Biomedical Signal Processing and Control 41 (2018) 152–160

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Feature fusion for imbalanced ECG data analysis

Wei Lu, Honghui Hou, Jinghui Chu*

School of Electronic Information Engineering, Tianjin University, Tianjin 300072, PR China

ARTICLE INFO

Article history:
Received 19 July 2017
Received in revised form 13 October 2017
Accepted 19 November 2017

Keywords:
Electrocardiogram signals
Imbalanced data set
Feature fusion
2D-convolutional neural network
Shallow characteristic
Random forest

ABSTRACT

World Health Organization (WHO) indicates that cardiovascular disease remains challenging in diagnosis and treatment. The electrocardiogram (ECG) is a very important diagnostic assistant for cardiac diseases. Traditionally, most of the ECG analysis methods are evaluated by their intra-patient performance, which however may not suitable for inter-patient cases. Here, we propose a complete classification system with excellent generalization ability. We first extract the 2D-convolutional and PQRST features of a single heartbeat after preliminary processing. We then balance the data with the Random Over Sampler algorithm after comparing several imbalanced algorithms. Finally, we use a Random Forest (RF) classifier to classify the data according to the Association for the Advancement of Medical Instrumentation (AAMI) standards (1988). Results show that $Recall_M$ (MR), $Precision_M$ (MP) and $Fscore_M$ (MF) of our proposal are all above 99%. In order to evaluate the performance of different methods, we designed inter-patient and intra-patient experiments separately. To further demonstrate the robust and adaptability of our model, we then transferred it to another data set and performed the experiment. In our experiments, the values of macro- and micro-metrics are up to 99%. All of the results are averages of five experiments, and the Average Accuracy (AA) of experiments applied here are above 99%, which illustrates that our proposal is a promising alternative and superior to most of the state-of-the-art methods.

1. Introduction

According to the data of the World Health Organization, 30% of global deaths are due to cardiovascular diseases (CVDs). Thus, the global burden of cardiovascular is still the most important healthcare issue around the world [1]. Arrhythmias are the most typical and important cardiovascular diseases, which may cause temporary shock and even sudden death. The best assistant to monitor and diagnose the arrhythmias is ECG, which is a visual signal captured or measured by placing electrodes on the surface of the body to detect voltage changes. Cardiologists often analyzed the ECG directly in the past, whereas computer-aided classification of arrhythmias has become popular more recently [2].

A fully automatic classification of arrhythmias includes four parts: ECG processing, heartbeat segmentation, feature extraction, and classification. The goal of ECG processing is to make the signal clearer and lay the foundation for subsequent experiments, i.e. ECG denoising [3–5], detection of characteristic points [6], etc. In the second phase, the ECG is divided into single or multiple periods of heartbeats by using the heartbeats frequency information [7,8]. In the last stage, common classification methods such as Artificial Neural Network (ANN) [9], Bayesian Network (BN) [10], Random Forest (RF) [11], Support Vector Machine (SVM) [12,13], etc. have been used to obtain the true class of the samples.

Feature extraction plays an important role throughout the process, and various methods have been proposed and validated. For classical methods, features are exacted from the time domain such as R-R intervals [14,15] and QRS width [16] as well as frequency domain for example S-transform [17,18], wavelet transform [9,15,19–24], Fourier transform [20,25,26], Modified Cosine Transform [19,20], etc. After feature exaction, there would be feature selection to remove related characteristics and reduce dimensions to improve the final accuracy, generally including principal component analysis (PCA) [9,21], linear discriminant analysis (LDA) [21], decision tree (DT) [22], independent component analysis (ICA) [14,15,21], etc. However, these methods are usually used for extracting handcrafted features from the ECG waveforms, and one of issues is incomplete use of information provided by the source data. Deep learning techniques can overcome this shortcoming. Previously [12], an electrocardiogram beat classification method was proposed based on Deep Belief Networks (DBN) with features extracted by DBN and timing interval. In [27], the authors used a labeled HRV data sets to train a Convolutional Neural Networks (CNN) model as a supervised approach, and used Stacked Autoencoders with Restricted Boltzmann Machines to obtain unlabeled features. According to the literature [28,29], 1D-CNN could

* Corresponding author.
E-mail addresses: luwei@tju.edu.cn (W. Lu), slyviahou@tju.edu.cn (H. Hou), cjh@tju.edu.cn (J. Chu).

https://doi.org/10.1016/j.bspc.2017.11.010
1746-8094/

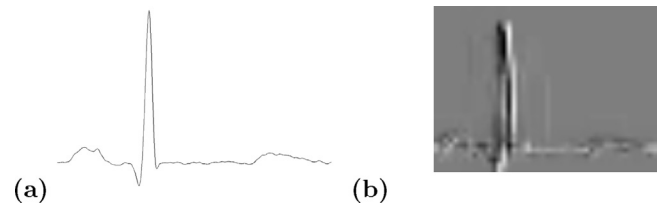

Fig. 1. The figures a and b stand for the original heartbeat and its feature map after convolution operation separately. The difference between each other is obvious.

be used for patient-specific classification with patient-specific and relatively small common data, which could save training time and be easily applied to the wearable device. An approach using Staked Denoising Autoencoders (SDAEs) with sparsity constraint for unsupervised feature representation and Active Learning (AL) to fine tuning Deep Neural Network (DNN) could efficiently match the statistical distribution of the data [30], etc.

All of the methods deal with 1D-ECG signals, and we put forward an idea of switching 1D-ECG signals into 2D-images for extracting convolutional features for the first time since others [28] proved that the simple convolutional neural network could improve the training accuracy compared with other methods. However, we found that the 2D-convolutional feature maps are blurred (Fig. 1), especially at the key points (i.e., PQRST positions) which can affect experimental results. Meanwhile, ECG data sets are extremely imbalanced, and this can reduce the recognition accuracy of the minority class. To reconcile those shortcomings, we proposed a classification system (Fig. 2). Firstly, single heartbeat signals are obtained by preprocessing and segmentation. Secondly, PQRST features are exacted from single heartbeat signals for making up the blurred issue of convolutional maps and 2D images are created by connecting the dots of the 1D single heartbeat signals. Then, the highly abstract features of a simple CNN are extracted and fused with PQRST features. After imbalanced processing, the fused features of balanced data set are classified by a simple RF classifier. This is the first time that deep features (CNN) are fused with shallow features (PQRST) for extracting representative features to the best of our knowledge although feature fusion has been reported in shallow-with-shallow [31] and deep-with-shallow [12]. To prove the generalization of the model, we use another data set to evaluate the performance of our model.

