartbeats (Ince et al., 2009). Feature extraction using DWT and classification using SVM were combined for four types of ECG heartbeats classification by Übeyli (2007). Giri et al. utilized DWT, PCA combined with SVM and K-Nearest Neighbours (KNN) to classify two types of ECG heartbeats (Giri et al., 2013). A method based on DWT and entropy was used to classify two types of ECG heartbeats by Jayachandran et al. (Jayachandran et al., 2010). Genetic algorithm (GA) and the back propagation neural network (BPNN) were employed to classify six types of ECG heartbeats by Li H et al. (Li et al., 2017). Q. Qin et al. used low-dimensional wavelet features and SVM for six types of ECG heartbeats classification (Qin et al., 2017). From the aforementioned studies, it can be seen that the computation become more complex and costly for obtaining the high accuracy. So a lightweight algorithms is demand to be proposed for wearable ECG classification.

In this paper, the motivation is to develop a system with low computational cost and best performance. Selection of features and classifiers is vital to realizing the lightweight detection algorithm. First of all, 24 features are extracted using DWT method in the five frequency subband. Then information gain (IG) criterion as a feature selection method is used to select the dominant features. At last, the selected features are then fed to the Random Forest (RF), KNN, and J48 for classification. The ECG heartbeats from different patients are used for the experiment, and the results show that our proposed approach would be a good solution for the diagnosis of CVD. Figure 1 shows the block diagram of the proposed method. The contributions of the proposed method include the following:

1. We propose the combined frequency analysis and Shannon entropy (CFASE) approach to extract 24 effective features for cardiac arrhythmia classification.

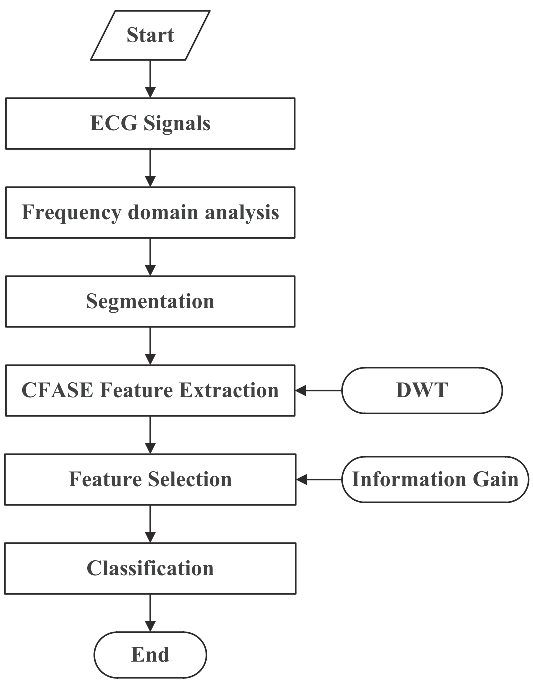

FIGURE 1 The block diagram of the proposed method

2. In wearable ECG monitoring, for the purpose of reducing computational complexity, we use IG to select most important features without losing classification accuracy.

The other sections are structured as follows: In Section 2 the materials are described in detail. Section 3 presents the proposed method including frequency domain analysis, segmentation, CFASE feature extraction, feature selection, and performance measures. In Section 4, the explanation of experiment and results are provided. Conclusion is drawn in Section 5.

2 | MATERIALS

In this study, ECG signals are obtained from the Massachusetts Institute of Technology-Beth Israel Hospital (MIT-BIH) arrhythmia database (Moody & Mark, 2002) where 48 ECG records from 47 subjects are included. Each record contains a 30-min ECG signal, and signals are sampled at 360 Hz with 11-bit resolution over the 10-mV range. There is an annotation file associated with each record, giving the interpretation for each heartbeat. In this work, five types of ECG heartbeats from the MIT-BIH arrhythmia database for one lead MLII from 10 subjects are collected, namely, normal beat (N), atrial premature contraction beat (A), premature ventricular contraction beat (V), left bundle branch block beat (L), and right bundle branch block beat (R). The normal beats dominate the MIT-BIH dataset, so we choose a part with normal beats orderly to avoid bias in the experiment result. Similarly, the remaining four types of beats we chose appropriately to avoid bias toward any one of the type. Details of data information are listed in Table 1. The specific information of the patients is as follows:

3 | METHODOLOGY

3.1 | Frequency domain analysis

The raw ECG signals are introduced high-frequency noise caused by power line interference and low-frequency noise, for example, baseline wander. At the beginning of the experiment, we design Butterworth filter to remove high-frequency component and low-frequency component. However, the ECG signals' characteristics are removed as well especially when eliminating the high-frequency noise. Thus, in this work, WT is utilized to analyse the component of frequency subbands and further remove the noise (Mallat, 1989; Li et al., 2018). First of all, the Daubechies-5 (db5) mother wavelet is adopted to decompose the signal into nine high-frequency subbands and one low-frequency subband as shown in Table 2. Moreover, we get rid of the top three high-frequency sub-bands and that the remaining wavelet coefficient subbands of fourth, fifth, sixth, seventh, eighth, and ninth levels are used to reconstruct by using wavelet inverse transform.

3.2 | Segmentation

According to the annotation file, we segregate the continuous ECG signal into individual heartbeats. First of all, the location of the R-peak as a reference point of different types of heartbeats is labelled. Then use R-peak to take 99 points forward and 200 points backward to form a sample (Pan & Tompkins, 1985). The ECG beat of 300 points that we obtain is shown in Figure 2, which demonstrates the characteristic of the ECG heartbeats in the time domain. The ECG heartbeats can be seen apparently, and the morphology of different types of heartbeats is very similar. The four subfigures in the first row of Figure 2 describe the normal signals, and the four subfigures in the second row of Figure 2 describe atrial

TABLE 1 Description of the ECG data in this work

Type of ECG beat	Number of beats	Name of record
Normal beat (N)	1,000	100, 101
Atrial premature contraction (A)	1,000	209, 232
Premature ventricular contraction (V)	1,000	106, 200
Left bundle branch block beat (L)	1,000	109, 111
Right bundle branch block beat (R)	1,000	118, 124

Abbreviation: ECG: electrocardiogram.

