Expert Systems with Applications: X 1 (2019) 100003

Contents lists available at ScienceDirect

Expert Systems with Applications: X

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

A novel multi-module neural network system for imbalanced heartbeats classification

Jing Jiang, Huaifeng Zhang, Dechang Pi*, Chenglong Dai

College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China

ARTICLE INFO

Article history:
Received 2 October 2018
Revised 13 March 2019
Accepted 14 March 2019
Available online 14 March 2019

Keywords:
Multi-module
Heartbeats classification
Imbalance problem
Convolutional neural network

ABSTRACT

In this paper, a novel multi-module neural network system named MMNNS is proposed to solve the imbalance problem in electrocardiogram (ECG) heartbeats classification. Four submodules are designed to construct the system: preprocessing, imbalance problem processing, feature extraction and classification. Imbalance problem processing module mainly introduces three methods: BLSM, CTFM and 2PT, which are proposed from three aspects of resampling, data feature and algorithm respectively. BLSM is used to synthesize virtual samples linearly around the minority samples. CTFM consists of DAE-based feature extraction part and QRS-based feature selection part, in which selected features and complete features are applied to determine the heartbeat class simultaneously. The processed data are fed into a convolutional neural network (CNN) by applying 2PT to train and fine-tune. MMNNS is trained on MIT-BIH Arrhythmia Database following AAMI standard, using intra-patient and inter-patient scheme, especially the latter which is strongly recommended. The comparisons with several state-of-the-art methods using standard criteria on three datasets demonstrate the superiority of MMNNS for improving detection of heartbeats and addressing imbalance in ECG heartbeats classification.

1. Introduction

Cardiovascular diseases (CVDs) are the leading cause of death globally. The number of people dying of CVD increases year by year, which resulted in 17.9 million deaths (32.1%) in 2015, up from 12.3 million (25.8%) in 1990 (Wang, Naghavi, Allen, Barber, & Bhutta, 2015). Cardiovascular diseases (CVDs) are disorders of the heart and blood vessels, including coronary heart disease, cerebrovascular disease, rheumatic heart disease and other conditions. Arrhythmia is one of the cardiovascular diseases during which the heart can beat too fast, too slowly or with an irregular rhythm. Generally speaking, arrhythmia can be mainly divided into two types. One is life-threatening ventricular fibrillation and tachycardia, requiring immediate treatment with a defibrillator. The other is the arrhythmia studied in this paper, being not immediately life-threatening but requiring further treatment.

Electrocardiogram (ECG) is a promising diagnostic tool for examining cardiac tissues and structures. It reflects the electrical activity of the heart recorded by electrodes placed on the skin over a period of time and consists of different waveforms representing the polarization or depolarization of the heart (Šarlija, Jurišić, & Popović, 2017). ECG contains a large number of information on not only the structure of the heart but also the function of its electrical conduction system. Moreover, it also provides data for diagnosis of diseases, classification of heartbeat, etc.

ECG widely applied in the fields of related disease classification, heartbeat type detection, biometric identification and emotion recognition (Kaplan Berkaya et al., 2018). The study in this paper aims to classify heartbeats, which is an important step in diagnosing arrhythmia. Recommended by the Association for the Advancement of Medical Instrumentation (AAMI), non-life-threatening arrhythmia can be divided into five subclasses (Table 1 lists the detailed heartbeat types): normal (N), supraventricular (S/SVEB), ventricular (V/VEB), fusion (F) and unknown beats (Q). Each type of beats has a great difference in morphology, and each type contains several subclasses with different shape, which brings a great challenge for physicians to analyze manually. To compensate for visual errors and manual interpretations, researchers have begun to develop computer-aided diagnosis (CAD) systems to automatically diagnose ECG.

With the emergence of CAD system in clinical medicine, the workload of cardiologists has been reduced, and the computational efficiency and accuracy of disease detection have been improved.