The rest of the paper is organized as follows: Section 2 introduces the main methods used in this paper involving feature extraction, imbalanced data processing, the classifier and learning rate. Section 3 explains the publicly available ECG data sets and the standards we used as well as performance metrics for experiments. Section 4 is results and discussion. Finally, conclusions and future directions are detailed in Section 5.

2. Methods

Our research concentrates on dealing with imbalanced data processing, feature fusion, parameter adjustment as well as classifier design. All of the algorithms in this section expound around these four parts.

For clearer presentation, we establish here some of the basic notions used in most of the subsections. Considering the original data set after selecting and heartbeat segmentation OS has m examples, we define: $OS = \{(x_i, y_i)\}$, $i = 1, \ldots, m$, where $x_i \in X = \{f_1, f_2, \ldots f_n\}$ is a n-dimensional instance. $y_i \in Y = \{1, 2, \ldots z\}$ is the label of x_i. Balanced data set based on original data $BOS = \{(x_{newi}, y_{newi})\}$, $i = 1, \ldots, M$ has N dimensions where M is the sum of balanced data set.

2.1. Feature fusion

2.1.1. Convolutional features

CNN is a typical method for deep learning based on its characteristics that automatically highlights and extracts the most valuable high dimensional features through convolutional operation from the input data. It pays more attention to the local features and their positions, i.e., the position among other features is determined when the local feature is extracted. Otherwise, on the same feature maps of CNN, the weights of neurons are the same resulting in network learning in parallel, which markedly saves learning time. In this paper, considering that samples belong to simple line types, we proposed a simple CNN that contains two convolutional layers, two Pooling layers, three Full-Connected (FC) layers and two Dropout layers (Fig. 3).

As explained in [27], the output of convolutional layers $C_{x,y}$ can be computed according to Eq. (1):

$$C_{(x,y)} = h\left(\sum_{i=1}^{k_m}\sum_{j=1}^{k_n} OS_{(x'+i,y'+i)} \times w_{(i,j)} + b_{(i,j)}\right), \tag{1}$$

where the two-dimensional input data is $OS_{(x,y)}$, the kernel size is (k_m, k_n) and the steps of convolution are (s_x, s_y), $w_{(i,j)}$, $b_{(i,j)}$ are weights and bias of the kernel and $x' = x \cdot s_x - 2$, $y' = y \cdot s_y - 2$.

Applying to our designed network, we used 96 * 72 two-dimensional gray images to simulate digital heartbeats obtained from MIT-BIH data set and utilized four convolutional kernels whose sizes are all $\left\{\left\{1, 1, 1\right\}, \left\{1, -7, 1\right\}, \left\{1, 1, 1\right\}\right\}$ to emphasize edge information in first layer, and eight convolutional kernels with (5, 5) size for second convolutional layer to detect the edges of the blurred feature maps of first layer. Because the texture information is more valuable than the background, we used a max-pooling method (selecting the max-value of neighborhood characteristics corresponding to the pooling window) and set $P_x = 2$, $P_y = 2$, which reduces the number of features in each feature map by four. In the fully connected part, we assembled a network with a hidden layer and set 200, 100, 5 neurons per layer. In that case, two dropout layers [32] are embedded into every two FC layers for the purpose of enhancing the network generation ability and preventing overfitting. The training and testing process of the dropout layers is presented in Eq. (2):

$$\begin{aligned} &Train: D_{(i,j)} = w_{(i,j)}\big|_p \times fc_{(i,j)} + b_{(i,j)}, \\ &Test: D_{(i,j)} = w_{(i,j)} \times fc_{(i,j)} \times P + b_{(i,j)}. \end{aligned} \tag{2}$$

Here, $fc_{(i,j)}$, $D_{(i,j)}$ are the input and output of the layer, $w_{(i,j)}\big|_p$ is the weight of selecting specific number of neurons according to retaining probability P, $w_{(i,j)}$ is the weight of all of neurons. Relu is adopted as the activation function of each layer except that softmax is used in the final classification.

The determination of the 2D-CNN consists of data feed forward pass and error back-propagation pass [35]. We used a cross-entropy cost function to adjust the network parameters, which were expressed as Algorithm 1:

Algorithm 1. Cross-entropy cost function for network parameter adjustment.

Require: $(x_1, y_1), \ldots, (x_m, y_m), \ldots, (x_n, y_n)$
Ensure: $w'_{i,j}$, $b'_{i,j}$
1: Randomly select k train samples from original data set;
2: Initialization: $w_{(i,j)} \to 0$, $b_{(i,j)} \to 0$, $lr \to r$;
3: **for** i=1 to k **do**
4: Storage target output S_i;
5: Calculate output vector $l_{(i,j)}$ of intermediate layer and actual output A_i of last layer;
6: Cost calculation:

$c = -\frac{1}{k}\sum_{x_i}[A_i \ln S_i + (1-A_i)\ln(1-S_i)]$;
Adjust weight and bias:
$w'_{(i,j)} = w_{(i,j)} + \Delta w_{(i,j)}$, $\Delta w_{(i,j)} = \frac{1}{k}\sum_x x_i(h(z) - A_i)$,
$b'_{(i,j)} = b_{(i,j)} + \Delta b_{(i,j)}$, $\Delta b_{(i,j)} = \frac{1}{k}\sum_x (h(z) - y)$.
Here, x is the input samples, z is the neuron input, $h()$ stands for activation function.
Switch $lr \rightarrow r'$ and repeat the steps above.
Until epoch equals its set point, and keep the best optimal value.
end for

2.1.2. PQRST features

As mentioned in Section 1, several basic approaches are frequently used as hand-craft features for classification. Because single-heartbeat images in our data set are blurred after CNN, we extracted PQRST features [5] as a supplement. The PQRST parameters are typical morphological features and are intercepted by three time windows of different sampling rates and start-end positions (Fig. 5).

We employed a 60 Hz time window to extract 10 points of QRS complex from $R-50$ ms to $R+100$ ms, where R is the position of the R wave of the electrocardiogram. For T wave features, we fixed our 20 Hz time window from $R+150$ ms to $R+500$ ms to get 8 feature points. At last, we extracted 7 points of P wave, whose window is from $R-200$ ms to $R-100$ ms with the 60 Hz sampling rate. Totally, 25 dimensional features are produced through above operations.

2.1.3. Feature fusion

Early fusion is an algorithm that belongs to feature-level fusion, which extracts features from each modality and concatenates these features into one large vector [34]. Based on this, we selected the features of a dense layer whose neurons are 200, and then fused the 200 dimensional CNN features with the 25 dimensional PQRST features to obtain a 225-dimensional feature vector for each sample. That is, features of both training and testing samples have 225 dimensions.