Record 100 (male; age 69; Medications: Aldomet, Inderal); Record 101 (female; age 75; Medications: Diapres); Record 106 (female; age 24; Medications: Inderal); Record 109 (male; age 64; Medications: Quinidine); Record 111 (female; age 47; Medications: Digoxin, Lasix); Record 118 (male; age 69; Medications: Digoxin, Norpace); Record 124 (male; age 77; Medications: Digoxin, Isordil, Quinidine); Record 200 (male; age 64; Medications: Digoxin, Quinidine); Record 209 (male; age 62; Medications: Aldomet, Hydrodiuril, Inderal); Record 232 (female; age 76; Medications: Aldomet, Inderal).

TABLE 2 Frequency band of ECG signal using nine-level decomposition

Decomposed signal	Decomposition level	Frequency band (Hz)
d1	1	180–360
d2	2	90–180
d3	3	45–90
d4	4	22.5–45
d5	5	11.25–22.5
d6	6	5.625–11.25
d7	7	2.8125–5.625
d8	8	1.40625–2.8125
d9	9	0.703125–1.40625
a9	9	0–0.703125

Abbreviation: ECG: electrocardiogram.

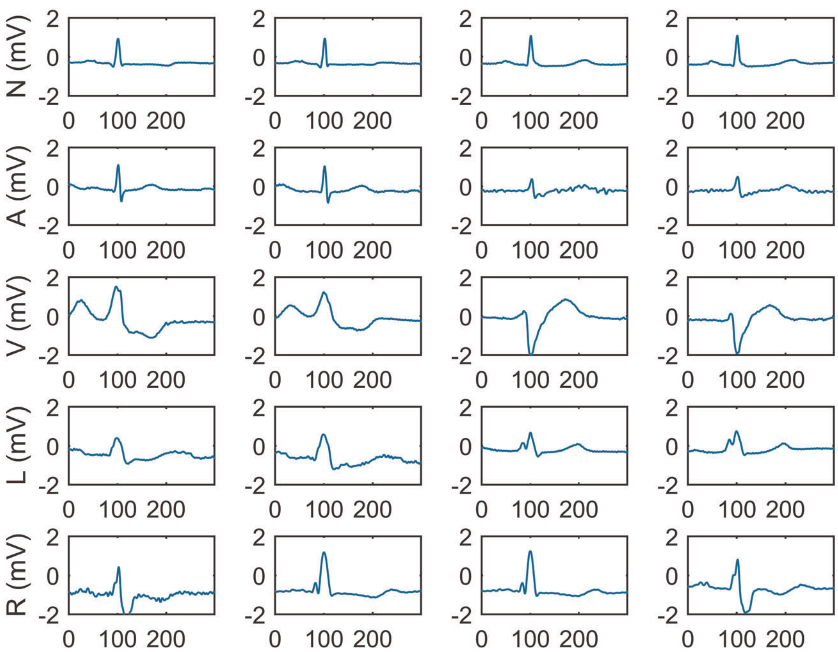

FIGURE 2 Segmented heartbeats from N, A, V, L, and R with 300 points (0.83 s)

premature contraction signals, etc. Therefore, selecting appropriate and effective features is particularly important for the different types of ECG heartbeats classification (Osowski & Linh, 2002).

3.3 | CFASE feature extraction

Biomedical signals such as ECG signals of the human body are nonlinear and nonstationary. Because of the nonstationary nature, Fourier transform is not suitable for analysing the ECG signals; subsequently, WT is proposed (Lai & Tsai, 2010; Abo-Zahhad, Al-Ajlouni, Ahmed, & Schilling, 2013). WT is a method that could be capable of analysing the ECG signals from the time domain and frequency domain simultaneously with multiresolution analysis of the function. WT consists of continuous WT and discrete WT (DWT). In many aspects, DWT can be used for the ECG signal feature extraction (Mohamed & Deriche, 2014; Hazarika, Chen, Tsoi, & Sergejew, 1997). DWT is defined as follows:

$$\mathrm{DWT}_x(j, k) = \int_R x(t)\overline{\psi_{j,k}(t)}dt, \tag{1}$$

$$\psi_{j,k}(t) = \frac{1}{\sqrt{2^j}}\psi\left(\frac{t-2^jk}{2^j}\right), \tag{2}$$

where $\psi_{j,k}(t)$ is a wavelet function, the scaling is parameter $a_j = 2^j$, and the translation parameter is$b_{j,k} = 2^jk$.

In this work, DWT is exploited to extract the statistical features. Each individual ECG heartbeat is divided into five subbands by applying Daubechies-2 mother wavelet (db2), Daubechies-4 mother wavelet (db4), and Daubechies-6 mother wavelet (db6). Feature extraction is within QRS complex frequency range from the detail wavelet coefficients (d^k,k = 1,2,3,4,5) and the approximation wavelet coefficients (a^5). Table 3 illustrates the corresponding frequencies to different levels of DWT for the ECG data. CFASE analysis method is employed to extract frequency features. A total of 18 frequency features are extracted by applying the frequency analysis, namely, median (Wmed), mean (Wmean), and standard deviation (Wstd) from DWT coefficients for each subband. Details of frequency features are shown in Table 4.