* Corresponding author.
E-mail address: nuaacs@126.com (D. Pi).

https://doi.org/10.1016/j.eswax.2019.100003

Table 1
ECG class description using AAMI standard.

AAMI heartbeat class	MIT-BIH heartbeat type
Normal beats (N)	Normal beat (N)
	Left bundle branch block (L)
	Right bundle branch block (R)
	Atrial escape beat (e)
	Nodal (junctional) escape beat (j)
Supraventricular ectopic beats (S)	Atrial premature contraction (A)
	Aberrated atrial premature beat (a)
	Nodal (junctional) premature beat (J)
	Supraventricular premature beat (S)
Ventricular ectopic beats (V)	Ventricular premature contraction (V)
	Ventricular escape Beat (E)
Fusion beats (F)	Fusion of ventricular and normal beat (F)
Unknown beats (Q)	Paced beat (/)
	Fusion of paced and normal beat (f)
	Unclassified beat (Q)

A CAD system automatically classifies heartbeats, which contains four steps of ECG signal preprocessing, heartbeat segmentation, feature extraction and learning/classification. Traditional methods first extract features from the original data, such as P-QRS-T complex features, statistical features, morphological features, wavelet features, etc. (Acharya et al., 2015; Khandoker, Palaniswami, & Karmakar, 2009; Li, Rajagopalan, & Clifford, 2014; Yücelbaş et al., 2017), and then feed them into conventional machine learning models, for instance, artificial neural networks, decision trees, support vector machine, linear discriminant analysis, k nearest neighbor and Bayesian algorithm (Chui, Tsang, Chi, Ling, & Wu, 2016; De, O'Dwyer, & Reilly, 2004; Martis, Acharya, Prasad, Chua, & Lim, 2013; Pławiak, 2018; Raj & Ray, 2018) to realize heartbeats classification. However, such features extracted manually may not accurately represent the optimal features in the signal and traditional machine learning methods can easily lead to overfitting (Acharya et al., 2017; Rahhal et al., 2016). Therefore, more accurate and efficient feature extraction and classification methods are critical to the overall diagnosis of the system.

As the great success in image recognition, speech recognition, natural language processing and other fields, deep learning model has been gradually applied to ECG analysis in recent years. Taking convolutional neural network (CNN) as an example, CNN integrates feature extraction and classification, classifying high-level features automatically learned by itself from original data. Li, Zhang, Zhang, and Wei (2017) feed the raw ECG signals directly into a 5-layer CNN for feature extraction and training. The idea of transfer learning is adopted in Isin and Ozdalili (2017) to take pre-trained AlexNet as a feature extractor, and the extracted features are then fed into a simple BP neural network to classify the heartbeats. Rahhal et al. (2016) uses active learning technology to improve the performance of the system. Zhai and Tin (2018) and Golrizkhatami and Acan (2018) process the input of the final classifier respectively, the former transforms the beats into dual beat couple matrix as 2-D inputs of CNN, and the latter fuses the 2D-convolutional and handcrafted features extracted from each beat. The accuracy of deep learning model is shown higher than that of the traditional classifier combined with manual feature extraction.

Although the aforementioned methods have achieved considerable classification accuracy, the evaluation metric of accuracy is incomprehensive when training data is skewed. Most of these papers fail to pay attention to the adverse effects of imbalance problem yet. As a result, solving the imbalance problem reasonably according to the characteristics of ECG data is the key to improve the classification performance of the model.