2.2. Imbalance to balance

The number of heartbeat samples from each class acquired from public database is unequal. The proportion of the majority class samples and the minority class samples is more than one hundred-fold, and this can cause minority class samples to be swallowed by the majority class. Inspired by [36] that has listed many practical methods in the two-class classification and in order to keep the same distribution of training and testing sets, we used several typical methods to balance the multi-class data set in this paper. These include Random Over Sampler (ROS), Random Under Sampler (RUS), Cluster Centroids (CC), Near Miss (NM), Edited Nearest Neighbours (ENN) [37], Repeated Edited Nearest Neighbours (RENN), Neighbourhood Cleaning Rule (NCR) [38] and One Sided Selection (OSS) [39].

Multi-class ROS belongs to over-sampling methods whose primary objective is to increase the number of the minority class. In this paper, samples are added to the four classes except for the class which has the maximum sample size. Each class increases the samples by the same operation, and as the result, each category contains the same number of samples as the majority class does.

The other methods mentioned above belong to under-sampling methods whose purpose mainly eliminates some samples of the majority class. Theoretical explanation of RUS, CC, NM is given in [35]. ENN removes samples whose most surrounding samples are different from itself. RENN is an improved algorithm based on ENN that repeats the ENN processing until there are no samples for deletion. NCR is roughly the same as ENN, but it removes only

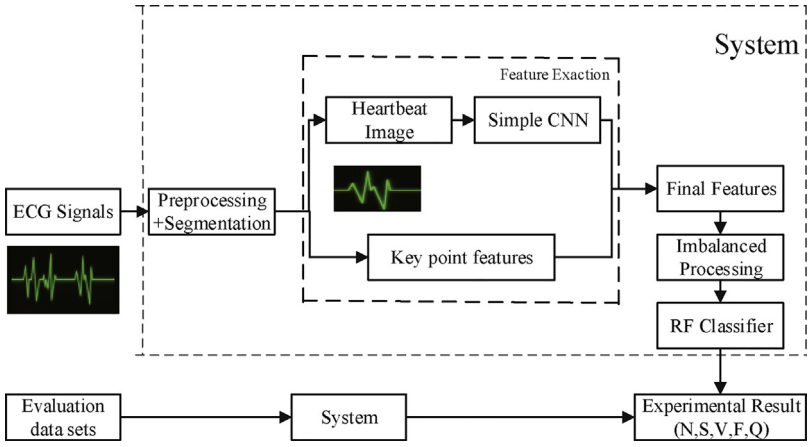

Fig. 2. Flowchart of the proposed system. Feature exaction steps are presented in the dashed box.

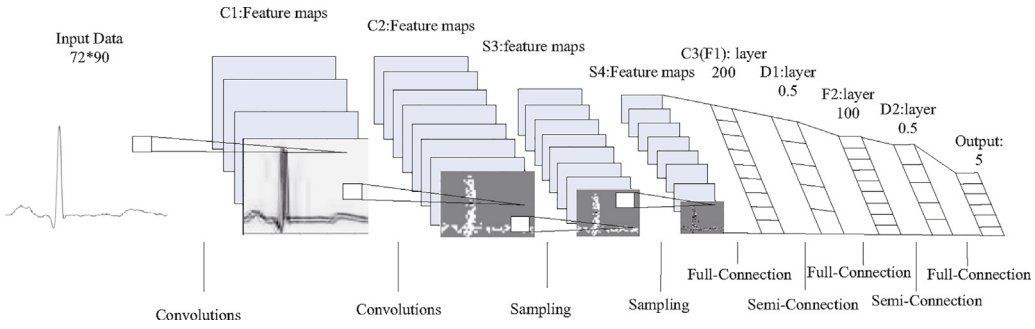

Fig. 3. The whole CNN model we proposed. The first full-connected layer is also a 1D-convolutional layer for classification. D1 and D2 both are Dropout layer with empirical value 0.5.

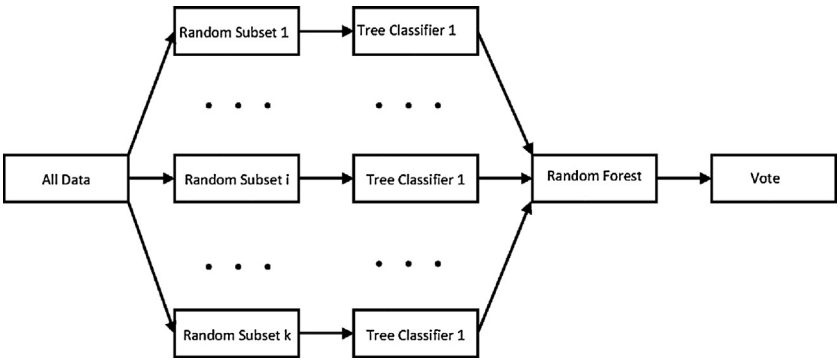

Fig. 4. The RF classifier framework.

the majority class samples whether the center is the minority most surrounding by the majority or the opposite. In addition, OSS can be expressed as

$$BOS = S_{majss} + S_{omin},$$

or

$$BOS = OS - (S_{majns} + S_{majbs} + S_{majrs}).$$

Here, S_{majns}, S_{majbs}, S_{majrs}, S_{majss} represent noise samples, boundary samples, redundant samples and safe samples respectively. Particularly, we treated the class which has a minimum number of samples as the minority class and keep it unchanged, and the rest four classes were regarded as the majority classes. Among NM methods, we employed nearmiss-2 proposed in [39] for comparison with other imbalanced algorithms.

Considering that the similarities and differences between the training set and the testing set can affect the final classification results. Besides, balanced training data set is helpful to get a model with good generalization capability and balanced testing data set can protect the minority class so as to improve the classification accuracy. We applied these algorithms to both training and testing data sets. Although these algorithms change the distribution of the original data sets somewhat, they will produce different results for the same data set even if we set the same random-state of them. Detailed applications of these are in Section 4.

2.3. Random Forest classifier

For the sake of classifying feature vectors with high accuracy, we used an easily implemented classifier named Random Forest (RF) [42], which is a large set of decision trees without pruning and the classification result is determined by voting for each tree classifier (Fig. 4).

Because of its rapid speed and high accuracy of classifying, we used RF to classify the final features which come from CNN and PQRST features.