Meanwhile, Shannon entropy (SE) analysis is applied in this work. The SE equation solves the problem of quantitative measurement of information. The definition of the formula is as follows:

$$\mathrm{SE} = -\sum_i s_i^2 \log_2 (s_i)^2, \tag{3}$$

where s is the signal and S_i is the coefficients of s in an orthonormal basis. The SE value is calculated from DWT coefficients for each subband. Six features of SE are obtained from each level (d1, d2, d3, d4, d5, and a5) that presents in Table 4. By using the CFASE feature extraction method, a total number of 24 statistic features are achieved.

3.4 | Feature selection

To reduce the computational cost, feature selection is essential, which is generally perpetrated after the quantification of the important features, and how to quantify the important feature becomes the biggest difference among the various approaches. In order to select effective features, IG criterion is adopted for the feature selection (Rahman et al., 2015). The measure of the importance about IG is to see how much information the feature can bring to the classification system, the more information it conveys, the more important. For each feature, the IG measures how much information is obtained about the heartbeat-class when the value of the feature is acquired. Equation 4 gives the entropy of Y, and Equation 5 gives the entropy of Y after observing X.

$$H(Y) = -\sum_{y \in Y} p(y) \log_2 p(y), \tag{4}$$

$$H(Y|X) = -\sum_{x \in X} p(x) \sum_{y \in Y} P(y|x) \log_2 p(y|x), \tag{5}$$

where X and Y are discrete random variables.

IG is the amount measured by the entropy of Y decreases that reflects the additional information of Y given by X (Quinlan, 1993). IG is defined as follows:

TABLE 3 Frequency band of ECG signal using five-level decomposition

Decomposed signal	Decomposition level	Frequency band (Hz)
d1	1	22.85–45
d2	2	11.77–22.85
d3	3	6.24–11.77
d4	4	3.47–6.24
d5	5	2.09–3.47
a5	5	0.70–2.09

Abbreviation: ECG: electrocardiogram.

TABLE 4 Description of the features in this work

No	Extracted features in each level
1	Median of the wavelet coefficients (Wmed)
2	Mean of the wavelet coefficients (Wmean)
3	Standard deviation of the wavelet coefficients (Wstd)
4	Shannon entropy of the wavelet coefficients (SE)

$$IG = H(X) - H(X|Y) = H(Y) - H(Y|X) \\ = H(X) + H(Y) - H(X, Y). \tag{6}$$

For a given value of the feature, it can be computed the difference between the unconditional entropy of the heartbeat-type and the conditional entropy of the heartbeat-type. After computing the IG for each feature, the one least-informative feature is deleted and repeated the tenfold cross-validation experiment. We continue reducing the procedure gradually deleting one feature at a time until we obtain the decline performance. The details of the selected 10 features that are gained by using IG are explained in Section 4.

3.5 | The performance measures

To evaluate the performance of the heartbeat classification, several performance measures have been used, namely, *sensitivity*, *specificity*, *precision*, and *accuracy* (Powers, 2011). Tenfold cross validation is selected to evaluate the performance measurement. For the tenfold cross validation, the entire dataset will be divided into 10 subsets. Every time, a single subset is used for testing the model and nine subsets are used for training. The cross-validation process is needed to repeat 10 times. At last, average measures of all folds are calculated. These performance measures are defined below:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\%, \tag{7}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100\%, \tag{8}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100\%, \tag{9}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}} \times 100\%, \tag{10}$$

where TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

4 | EXPERIMENT AND RESULTS

4.1 | Experiment

The proposed model has been trained and tested with the ECG dataset on PC workstation with 2.50-GHz CPU and 8-GB RAM. In our work, the training delay was up to 3.58 s for the RF classifier, 0.004 s for KNN classifier, and 0.14 s for J48 classifier when using the selected 10 features and db6.

In this section, the experiment and results based on the proposed method are described in the details. The first step is frequency domain analysis, and through this step, the raw ECG signals become clean. Afterwards, the denoised ECG signals would be divided into individual heartbeats. Then for extracting significant features, the CFASE feature extraction method is used to complete this step. In this approach, the total 24 features are extracted from db2, db4, and db6, which described in section 3.3. The 10 selected features by IG are presented in Table 5. By employing classifiers RF, K-NN, and J48 classifier we compare the performance of the features extracted from three mother wavelets db2, db4, and db6

TABLE 5 The features are selected using information gain method

Level	Features
d1	SE1, Wstd1
d2	SE2, Wstd2
d4	Wmean4
d5	SE5, Wmean5, Wstd5
a5	SE6, Wstd6

on aspects of sensitivity, specificity, precision, and accuracy, as shown in Tables 6–8. In addition, a confusion matrix is used to analyse the results (Pal, 2005; Zhang & Zhou, 2007; Townsend, 1971).

4.2 | Results

It is observed from Tables 6 to 8 that db6 obtains the best performance compared with db2 and db4 using the extracted features and RF classifier, which provides the highest accuracy of 98.34%, sensitivity of 98.3%, specificity of 99.6%, and precision of 98.3%.

Accuracy with varying number of features in different classifiers based on db6 is shown in Figure 3. It can be realized that the best performance is achieved by 10 features approximately. In addition, the classification accuracy of RF and KNN is better than that of the J48.

The features in Table 9 are extracted by db6, and 10 features are selected using IG. It is observed from Table 9 that db6 combined with RF classifier obtained the best performance, and it provided an average accuracy of 98.08%, average sensitivity of 98.1%, average specificity of 99.5%, and average precision of 98.1%.

Detailed analysis of the relationship between performance measures and feature number using RF, KNN, and J48 based on db6 is indicated in Figures 4–6. These results illustrate the necessity of the feature selection. When we select 10 features, the values of both sensitivity and precision are around 98% and it almost invariable. Meanwhile, the performance measure of specificity gain 99.6% lightly changed with more than five selected features. The above analysis demonstrates that it is the number of 10 features that is appropriate using db6 and RF.