To overcome the deficiencies mentioned above, this paper proposed a novel multi-module neural network system for imbalanced heartbeats classification, which is called MMNNS. The main contributions of our proposed system are highlighted as follows:

- Four submodules are designed to construct the whole system, of which Imbalance Problem Processing Module, the core part, introduces three methods to eliminate the negative effect of imbalanced data distribution, named Borderline-SMOTE (BLSM), Context-Feature Module (CTFM) and Two-Phase Training (2PT). These methods are from the views of resampling, data feature and algorithm respectively. To the best of our knowledge, this is the first try to combine such three modules together to overcome imbalance problem.
- Denoising autoencoder (DAE) and convolutional neural network (CNN) are used to extract high-level features in turn from ECG signal automatically instead of manually extracting features, thus simplifying the feature extraction process and improving the accuracy of extracted features.
- MMNNS is trained on two data division schemes of intra-patient and inter-patient simultaneously, especially the latter which is more convincing but has not been adopted by most scholars. To validate the effectiveness of our model, extensive experiments are carried out on three datasets using seven classification assessment metrics and two statistical measures. Besides, a large number of comparisons have been made between MMNNS and state-of-the-arts.

The rest of paper is organized as follows. In Section 2, a succinct background of denoised autoencoder and convolutional neural network is introduced. Detailed descriptions of the proposed method are demonstrated in Section 3. Section 4 presents the experimental configuration and reports the experimental results. Conclusion and future direction are drawn in Section 5.

2. Preliminaries

Denoising autoencoder and convolutional neural network, as two classical deep learning models, are adopted to extract high-level features from ECG heartbeats in this study. Detailed introductions are made in this section.

2.1. Denoising autoencoder

Autoencoder is a neural network capable of learning effective features from input data in an unsupervised manner (Géron, 2017). It can reconstruct ECG data by extracting the most useful sparse high-level features. To make the autoencoder learn more useful features instead of simply copying the input data, we usually add some restrictions to the network and force the model to consider which parts of the input data need to be copied first. In this paper, a denoising autoencoder (DAE) is used to predict the original data by training corrupted data.

A corruption process $C(\tilde{x}|x)$ is introduced in the model firstly (see Fig. 1). Then the corrupted data is fed into the network. The network consists of two parts: one is the encoder represented by the function $h = f(x)$, which converts input into internal features; the other is a decoder that generates reconfiguration, $r = g(h)$, which converts internal features into output.

$$h = f(\tilde{x}) = f(W_1\tilde{x} + b_1) \tag{1}$$

$$f(z) = \max(0, z) \tag{2}$$

$$r = g(h) = W_2h + b_2 \tag{3}$$

where $\tilde{x}$ is the copy of the input x after adding the Additive white Gaussian noise (AWGN). W_1 and W_2 are the weight matrices of encoder and decoder respectively. Because the autoencoder built here

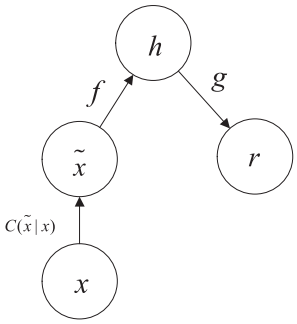


Fig. 1. DAE structure.

is neatly symmetrical, we tie the weights of the decoder layers to the weights of encoder layers, i.e. $W_1 = W_2^T = W$. This technique halves the number of weights, speeding up training and limiting the risk of overfitting. b_1 and b_2 are the bias vectors of input and output layers, respectively. Activation function ReLU is used in coding layer to get nonlinearization of the input, while the decoding layer is a linear process and does not require the activation function.

We usually adjust the hyperparameters by minimizing the cost function of Eq. (4), so that the DAE can well represent the input signal.

$$L(w, b_1, b_2) = \sum \|r - x\|_2^2 \tag{4}$$

2.2. Convolutional neural network

Convolutional neural networks (CNNs) refer to the neural networks that use convolution operations at least one layer of the network instead of general matrix multiplication, specifically for processing data with grid-like topology (Ian Goodfellow, 2016). It integrates feature extraction and classification together, different from the traditional classifier which needs to input pre-extracted features. A typical layer in a CNN consists of a convolutional layer and a pooling layer. Feature maps of previous layer are computed with several convolutions in parallel to generate a set of linear activation responses. Then, they passed through a nonlinear activation function to generate a feature map for next layer. Finally, the pooling function is used to further adjust the output. CNNs are stacked with several typical layers, which can be used to extract high-level features.