2.4. Learning rate

Learning rate *lr* is usually used in Gradient descent algorithms for weight updates to optimize the model. It influences the range of weight adjustment seriously. To make the gradient descent method have a better performance, *lr* should be set in an appropriate range because *lr* determines the speed at which the parameter moves to the optimal value. An excessively large *lr* value may result in cost function steps over the optimal value as Fig. 7(a) shows. On the contrary, a too small *lr* value leads to low efficiency optimization or cost function divergence. We chose the relatively better *lr* with many experiments, and the results are shown in Experiment One.

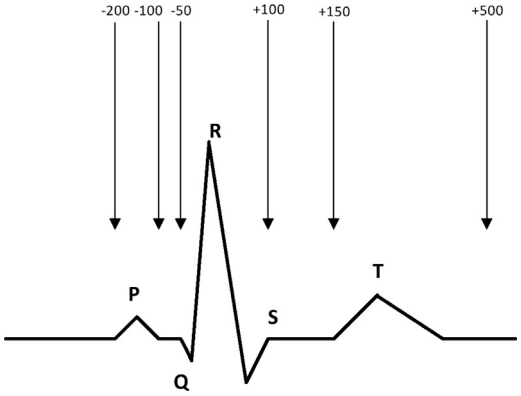

Fig. 5. The graphic annotation of the fiducial points of a heartbeat. "+" or "−" presents the number of milliseconds off from R peak.

The four main parts in this section are organized in accordance with Fig. 2 for our work. Algorithm 2 further explains the system process from acquiring ECG signal to calculating the accuracy of disease category classification.

Algorithm 2. System flow.

Require: Continuous ECG signals.
Ensure: Value of Precision_u, Recall_u, Fscore_u, Precision_M, Recall_M, Fscore_M, Error Rate and Average Accuracy.

1: *Step*1: Signal preprocessing including signal denoising and segmentation to obtain completely pure heartbeat data;
2: *Step*2: Get system features:
   *Step*2.1: Extract PQRST features of data, labeled F1;
   *Step*2.2: Extract CNN features of data, labeled F2;
   *Step*2.3: Fuse F1 with F2 as final exacted features F according to the early fusion;
3: *Step*3: Apply several methods such as ROS, RENN etc. to F to adjust the distribution and select the optimal algorithm for training system;
4: *Step*4: System completion via an RF classifier after imbalanced treatment;
5: *Step*5: Acquire the optimized model by constantly training and adjusting parameters;
6: *Step*6: Test the model to judge whether they are qualified or not according to the metrics;
7: *Step*7: Output test accuracy with the final model.

3. Experimental materials

In this section, we mainly introduce the preprocessing of the data used in the experiments and the criteria for evaluating the experimental results. In Section 3.1, we describe the data sets and its processing for our experiment. In Section 3.2, we briefly explain the metrics we applied to evaluate experimental results.

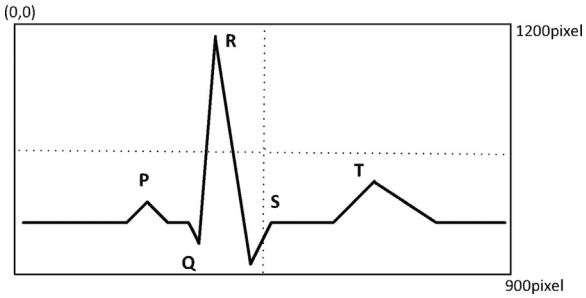

Fig. 6. One heartbeat is bounded in a window which is 1200 pixels in width and 900 pixels in height.

3.1. Data sets

3.1.1. MIT-BIH Arrhythmia Database (AD)

This publicly available arrhythmia ECG database contains 48 half-hour excerpts of two-channel ambulatory ECG recordings [30]. Those recordings were sampled through a filter at 360 Hz per second and 11-bit resolution for each channel. There are over 109,000 labels provided by two or more cardiologists independently for 15 different ECG beat types, and each record has its own label.

In this data set, we prepared inter-patient and intra-patient data separately. According to [41], in the intra-patient experiment, the training and test data sets were randomly selected from all heartbeats. In the inter-patient experiment, the data set were divided into training and test data sets according to the traditional partition rule. Namely, the training set (DS1) was composed of all heartbeats of records: 101, 106, 108, 109, 112, 114, 115, 116, 118, 119, 122, 124, 201, 203, 205, 207, 208, 209, 215, 220, 223, and 230, while the test set (DS2) was composed of all heartbeats of records: 100, 103, 105, 111, 113, 117, 121, 123, 200, 202, 210, 212, 213, 214, 219, 221, 222, 228, 231, 232, 233, and 234.

3.1.2. MIT-BIH ST Change Database (STCD).

This database includes 28 ECG recordings of varying lengths. Most of these were recorded during exercise stress tests. They exhibit transient ST depression. Records 323 through 327 are excerpts of long-term ECG recordings and exhibit ST elevation.

In this paper, only one channel of both two databases was selected to design the experiment. The records were split into a large number of beats, and each beat contained full PQRST values (Fig. 5). The records of the channel were classified into five beat types according to the AAMI standards: normal beats (N), supraventricular ectopic beats (S), ventricular ectopic beats (V), fusion beats (F), and unclassifiable beats (Q). The description of AAMI standards is given in Table 1. Specifically, we used the median filter to remove the baseline drift and the low-pass filter to remove the power line interference as well as high frequency noise [5]. Each heartbeat in our experimental data consists of 90 points on the left of P peak and 198 points on the right. We used spline interpolation to generate an ECG curve. Then, we used a window which is 1200 pixels in width and 900 pixels in height to bound the curve just like the screen of an ECG device. By this means, we acquired an 1200 * 900 image as shown in Fig. 6. The *Y*-axis corresponds to the normalized amplitude of the ECG signal. In order to reduce the computational cost of CNN, we performed a series of experiments to choose a proper smaller image size. Finally, the ECG image as the input of 2D-CNN is 96 * 72. The results indicate that the method based on physiology can improve classification accuracy.