In addition, performance measurement of the classifier can be evaluated using confusion matrices. Confusion matrices are calculated for the proposed method with the different classifiers (RF, KNN, and J48), which is depicted in Table 10. In confusion matrix, diagonal elements represent the appropriately classified ECG heartbeats. From this table, we can observe that db6 mother wavelet with RF classifier gained the best performance. The classification results describe above show that we can obtain high-performance outcomes while identifying patients with cardiac arrhythmia by classifying individual ECG heartbeats using a set of 24 features.

TABLE 6 The classification results obtained for db2 when the ECG heartbeats were classified using RF, KNN, and J48

Classifier	Sensitivity	Specificity	Precision	F1-score	ROC area	Accuracy
RF	0.98	0.995	0.981	0.98	0.999	0.9804
KNN	0.973	0.993	0.973	0.973	0.983	0.973
J48	0.95	0.987	0.95	0.95	0.975	0.9498

Abbreviations: db2: Daubechies-2 mother wavelet; ECG: electrocardiogram; KNN: K-Nearest Neighbours; RF: Random Forest; ROC: receiver operating characteristic.

TABLE 7 The classification results obtained for db4 when the ECG heartbeats were classified using RF, KNN, and J48

Classifier	Sensitivity	Specificity	Precision	F1-score	ROC area	Accuracy
RF	0.98	0.995	0.98	0.98	1	0.9798
KNN	0.972	0.993	0.972	0.972	0.983	0.9722
J48	0.966	0.991	0.966	0.966	0.982	0.9656

Abbreviations: db4: Daubechies-4 mother wavelet; ECG: electrocardiogram; KNN: K-Nearest Neighbours; RF: Random Forest; ROC: receiver operating characteristic.

TABLE 8 The classification results obtained for db6 when the ECG heartbeats are classified using RF, KNN, and J48

Classifier	Sensitivity	Specificity	Precision	F1-score	ROC area	Accuracy
RF	0.983	0.996	0.983	0.983	1	0.9834
KNN	0.975	0.994	0.975	0.975	0.984	0.9748
J48	0.968	0.992	0.968	0.968	0.981	0.9678

Abbreviations: db6: Daubechies-6 mother wavelet; ECG: electrocardiogram; KNN: K-Nearest Neighbours; RF: Random Forest; ROC: receiver operating characteristic.

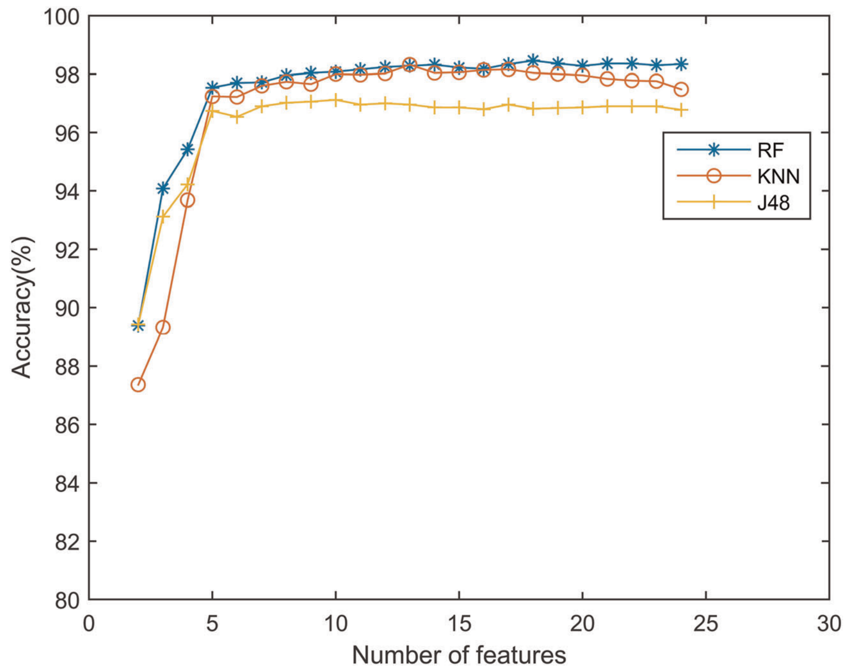

FIGURE 3 Accuracy with varying number of features in different classifiers

TABLE 9 The performance statistics are obtained from RF, KNN, and J48 classifier using db6 wavelet

db6-RF							
Class	**Sensitivity**	**Specificity**	**Precision**	**F1 score**	**ROC area**	**Sensitivity**	**Specificity**
N	0.987	0.995	0.979	0.983	0.999	0.991	0.994
A	0.969	0.997	0.99	0.979	0.999	0.969	0.997
V	0.968	0.993	0.973	0.97	0.998	0.957	0.996
L	0.992	0.994	0.976	0.984	1	0.992	0.994
R	0.988	0.996	0.986	0.987	0.999	0.991	0.994
Average	0.981	0.995	0.981	0.981	0.999	0.98	0.995
Accuracy			0.9808				

db6-KNN					**db6-J48**		
Precision	**F1-score**	**ROC area**	**Sensitivity**	**Specificity**	**Precision**	**F1 score**	**ROC area**
0.977	0.984	0.993	0.98	0.994	0.978	0.979	0.987
0.987	0.978	0.983	0.965	0.992	0.97	0.967	0.98
0.983	0.97	0.976	0.952	0.99	0.961	0.956	0.975
0.977	0.985	0.993	0.975	0.993	0.974	0.975	0.987
0.976	0.984	0.993	0.984	0.993	0.973	0.979	0.992
0.98	0.98	0.988	0.971	0.993	0.971	0.971	0.984
0.98				0.9712			

Abbreviations: db6: Daubechies-6 mother wavelet; ECG: electrocardiogram; KNN: K-Nearest Neighbours; RF: Random Forest; ROC: receiver operating characteristic.