Each unit in the convolutional layer is connected to a local receptive field of the previous layer's feature map by a set of weights called filters (convolution kernels) and each convolution kernel obtains a mapping of a class of features. Different feature maps use different filters, while all neurons in same feature map share same filters (Lecun, Bengio, & Hinton, 2015). The formula for the output of a neuron in a convolutional layer is

$$c_{i,j,k} = \sigma\left(b_k + \sum_{u=1}^{f_h}\sum_{v=1}^{f_w}\sum_{k'=1}^{f_{n'}} x_{i',j',k'} \cdot w_{u,v,k,k'}\right) \text{ with } \begin{cases} i' = u \cdot s_h + f_h - 1 \\ j' = v \cdot s_w + f_w - 1 \end{cases} \tag{5}$$

where $c_{i,j,k}$ is the output of the neuron located in row i column j in feature map k of the convolutional layer (layer l). $x_{i',j',k'}$ is the output of the neuron located in layer $l-1$, row i', column j', feature map k' (or channel k' if the previous layer is the input layer). $w_{u,v,k,k'}$ is the connection weight between any neuron in feature map k of the layer l and its input located at row u, column v (relative to the neuron's receptive field), and feature map k'.b_k is the bias for feature map k (in layer l) and σ is the activation function which produce non-linearity. f_h and f_w are the height and width of the receptive field, and $f_{n'}$ is the number of feature maps in the previous layer (layer $l-1$). s_h and s_w are the vertical and horizontal strides respectively (Géron, 2017).

In this study, the input heartbeat data is one-dimensional vectors with one channel, so we modify the output of a neuron in a convolutional layer as Eq. (6).

$$c_{j,k} = \sigma\left(b_k + \sum_{v=1}^{f_w}\sum_{k'=1}^{f_{n'}} x_{j',k'} w_{v,k,k'}\right) \text{ with } j' = v \cdot s_w + f_w - 1 \tag{6}$$

Pooling layer, also called subsampling layer, is aimed to subsample the input data in order to reduce computational load and the memory usage. Max pooling is used in this paper to preserve the most prominent features by offering the maximum output within a rectangular neighborhood (Ian Goodfellow, 2016).

The training of CNN is similar to that of traditional fully-connected neural network, which is usually trained through back propagation algorithm to minimize the loss function.

3. Multi-module neural network system

This section introduces the whole system structure, mainly including four modules: signal preprocessing, imbalance problem processing, feature extraction and classification. In the process of preprocessing, the original data is denoised and segmented into heartbeat segment of equal length. Imbalance problem processing is the main emphasis of the Algorithm. This part combines the characteristic of both ECG data and classification model, taking a series of measures to process the imbalanced data after segmentation based on data level and algorithm level. Then, the processed imbalanced data are fed into the CNN model for feature extraction and classification.

The flow chart is described in Fig. 2.

3.1. Preprocessing

ECG recordings are usually contaminated by different types of noises and artifacts, which will affect the subsequent experiments and the final classification results. It is essential to adopt reasonable method to pre-process the original signal without losing useful information. Although the ECG signals obtained from public datasets do not contain as much noise as the ECG data obtained directly from the patient, there are still some noises in the signal spectrum that overlap with useful information. Given that ECG signal is the raw data of almost all cardiac disease diagnosis and analysis, we should utilize a reasonable method to preprocess the original signal without losing useful information.

It is pointed out in Haritha, Ganesan, and Sumesh (2016) that the low-frequency noise, mainly coming from baseline wander, can be removed by median filter effectively. As in De et al. (2004), we use a 200 ms width median filter to remove P wave and QRS complex wave, then use a 600 ms width median filter to remove T wave, and afterwards subtract the filtered signal from the original signal to get the baseline correction signal. The 12-tap low-pass filter is used to remove high-frequency noise and power-line interference.