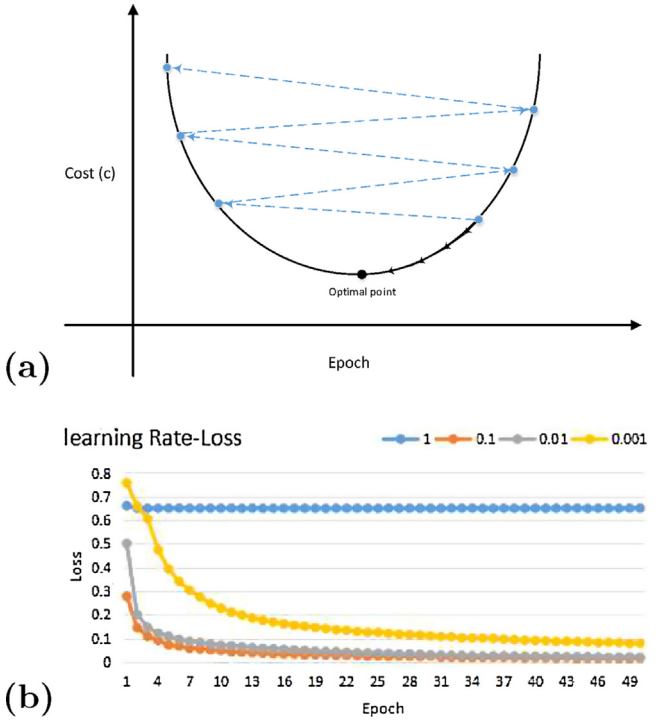

Fig. 7. (a) The dotted and solid line represent the real and ideal situation separately. The optimal value is the black point at the bottom. (b) The loss along with epoch at different learning rate are presented here. Normally, the loss at smaller learning rate drops slower.

3.2. Performance metrics

In order to express the classification results accurately and compare it with the existing experimental results conveniently, we used some common metrics [42], which are defined as follows:

$$AverageAccuracy(AA) = \frac{\sum_{i=1}^{l}(tp_i + tn_i/tp_i + fn_i + fp_i + tn_i)}{l}, \quad (3)$$

$$Errorrate(ER) = \frac{\sum_{i=1}^{l}(fp_i + fn_i/tp_i + fn_i + fp_i + tn_i)}{l}, \quad (4)$$

$$Precision_u(uP) = \frac{\sum_{i=1}^{l} tp_i}{\sum_{i=1}^{l}(tp_i + fp_i)}, \quad (5)$$

$$Recall_u(uR) = \frac{\sum_{i=1}^{l} tp_i}{\sum_{i=1}^{l}(tp_i + fn_i)}, \quad (6)$$

$$Fscore_u(uF) = \frac{\left(\beta^2 + 1\right)(uP)(uR)}{\beta^2(uP) + (uR)}, \quad (7)$$

$$Presion_M(MP) = \frac{\sum_{i=1}^{l}(tp_i/tp_i + fp_i)}{l}, \quad (8)$$

$$Recall_M(MR) = \frac{\sum_{i=1}^{l}(tp_i/tp_i + fn_i)}{l}, \quad (9)$$

$$Fscore_M(MF) = \frac{\left(\beta^2 + 1\right)(MP)(MR)}{\beta^2(MP) + (MR)}, \quad \beta = 1. \quad (10)$$

Those metrics are used for multi-classes $Y(i)$ and they involve four representations: true positive for $Y(i)$ (tp_i), False Positive (fp_i), True Negative (tn_i) and False Negative (fn_i). u and M indices represent micro- and macro-averaging. Of course, the confusion matrix is necessary to represent the classification intuitively. A detailed utilization is illustrated in the next section.

Table 1
Advancement of Medical Instrumentation (AAMI) standards (1988).

Heartbeat class	N	F	V	F	Q
Heartbeat type	Normal beat (N)	Atrial premature beat (A)	Premature ventricular contraction (V)	Fusion of ventricular and normal beat (F)	Paced beat (/)
	Left bundle branch block beat (L)	Aberrated atrial premature beat (a)	Ventricular escape beat (E)		Fusion of paces and normal beat (f)
	Right bundle branch block beat (R)	Nodal (junctional) premature beat (J)			Unclassified beat (Q)
	Atrial escape beats (e)	Supraventricular premature beat (S)			
	Nodal (junctional) escape beat (j)				

Table 2
The comparison between 1D-CNN and 2D-CNN. Four metrics are used.

	MP	MR	MF	AA
1D-CNN	0.9440	0.9600	0.9519	0.9400
2D-CNN	0.9600	0.9680	0.9640	0.9600

4. Results and discussion

In this section, we present partial parameter setting and the experimental results which prove the superiority of our model. In Experiment One, we illustrate the selection process of learning rate. In Experiment Two, 1D-CNN and 2D-CNN are compared. In Experiment Three, we compare the classification performance of CNN features and its combination with PQRST features. In Experiment Four, the results for several imbalanced processing algorithms applied to the optimal features of previous experiments are presented. Finally, we explain the different behaviors of our proposed model versus other methods proposed for classifying the same data set.

4.1. Experiment one: learning rate determination

Section 2.4 explains the importance of a suitable learning rate. As Fig. 7(b) shows, three different *lr* have different convergence epoches.

4.2. Experiment two: 1D-CNN and 2D-CNN

Others [28] have shown that 1D-CNN could acquire highly abstract features without dimensional reduction easily, but it could only get convolutional numerical results. Though training time in 2D-CNN is more significant than in 1D-CNN, theoretically more valuable information can be obtained when changing from 1D-CNN to 2D-CNN. We carried out a comparison experiment involving 1D-CNN and 2D-CNN. The results is shown in Table 2.

Table 3
Comparison of multi-class imbalanced processing methods. The best result in each metric is in bold.

	MR	MP	MF	AA
CC	0.8507	0.8763	0.8633	0.8658
RENN	0.9909	0.9848	0.9879	0.9853
ENN	0.9867	0.9795	0.9831	0.9806
NM	0.9850	0.9795	0.9823	0.9827
NCR	0.9888	0.9822	0.9854	0.9824
OSS	0.9814	0.9697	0.9755	0.9702
RUS	0.9027	0.8602	0.8809	0.8918
ROS	**0.9992**	**0.9998**	**0.9995**	**0.9996**

4.3. Experiment three: CNN and CNN+PQRST

As illustrated above, the heartbeat image data has many unique features that are not available in the original 1-D ECG data set. The heartbeat image features extracted by two-dimensional convolutional operation have more significant local characteristics. Furthermore, adding recognizable PQRST features to convolutional features remedies the shortcomings of convolution operation theoretically. Our experiment proved that the fusion algorithm improved classification accuracy by about one percent. Fig. 8 presents a detailed comparison information of CNN and CNN+PQRST.

4.4. Experiment four: imbalanced treatment

Recall that there is an extremely imbalanced phenomenon in the ECG data set, which may lead to misclassification. Because commonly used classifiers are more suitable for balanced data than for imbalanced data [44], changing imbalanced data into balanced ones may improve the accuracy somewhat. However, not all imbalanced treatments achieve the same results for their own unique applicable conditions. Table 3 summarizes several applications of imbalanced-dealing methods mentioned in Section 2.3. For each metric, the table shows the average value of every algorithm on AD we used.