Table 11 presents the comparison of different methods which use the same MIT-BIH database. The selected 10 effective statistical features are obtained by using the CFASE method and IG. The experimental results show the accuracy of 98.08% for the five types of ECG heartbeat classification by using our proposed method, and it could be an upright method for ECG heartbeat classification.

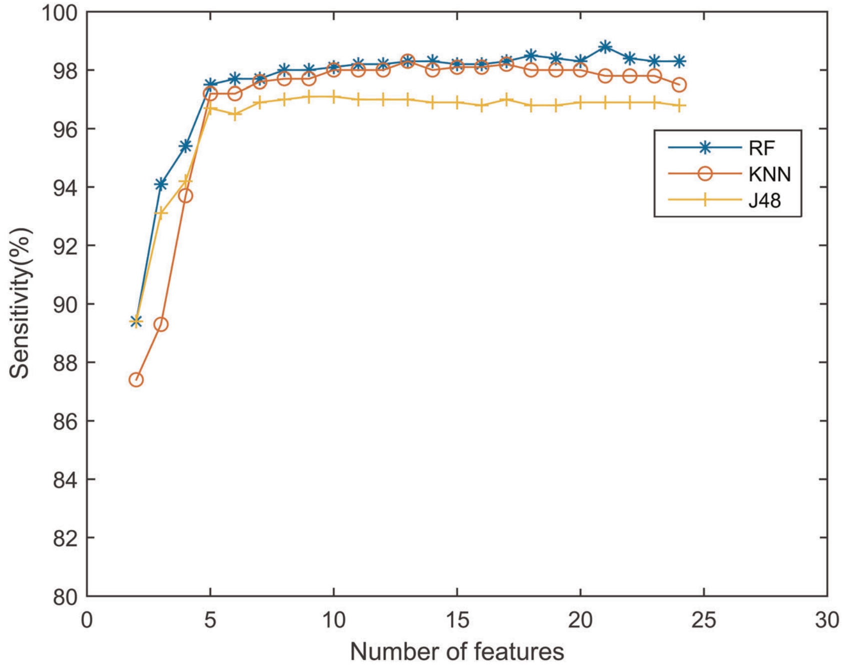

FIGURE 4 Sensitivity with varying number of features in different classifiers

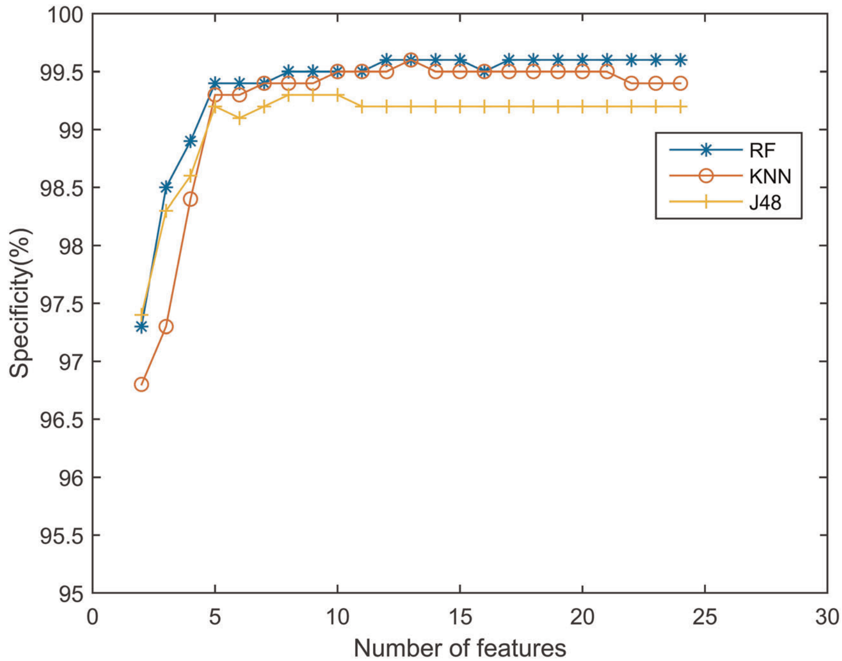

FIGURE 5 Specificity with varying number of features in different classifiers

4.3 | Discussion

The target of this work is to assess the performance of CFASE feature extraction and dimensionality reduction method IG for classification of five ECG heartbeats types. In our approach, DWT is selected to analyse the ECG signals and DWT is suitable to investigate the nonlinear and nonstationary signals. Meanwhile, WT denoising is able to protect useful signal spikes and mutation signals that distinguish high-frequency noise from high-frequency information effectively. To quantify the important features, IG criterion is the best fit to choose a reduced set of features because it can compute the features with more information and select the more appropriate features in the experiment. For performance measures, the results of using 10 features were similar to that of using 24 features. The features extraction, classification methods, and accuracy in previous works and present study are as illustrated in Table 11. As shown in Table 11, the proposed method exhibited higher accuracy value than methods proposed in previous studies.