Then Pan-Tompkins algorithm (Pan & Tompkins, 2007) is used to detect R-peak. The signal is segmented into pieces with the length of 300 sampling points by taking the first 130 samples and the last 170 points including R-peak as fiducial point. Subsequently, we apply Z-score normalization to eliminate the effects of offset and amplitude scaling and group the segmented pieces by categories, after which is easy to facilitate the following operations.

Algorithm: BLSM(T, M, r, k, s).

Input: T-Training set
M-Minority examples set
r, k-Number of nearest neighbors
s-Number of synthetic examples that account for the number of original examples in the given class
Output: Synthetic minority samples set: M'

```
D = Φ                                   //D is a set containing borderline samples
for all m_i in M do
    N_{m_i} ← r nearest neighbors of m_i in T
    n ← the number of samples in N_{m_i} and not in M
    if r/2 ≤ n < r then                 //m_i is a borderline sample
        add m_i to D
    end if
end for
M' = Φ                                  //M' is a set containing synthetic samples
for all d_i in D do
    N_{d_i} ← k nearest neighbors of d_i in M
    for i = 1 to s do
        m ← choose a random sample from N_{d_i}
        d'_i ← d_i + p * (d_i − m)      //p is a random number in(0, 1), d'_i is a synthetic sample
        add d'_i to M'
    end for
end for
M' = M' ∪ M                             //M' is the union of minority samples and synthetic samples
return M'
```

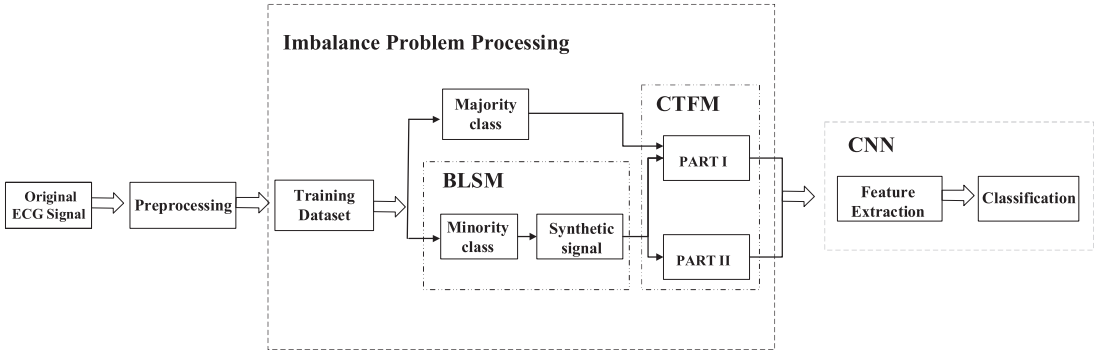


Fig. 2. Schematic overview of MMNNS.

3.2. Imbalance problem processing

Data imbalance refers to the phenomenon that the number of certain class in training set is overrepresented compared to other classes. The class with too many numbers is called majority class, while others are called minority class. Most real-world datasets are imbalanced. For example, the number of N-type heartbeats in MIT-BIH Arrhythmia Database is more than 8000 times that of Q-type and 100 times that of F-type. In medical diagnosis, the cost of erroneously misclassifying the minority (i.e. abnormal samples), which will delay the best treatment time, is much higher than that of majority classes (i.e. normal samples). In addition, most machine learning algorithms are designed under the assumption that the underlying training set is balanced. Such highly skewed distribution of training data will be prone to force the learning algorithm biased towards majority class. This not only limits the convergence during the training phase, but also affects the generalization of the model and the accuracy on testing set. As a result, the importance of minority class should not be neglected and it is the key to solve the imbalanced problem in this study.

Methods for solving class imbalance can be divided into three main categories (Buda, Maki, & Mazurowski, 2017; Guo, Li, Shang, Gu, Huang, & Gong, 2017).