Table 4
With the intra-patient scheme, results obtained by different methods according to the AAMI recommendations on the MBAD data set. The best result is in bold.

Method	Features	Classifier	Average Accuracy (%)
Tang et al.	WT	QNN	92.8300
Muhammad Zubair et al.	1D-CNN	softmax	94.6200
Martis et al.	DWT + PCA	SVM-RBF	97.3200
	DWT + PCA	NN	98.9800
Martis et al.	Cumulant + PCA	NN	95.5500
Elhajo et al.	PCA + DWT + HOS + ICA	SVM-RBF	99.0400
	PCA + DWT + HOS + ICA	NN	99.1000
Li et al.	WPE	RF	94.6100
Proposed	CNN + PQRST + balance	RF	**99.9000**

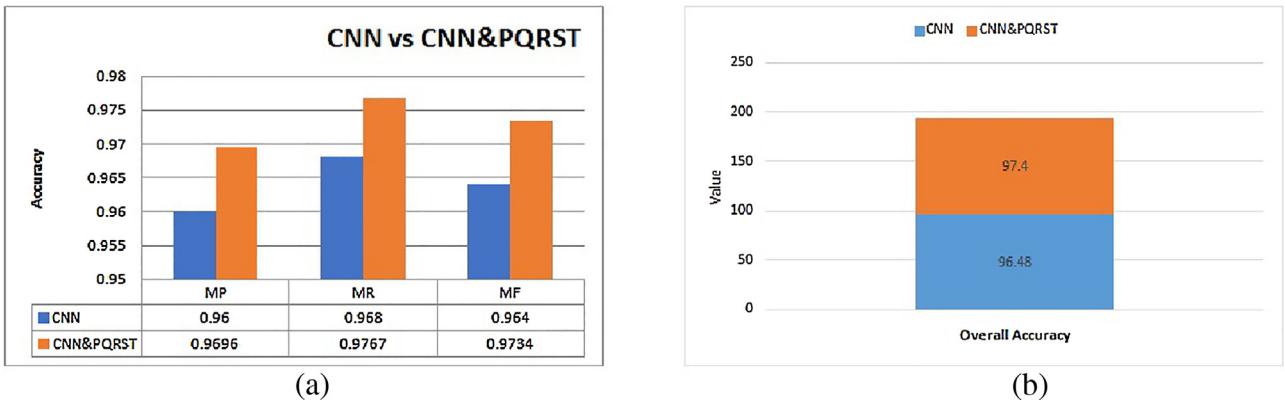

Fig. 8. The evaluating indicator of multi-class macro-average (a) and entirety (b).

Table 5
With the inter-patient scheme, results obtained by different methods according to the AAMI recommendations on the MBAD data set. The best result is in bold.

Method	MR	MP	MF	AA
Ye et al.	62.3700	65.7700	64.0200	65.2100
Yu and Chou	70.1200	72.3000	71.1900	75.2000
Song et al.	80.5500	84.3700	82.4200	81.8800
Güler and Übeyli	52.5700	43.2200	47.4400	60.1900
Proposed	**99.1300**	**98.5000**	**98.8100**	**99.5300**

4.5. Experiment five: system generalization ability

To further demonstrate the superiority of our system as well as its generalization, several previous works that used the same data set and standards were compared. According to the scheme of heartbeat selection, our contrast experiments consist of two groups: inter-patient (Table 5) and intra-patient (Table 4). All of the contrast experiments of each table were carried out using the same training and testing data sets as ours.

Considering the generalization ability and stability of the model, we tested our model with the other data set: STCD, whose division standards are the same as AD. Table 6 is the experiment we carried out based on STCD. In Table 6(a) (confusion matrix is Table 6(b)), all heartbeats in the STCD were used to test the SD1 training model, while the remaining steps remained unchanged.

4.6. Experiment analysis and discussion

From the above experiments, we can draw the following conclusions:

(1) A variable learning rate is more beneficial to the training model than a constant learning rate.

In Experiment One, we can notice that when $lr = 1$, the cost function with a high loss is almost unchanged. At $lr = 0.001$, the cost value changes slowly along with epoch and the final loss is higher than the loss at $lr = 0.1$ or $lr = 0.01$. Though the loss function at $lr = 0.1$ decreases sharply in first epoch, the loss of $lr = 0.01$ is lower than $lr = 0.1$. Over all consideration, we chose $lr = 0.1$ in the first 25 iterations, and $lr = 0.01$ for other iterations. This provides good performance of training model because val_cost (the validation cost of training model) decreases faster at the beginning but changes slower near the optimal value.

(2) The method that treats 1D ECG signal as 2D images is feasible.

It can be observed from Table 3 that the 2D-CNN improves the performance of 1D-CNN by 1-2 percentage points in terms of MP, MR, MF and AA. Especially in Precision_M and Average Accuracy, the accuracy of 2D-CNN is two percentage points higher than 1D-CNN, which indicates that our proposal has a high precision and discrimination ability for all the samples. Thus, we can infer that 2D-CNN is feasible even though the training time of 2D is longer than 1D. In addition, the performance of 1D-CNN is not satisfactory as in [28], which is probably because the data set and the parameters in our experiment are different with those used in [28].

(3) The system with fused features is better than the one with single CNN in terms of classification.

In Fig. 8, the classification performance of CNN features combined with PQRST features is better than CNN features alone. Though the difference in Average Accuracy is only about 1%, this proves that the correlation between the two kinds of features is small and fused features are helpful for classification.

(4) The Random Over Sampler method yields better classification results than other imbalanced methods in our research.

It can be observed from Experiment Four that nearly all methods improve the classification accuracy, especially Random Over Sampler method whose Average Accuracy reaches 99.96%. The Recall_M, Precision_M and Fscore_M rise to nearly 99% to 100%. On the contrary, the under sampling methods are far from satisfactory for 3 reasons: (1) they decrease the quantity of the majority class; (2) some methods do not change the overall distribution of samples (such as CC algorithm); (3) they tend to ignore some minority class samples as noise in the removal process (such as OSS algorithm). Therefore, the under sampling methods are likely not only to lose plenty of useful information of the majority class (such as RUS algorithm) and cause the minority class to still be at disadvantage, but also to trap their loss functions in a local optimum.

Thus, expanding the sample size to change the imbalanced distribution is helpful to avoid stepping into local optimization and protect the minority class. Therefore, we adopted the Random Over Sampler algorithm to balance our data set in our final model and in the following experiments.