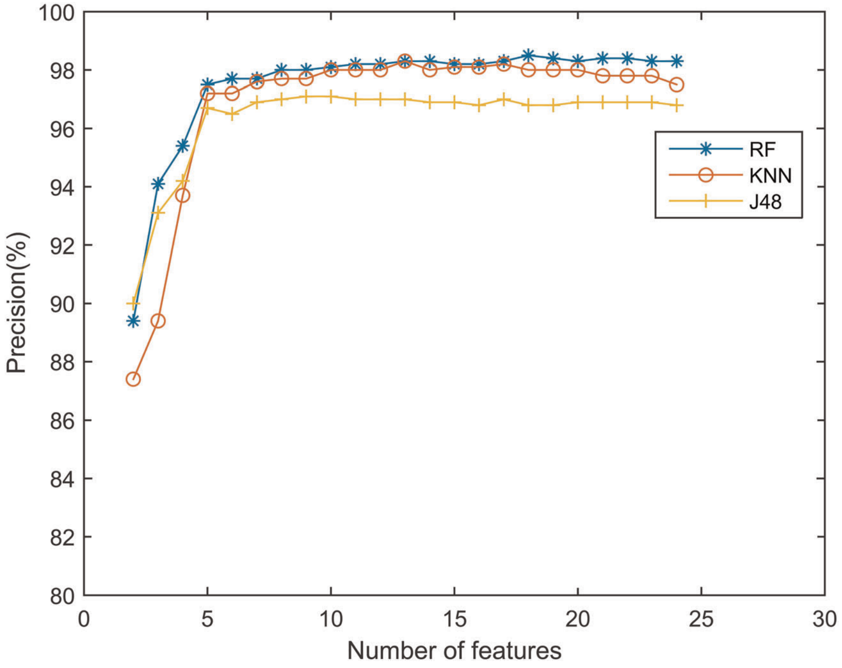

FIGURE 6 Precision with varying number of features in different classifiers

TABLE 10 Confusion matrices are obtained with different classifiers (RF, KNN, and J48) for ECG heartbeats classification

	db6-RF					db6-KNN					db6-J48				
	N	A	V	L	R	N	A	V	L	R	N	A	V	L	R
N	987	6	5	1	1	991	3	2	2	2	980	12	6	0	2
A	16	969	6	9	0	17	969	4	10	0	17	965	8	10	0
V	5	2	968	13	12	6	5	957	10	22	5	6	952	14	23
L	0	2	5	992	1	0	4	4	992	0	0	12	11	975	2
R	0	0	11	1	988	0	1	7	1	991	0	0	14	2	984

Abbreviations: db6: Daubechies-6 mother wavelet; ECG: electrocardiogram; KNN: K-Nearest Neighbours; RF: Random Forest.

TABLE 11 Compare with other methods based on the MIT-BIH arrhythmia database.

Work	Class	Features	Classifier	Accuracy (%)
Bogovski, P et al.	5	Waveform features	LIBSVM	95.21
Huang et al.	5	RR-intervals	SVM	94
Lin et al.	5	Normalized RR-intervals	LD	93
Jatmiko et al.	4	Wavelet transform coefficients	FLVQ	95.50
Ince et al.	5	Waveform features	Multidimensional PSO	95.28
Li H et al.	6	Statistical features of WPCs	GA-BPNN	97.78
Proposed method	5	CFASE features of DWT coefficients	RF	98.08

Abbreviations: CFASE: combined method of frequency analysis and Shannon entropy; DWT: discrete wavelet transform; FLVQ: fuzzy-neuro learning vector quantization; GA-BPNN: genetic algorithm-back propagation neural network; LD: linear discriminant; MIT-BIH: Massachusetts Institute of Technology-Beth Israel Hospital; PSO: particle swarm optimization; RF: Random Forest; SVM: support vector machine; WPCs: wavelet packet coefficients.

In addition, there remain certain limitations. For instance, more types of heartbeats could be put into consideration for fully understanding of the heart rhythm. Meanwhile, the number of samples also needs extension for exploring deeper or hidden features. However, the proposed method may not be efficient enough in dealing with large dataset. Artificial neural network (ANN) exhibits great flexibility and adaptability to handle large amounts of data (Jadhav, Nalbalwar, & Ghatol, 2010). Compared with conventional methods, ANN performs obvious advantages in dealing with fuzzy data, random data, and nonlinear data. Therefore, it can be a good alternative for large-scale ECG data analysis in future work.

There is a possibility of involving the same patient information in both training and testing data that not applying subject-based validation scheme. Another significant limitation is not analysing ECG fragments in more leads.

5 | CONCLUSION

We have proposed an approach for classifying five types ECG heartbeats, namely, N, A, V, L, and R, which come from MIT-BIH arrhythmia database. The db5 mother wavelet is used for signal frequency domain analysis that the signal is denoised and became clean. CFASE feature extraction method combined with db6 mother wavelet is used for the feature extraction. A comprehensive of 24 features are extracted for the heartbeat classification, and the dimension of the full features is reduced by IG method to obtain highly informative features. The proposed method has yielded the highest sensitivity of 98.1%, the specificity of 99.5%, and the precision of 98.1% and the accuracy of 98.08%. It makes contribution to reduce the computational cost and improves the classification efficiency.

In the future, we plan to increase the classification accuracy based on the subject training scheme: (a) increase the number of samples and analyse various types ECG heartbeats from different databases, (b) extract more number of suitable features for ECG signal classification, and (c) using deep learning methods on classification for more different types of ECG heartbeats.

ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China under grant 61572231 and in part by the Shandong Provincial Key Research and Development Project under grant 2017GGX10141.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

ORCID

Yuan Zhang https://orcid.org/0000-0003-2726-2855
Benny Lo https://orcid.org/0000-0002-5080-108X
Wenyao Xu https://orcid.org/0000-0001-6444-9411

REFERENCES

Abo-Zahhad, M., Al-Ajlouni, A. F., Ahmed, S. M., & Schilling, R. J. (2013). A new algorithm for the compression of ecg signals based on mother wavelet parameterization and best-threshold levels selection. *Digital Signal Processing*, *23*(3), 1002–1011. https://doi.org/10.1016/j.dsp.2012.11.005

Bie, R., Zhang, G., Sun, Y., Xu, S., Li, Z., & Song, H. (2017). Smart assisted diagnosis solution with multi-sensor holter. *Neurocomputing*, *220*, 67–75. https://doi.org/10.1016/j.neucom.2016.06.074

Bogovski, P. (2012). Analysis and classification of cardiac arrhythmia using ECG signals. *International Journal of Computer Applications*, *38*(1), 37–40. https://doi.org/10.5120/4575-6742

Giri, D., Acharya, U. R., Martis, R. J., Sree, S. V., Lim, T. C., Thajudin Ahamed, V. I., & Suri, J. S. (2013). Automated diagnosis of coronary artery disease affected patients using lda, pca, ica and discrete wavelet transform. *Knowledge-Based Systems*, *37*(2), 274–282. https://doi.org/10.1016/j.knosys.2012.08.011

Hazarika, N., Chen, J. Z., Tsoi, A. C., & Sergejew, A. (1997). Classification of EEG signals using the wavelet transform. *Signal Processing*, *59*(1), 61–72. https://doi.org/10.1016/S0165-1684(97)00038-8

Hu, S., Shao, Z., & Tan, J. (2012). A real-time cardiac arrhythmia classification system with wearable electrocardiogram. *Sensors*, *12*(9), 12844–12869. https://doi.org/10.3390/s120912844

Huang, H., Liu, J., Zhu, Q., Wang, R., & Hu, G. (2014). A new hierarchical method for inter-patient heartbeat classification using random projections and RR intervals. *Biomedical Engineering Online*, *13*(1), 90. https://doi.org/10.1186/1475-925X-13-90

Ince, T., Kiranyaz, S., & Gabbouj, M. (2009). A generic and robust system for automated patient-specific classification of ECG signals. *IEEE Transactions on Biomedical Engineering*, *56*(5), 1415–1426. https://doi.org/10.1109/TBME.2009.2013934

Jadhav, S. M., Nalbalwar, S. L., & Ghatol, A. (2010). Artificial neural network based cardiac arrhythmia classification using ECG signal data. Electronics and Information Engineering (ICEIE), 2010 International Conference On. IEEE. doi:https://doi.org/10.1109/ICEIE.2010.5559887

Jatmiko, W., Nulad, W. P., Matul, I. E., Setiawan, I. M. A., & Mursanto, P. (2011). Heart beat classification using wavelet feature based on neural network. *Wseas Transactions on Systems*, *10*(1), 17–26.

Jayachandran, E. S., Joseph, K. P., & Acharya, U. R. (2010). Analysis of myocardial infarction using discrete wavelet transform. *Journal of Medical Systems*, *34*(6), 985–992. https://doi.org/10.1007/s10916-009-9314-5

Knvps, R., & Dhuli, R. (2017). Classification of ECG heartbeats using nonlinear decomposition methods and support vector machine. *Computers in Biology and Medicine*, *87*, 271–284. https://doi.org/10.1016/j.compbiomed.2017.06.006

Lai, C. C., & Tsai, C. C. (2010). Digital image watermarking using discrete wavelet transform and singular value decomposition. *IEEE Transactions on Instrumentation and Measurement*, *59*(11), 3060–3063. https://doi.org/10.1109/TIM.2010.2066770

Li, H., Yuan, D., Ma, X., Cui, D., & Cao, L. (2017). Genetic algorithm for the optimization of features and neural networks in ECG signals classification. *Scientific Reports*, *7*, 41011. https://doi.org/10.1038/srep41011

Li, J. Q., Li, R., Chen, Z. Z., Deng, G. Q., Wang, H., & Mavromoustakis, C. X. (2018). Design of a continuous blood pressure measurement system based on pulse wave and ecg signals. *IEEE Journal of Translational Engineering in Health and Medicine*, *6*, 1–14. https://doi.org/10.1109/JTEHM.2017.2788885

Lin, C. C., & Yang, C. M. (2014). Heartbeat classification using normalized RR intervals and wavelet features. *International Symposium on Computer*, *2014*, 650–653. https://doi.org/10.1109/IS3C.2014.175

Mallat, S. G. (1989). A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *11*(7), 674–693. https://doi.org/10.1109/34.192463

Mohamed, M. A., & Deriche, M. A. (2014). An approach for ECG feature extraction using daubechies 4 (db4) wavelet. *International Journal of Computer Applications*, *96*(12), 36–41. https://doi.org/10.5120/16850-6712

Moody, G. B., & Mark, R. G. (2002). The impact of the MIT-BIH arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine*, *20*(3), 45–50. https://doi.org/10.1109/51.932724

Osowski, S., & Linh, T. H. (2002). Ecg beat recognition using fuzzy hybrid neural network. *IEEE Transactions on Biomedical Engineering*, *48*(11), 1265–1271. https://doi.org/10.1109/10.959322

Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, *26*(1), 217–222. https://doi.org/10.1080/01431160412331269698

Pan, J., & Tompkins, W. J. (1985). A real-time QRS detection algorithm. *IEEE Transactions on Biomedical Engineering*, *BME-32*(3), 230–236. https://doi.org/10.1109/TBME.1985.325532

Powers, D. M. W. (2011). Evaluation: from precision, recall and f-factor to roc, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, *2*, 2229–3981. https://doi.org/10.9735/2229-3981

Qin, Q., Li, J., Zhang, L., Yue, Y., & Liu, C. (2017). Combining low-dimensional wavelet features and support vector machine for arrhythmia beat classification. *Scientific Reports*, *7*(1), 6067. https://doi.org/10.1038/s41598-017-06596-z

Quinlan, J. (1993). *Program for machine learning*. San Mateo: Morgan Kaufmann.