(5) Our system is superior to the state-of-the-art methods in both intra-patient and inter-patient experiments, and the model trained with the AD data set shows good robustness and adaptability when it is transferred to the STCD data set.

Table 4 presents the comparative results of intra-patient experiments. Here, Tang et al. [24] used wavelet transform (WT) for feature extraction after normalization and used rough sets (RS) as well as quantum neural network (QNN) to recognize electrocardiogram signals, whose optimal recognition rate is 92.83%. Zubair et al. [29] introduced 1D-CNN to extract features and then classified the MIT-BIH data set into five classes recommended by AAMI standards. Martis et al. proposed linear [21] and nonlinear [45] methods, which achieved an accuracy of

Table 6
Results obtained by testing all of the STCD data on the SD1 training model.

(a) Classification accuracy of SD1 training model

(%)	ER	uP	uR	uF	MP	MR	MF	AA
Value	2.3709×10^{-2}	99.9640	99.9640	99.9640	99.9650	99.9640	99.9650	99.9760

(b) Confusion matrix of SD1 training model

	N prediction	S prediction	V prediction
N instance	22,440	23	1
S instance	0	22,464	0
V instance	0	0	22,557

above 98%, outstanding among classical ECG classification algorithms. Later, Elhajo et al. [31] combined linear features with nonlinear features, which improved the accuracy to 99.04%. Li et al. [11] decomposed the ECG signals by wave package decomposition (WPD) and calculated its coefficients entropy as representative features. This is the first time that WPF and RF used in inter-patient ECG classification according to AAMI standards. Table 5 shows several comparative inter-patient results. Ye et al. [15] concatenated RR interval features with features obtained from Wavelet Transform and Independent Component Analysis (ICA) as final features to classify. Yu and Chou [14] simply pushed the combined features coming from RR interval features and ICA features into neural network to classify. Song et al. [13] used a SVM classifier to classify the representative features extracted from wavelet coefficients and LDA. Güler and Übeyli [46] designed a two-level network employing time-frequency and statistical methods, in which the outputs of the first level were treated as the inputs of the second level.

It can be observed from Tables 4 and 5 that the results of intra-patient experiment are significantly superior to the results of inter-patient experiment. The reason is that the intra-patient model has patient specificity, which greatly improved the test results. To perform a fair evaluation of ECG classification performance and accord with a realistic scenario, heartbeats of training and testing sets should be from different people. The Average Accuracy of our system in both inter-patient and intra-patient experiments are above 99%. When we tested STCD data set on the SD1 model, the Error Rate is close to 0 and the seven other metrics are all above 99%, which indicates that almost all the samples were accurately classified.

From above, it can be summarized that our proposal has desirable adaptability and stability for inter-patient classification and for different data sets. It is superior to the state-of-the-art methods in terms of classification accuracy and category specificity.

5. Conclusion and directions

In this paper, a new ECG classification system based on feature fusion and imbalanced processing has been proposed. Comparing with the state-of-the-art approaches based on deep or shallow architectures, this system has some novelties:

- It handles two-dimensional ECG images rather than one-dimensional signals.
- It is the first report to fuse simple CNN features with traditional features as the inputs of the final Random Forest classifier.
- Several methods of imbalanced processing have been applied to balance sample distribution.
- In the experiment of model stability, we introduced the idea of transfer learning to enhance experimental reliability.

We carried out inter-patient and intra-patient experiments, as well as a test based on STCD for contrast. The experimental results show the Average Accuracy is above 99%, which indicates our system has strong adaptability and is superior to the state-of-the-art methods using the same data set and standards. For future developments, we plan to work in the following directions: (1) Expanding the scope of the model adaption and paying more attention to patient-specific data. (2) Improving the model with semi-supervised or unsupervised learning. (3) Introducing the reinforcement learning into the parameter tuning to reduce the time for classification.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 61271069).

The authors would like to thank Suyao Wang for implementation of 1D-CNN as contrast model. We would also like to thank the anonymous reviewers and the Associate Editor for the constructive evaluation of this paper.

References

[1] C. Deaton, E.S. Froelicher, L.H. Wu, C. Ho, K. Shishani, T. Jaarsma, The global burden of cardiovascular disease, Eur. J. Cardiovasc. Nurs. 10 (2 Suppl) (2011) S5–S13.
[2] M.M. Hadhoud, M.I. Eladawy, A. Farag, Computer aided diagnosis of cardiac arrhythmias, The 2006 International Conference on Computer Engineering and Systems, IEEE (2006) 262–265.
[3] O. Sayadi, M.B. Shamsollahi, Multiadaptive bionic wavelet transform: application to ECG denoising and baseline wandering reduction, EURASIP J. Adv. Signal Process. 2007 (1) (2007) 041274.
[4] O. Sayadi, M.B. Shamsollahi, ECG denoising and compression using a modified extended Kalman filter structure, IEEE Trans. Biomed. Eng. 55 (9) (2008) 2240–2248.
[5] P. De Chazal, M. O'Dwyer, R.B. Reilly, Automatic classification of heartbeats using ECG morphology and heartbeat interval features, IEEE Trans. Biomed. Eng. 51 (7) (2004) 1196–1206.
[6] C. Li, C. Zheng, C. Tai, Detection of ECG characteristic points using wavelet transforms, IEEE Trans. Biomed. Eng. 42 (1) (1995) 21–28.
[7] Y. Li, X. Tang, Z. Xu, An approach of heartbeat segmentation in seismocardiogram by matched-filtering, 2015 7th International Conference on Intelligent Human–Machine Systems and Cybernetics (IHMSC), vol. 2, IEEE (2015) 47–51.
[8] E. Pinheiro, O. Postolache, P. Girão, Method for segmentation of cardiac signals based on four parameter sine fitting, EUROCON-International Conference on Computer as a Tool (EUROCON), 2011 IEEE, IEEE (2011) 1–4.
[9] R. Ceylan, Y. Özbay, Comparison of FCM, PCA and wt techniques for classification ECG arrhythmias using artificial neural network, Expert Syst. Appl. 33 (2) (2007) 286–295.
[10] L.S. de Oliveira, R.V. Andreão, M. Sarcinelli-Filho, Detection of premature ventricular beats in ecg records using bayesian networks involving the p-wave and fusion of results, 2010 Annual International Conference of the IEEE, Engineering in Medicine and Biology Society (EMBC), IEEE (2010) 1131–1134.
[11] T. Li, M. Zhou, ECG classification using wavelet packet entropy and random forests, Entropy 18 (8) (2016) 285.
[12] M. Huanhuan, Z. Yue, Classification of electrocardiogram signals with deep belief networks, 2014 IEEE 17th International Conference on Computational Science and Engineering (CSE), IEEE (2014) 7–12.

[13] M.H. Song, J. Lee, S.P. Cho, K.J. Lee, S.K. Yoo, Support vector machine based arrhythmia classification using reduced features, Int. J. Control Autom. Syst. 3 (4) (2005) 571.
[14] S.-N. Yu, K.-T. Chou, Integration of independent component analysis and neural networks for ECG beat classification, Expert Syst. Appl. 34 (4) (2008) 2841–2846.
[15] C. Ye, M.T. Coimbra, B.V. Kumar, Arrhythmia detection and classification using morphological and dynamic features of ECG signals, 2010 Annual International Conference of the IEEE, Engineering in Medicine and Biology Society (EMBC), IEEE (2010) 1918–1921.
[16] J.M. Hsing, K.A. Selzman, C. Leclercq, L.A. Pires, M.G. McLaughlin, S.E. McRae, B.J. Peterson, P.J. Zimetbaum, Paced left ventricular QRS width and ECG parameters predict outcomes after cardiac resynchronization therapy clinical perspective, Circ.: Arrhythm. Electrophysiol. 4 (6) (2011) 851–857.
[17] M.K. Das, D.K. Ghosh, S. Ari, Electrocardiogram (ECG) signal classification using s-transform, genetic algorithm and neural network, 2013 IEEE 1st International Conference on Condition Assessment Techniques in Electrical Systems (CATCON), IEEE (2013) 353–357.
[18] J. Agrawal, R. Vijay, Time-frequency filtering with the s-transform of ECG signals, Int. J. Sci. Res. 3 (2013) 1–5.
[19] S. Ilić, Comparison of compression ratios for ecg signals by using three time-frequency transformations, Facta Univ.-Ser.: Electron. Energ. 20 (2) (2007) 223–232.
[20] A.A. Shinde, P. Kanjalkar, The comparison of different transform based methods for ECG data compression, 2011 International Conference on Signal Processing, Communication, Computing and Networking Technologies (ICSCCN), IEEE (2011) 332–335.
[21] R.J. Martis, U.R. Acharya, L.C. Min, ECG beat classification using PCA, LDA, ICA and discrete wavelet transform, Biomed. Signal Process. Control 8 (5) (2013) 437–448.
[22] L. Zhang, H. Peng, C. Yu, An approach for ECG classification based on wavelet feature extraction and decision tree, 2010 International Conference on Wireless Communications and Signal Processing (WCSP), IEEE (2010) 1–4.
[23] N. Emanet, ECG beat classification by using discrete wavelet transform and random forest algorithm, ICSCCW 2009. Fifth International Conference on Soft Computing, Computing with Words and Perceptions in System Analysis, Decision and Control, 2009, IEEE (2009) 1–4.
[24] X. Tang, L. Shu, Classification of electrocardiogram signals with RS and quantum neural networks, Int. J. Multimed. Ubiquitous Eng. 9 (2) (2014) 363–372.
[25] Z. Gu, N. Zhang, S. Qian, Analyzing electrocardiogram signals with multiscale short-time Fourier transforms, uS Patent 8,494,622 (July 23 2013).
[26] E. Uslu, G. Bilgin70, Exploiting locality based Fourier transform for ECG signal diagnosis, 2012 International Conference on Applied Electronics (AE), IEEE (2012) 323–326.
[27] I.A. Dmitrievich, Deep learning in information analysis of electrocardiogram signals for disease diagnostics, Ph.D. thesis), Moscow Institute of Physics and Technology, 2015.
[28] S. Kiranyaz, T. Ince, R. Hamila, M. Gabbouj, Convolutional neural networks for patient-specific ECG classification, 2015 37th Annual International Conference of the IEEE, Engineering in Medicine and Biology Society (EMBC), IEEE (2015) 2608–2611.
[29] M. Zubair, J. Kim, C. Yoon, An automated ECG beat classification system using convolutional neural networks, 2016 6th International Conference on IT Convergence and Security (ICITCS), IEEE (2016) 1–5.
[30] G.B. Moody, R.G. Mark, The impact of the MIT-BIH arrhythmia database, IEEE Eng. Med. Biol. Mag. 20 (3) (2001) 45–50.
[31] F.A. Elhaj, N. Salim, A.R. Harris, T.T. Swee, T. Ahmed, Arrhythmia recognition and classification using combined linear and nonlinear features of ECG signals, Comput. Methods Progr. Biomed. 127 (2016) 52–63.
[32] G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R.R. Salakhutdinov, Improving neural networks by preventing co-adaptation of feature detectors, arXiv:1207.0580.
[34] R.E. Schapire, A Brief Introduction to Boosting, Ijcai, vol. 99 (1999) 1401–1406.
[35] M.T. Hagan, M.B. Menhaj, Training feedforward networks with the Marquardt algorithm, IEEE Trans. Neural Netw. 5 (6) (1994) 989–993.
[36] H. He, E.A. Garcia, Learning from imbalanced data, IEEE Trans. Knowl. Data Eng. 21 (9) (2009) 1263–1284.
[37] D.L. Wilson, Asymptotic properties of nearest neighbor rules using edited data, IEEE Trans. Syst. Man Cybern. 2 (3) (1972) 408–421.
[38] A. Fernández, S. Garc’í, M.J. del Jesus, F. Herrera, A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data-sets, Fuzzy Sets Syst. 159 (18) (2008) 2378–2398.
[39] M. Kubat, S. Matwin, et al., Addressing the curse of imbalanced training sets: one-sided selection, in: ICML, vol. 97, Nashville, USA, 1997, pp. 179–186.
[41] I. Mani, I. Zhang, KNN approach to unbalanced data distributions: a case study involving information extraction, Proceedings of Workshop on Learning from Imbalanced Datasets (2003) 42–48.
[42] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[44] A.K. Jain, R.P.W. Duin, J. Mao, Statistical pattern recognition: a review, IEEE Trans. Pattern Anal. Mach. Intell. 22 (1) (2000) 4–37.
[45] R.J. Martis, U.R. Acharya, C.M. Lim, K. Mandana, A.K. Ray, C. Chakraborty, Application of higher order cumulant features for cardiac health diagnosis using ECG signals, Int. J. Neural Syst. 23 (04) (2013) 1350014.
[46] İ. Güler, E.D. Übeylı, ECG beat classifier designed by combined neural network model, Pattern Recogn. 38 (2) (2005) 199–208.