Rahman, Q. A., Tereshchenko, L. G., Kongkatong, M., Abraham, T., Abraham, M. R., & Shatkay, H. (2015). Utilizing ECG-based heartbeat classification for hypertrophic cardiomyopathy identification. *IEEE Transactions on Nanobioscience*, *14*(5), 505–512. https://doi.org/10.1109/TNB.2015.2426213

Theis, F. J., & Meyerbäse, A. (2002). Biomedical signal analysis. *Leonardo*, *2*(46), 19. https://doi.org/10.1049/ic:20010049

Townsend, J. T. (1971). Theoretical analysis of an alphabetic confusion matrix. *Perception & Psychophysics*, *9*(1), 40–50. https://doi.org/10.3758/BF03213026

Übeyli, E. D. (2007). ECG beats classification using multiclass support vector machines with error correcting output codes. *Digital Signal Processing*, *17*(3), 675–684. https://doi.org/10.1016/j.dsp.2006.11.009

Weiwei, C., Runlin, G., Lisheng, L., Manlu, Z., Wen, W., Yongjun, W., et al. (2016). Outline of the report on cardiovascular diseases in china, 2014. *European Heart Journal Supplements: Journal of the European Society of Cardiology*, *18*(Suppl F), F2. https://doi.org/10.1093/eurheartj/suw030

World Health Organization (2017, July 5), Cardiovascular diseases (CVDs), [Online]. Retrieved from:http://www.who.int/mediacentre/factsheets/fs317/en.

Zhang, M. L., & Zhou, Z. H. (2007). Ml-knn: A lazy learning approach to multi-label learning. *Pattern Recognition*, *40*(7), 2038–2048. https://doi.org/10.1016/j.patcog.2006.12.019

Zhang, Y., Sun, L., Song, H., & Cao, X. (2014). Ubiquitous WSN for healthcare: Recent advances and future prospects. *IEEE Internet of Things Journal*, *1*(1), 311–318. https://doi.org/10.1109/JIOT.2014.2329462

AUTHOR BIOGRAPHIES

Yuwei Zhang received her BS in Communication Engineering in 2016 from University of Jinan, China. She is a master student majoring in computer science at Shandong Provincial Key Laboratory of Network Based Intelligent Computing, University of Jinan. She is a member of both ACM and CCF. Her current research interests include biomedical signal/image processing machine learning and pattern recognition.

Yuan Zhang received his M.S. degree in communication systems and the Ph.D. degree in control theory and engineering both from Shandong University, China, in 2003 and 2012, respectively. He is currently a Professor with the University of Jinan, China. Dr. Zhang was a visiting professor with the Computer Science Department, Georgia State University, USA, in 2014. As the first author or corresponding author he has published more than 60 peer reviewed papers in international journals and conference proceedings, 1 book chapter, and 6 patents in the areas of Smart Health and Biomedical Big Data Analytics. He has served as Leading Guest Editor for six special issues of IEEE, Elsevier, Springer and InderScience publications, including *IEEE Internet of Things Journal* special issue on Wearable Sensor Based Big Data Analysis for Smart Health

and *IEEE Journal of Biomedical and Health Informatics* (JBHI) special issue on Pervasive Sensing and Machine Learning for Mental Health. He has served on the technical program committee for numerous international conferences. He is an associate editor for IEEE Access and an Editor for the Elsevier journal Internet of Things. Dr. Zhang's research interests include wearable sensing for smart health, machine learning for auxiliary diagnosis, and biomedical big data analytics. His research has been extensively supported by the Natural Science Foundation of China, China Postdoctoral Science Foundation, and Natural Science Foundation of Shandong Province with total grant funding over 2 million RMB. Dr. Zhang is a Senior Member of both IEEE and ACM.

Dr. Benny Lo is a Senior Lecturer of the Hamlyn Centre and the Department of Surgery and Cancer, Imperial College London. Dr Lo's research mainly focuses on body sensor networks and wearable technologies for health care and well-being. Much of his research has translated for clinical applications, and he has also led a few pioneering clinical trials of novel sensing technologies, such as the clinical validation of a new spectroscopic sensor for blood perfusion monitoring for microvascular surgeries. He is a Senior Member of IEEE, an Associate Editor of the IEEE Journal of Biomedical Health Informatics, Chair of the IEEE EMBS Wearable Biomedical Sensors and Systems Technical Committee, and a member of the IEEE EMBS Standards Committee.

Wenyao Xu received the Ph.D. degree from the University of California at Los Angeles, Los Angeles, USA, and both the Master and Bachelor degree from Zhejiang University, China. Wenyao Xu is an Associate Professor with tenure of the Computer Science and Engineering Department, University at Buffalo (SUNY). His research has focused on exploring novel sensing and computing technologies to build up innovative Internet-of-Things (IoT) systems for high-impact human-technology applications in the fields of Smart Health and Cyber-Security. Results have been published in peer-reviewed top research venues across multiple disciplines. To date, his group has published over peer-reviewed 160 papers, won seven best paper awards, two best paper nominations and three international best design awards. His inventions have been filed within U.S. and internationally as patents, and have been licensed to industrial players. His research has been reported in high-impact media outlets. Currently, Dr. Xu serves as an Associate Editor of *IEEE Transactions on Biomedical Circuits and Systems*, the technical program committee of numerous conferences in the field of Smart Health and Internet of Things, and has been a Technical Program Committee co-chair of IEEE Body Sensor Networks in 2018.

How to cite this article: Zhang Y, Zhang Y, Lo B, Xu W. Wearable ECG signal processing for automated cardiac arrhythmia classification using CFASE-based feature selection. *Expert Systems*. 2019;e12432. https://doi.org/10.1111/exsy.12432