Data level methods mainly change the distribution of training data by oversampling or undersampling. The basic way of oversampling is to simply replicates the samples selected from minority class randomly, which is called random oversampling. However, random oversampling is prone to cause overfitting. As a result, more advanced techniques are proposed, such as SMOTE, Cluster-based oversampling, DataBoost-IM, Class-aware sampling and other adjusted SMOTE strategies (Chawla, Bowyer, Hall, & Kegelmeyer, 2002; Guo & Viktor, 2004; Jo & Japkowicz, 2004; Maldonado, López, & Vairetti, 2019; Shen, Lin, & Huang, 2016). Undersampling, in contrast to oversampling, randomly removes samples from majority until the distribution gets balanced. More attention must be paid to avoid losing useful information in data selection when using undersampling.

Algorithm level methods overcome class imbalance by adjusting algorithm with training data distribution unchanged, including Thresholding, Cost-sensitive learning, Ensemble methods and One-class classification (Elkan, 2001; Lee & Cho, 2006; Sun, Kamel, Wong, & Wang, 2007; Zhou & Liu, 2006). Although Cost-sensitive learning can significantly improve the classification performance, they are only applicable to the cases where misclassification costs are known. Unfortunately, it is quite challenging and even impossible to determine the cost of misclassification in some particular domains (Wang et al., 2016). As for ensemble methods, requiring considerable time to train multiple classifiers, it is not practical when we use deep neural network as the base classifier. Thus new methods such as second order cone programing SVM (Maldonado & López, 2014) have been proposed.

Hybrid methods combine methods in data level and algorithm level, such as EasyEnsemble, BalanceCasecade, SMOTEBoost, Two-Phase Training, etc. (Chawla, Lazarevic, Hall, & Bowyer, 2003;

Havaei et al., 2017; Liu, Wu, & Zhou, 2009). Compared with data level or algorithm level methods only, it is shown that hybrid methods are more effective in imbalanced learning.

Besides, performance measures are essential for effectiveness evaluations and guidance of learning. As stated at the beginning of this section, accuracy, a commonly used metric, favors majority class and easily leads to misleading assessment. Metrics like precision, specificity, sensitivity, G-mean, F-measure are introduced into imbalanced learning field. Some works even focus on novel evaluation metrics, such as Adjusted F-measure (Maratea, Petrosino, & Manzo, 2014).

Neural networks and deep learning have gained lots of interests recently, also facing the imbalance problem. In standard back propagation algorithm, weights are updated by minimizing overall error, which are contributed mainly from majority class. Thus, we achieve biased classification results. To tackle this problem, two novel loss functions named MFE and MSFE are proposed in Wang et al. (2016). A stacked denoising autoencoder neural networks algorithm based on cost-sensitive oversampling is designed in Zhang, Gao, Song, and Jiang (2016). Khan, Hayat, Bennamoun, Sohel, and Togneri (2018) present a cost sensitive deep neural network which can automatically learn robust feature representations for both of the majority and minority classes. Raj, Magg, and Wermter (2016) integrate different methods into a novel approach using cost sensitive neural networks to improve the performance on imbalanced datasets.

We design a structure to deal with imbalanced training data combining both data and algorithm level. Above all, from data point of view, we apply Borderline-SMOTE (BLSM) algorithm to one-dimensional ECG time series data, and oversample minority samples by linear synthesis. Secondly, considering the features extracted from skewed data are biased towards majority class, a novel context-feature module (CTFM) is introduced in this paper. CTFM integrates feature extraction and feature selection, feeding prominent feature of each heartbeat segmentation and the features in the larger region of its context to the classifier simultaneously. CTFM doubles the number of minority samples, and enhances both the accuracy of model recognition and reliability for feature extraction. Finally, we adopt Two-Phase Training, referred to as 2PT. In the first training stage, the convolutional neural network (CNN) is trained with balanced data, and in second stage, the model is fine-tuned with original skewed data.

In order to make the test results closer to the real results and convenient to compare with other literature, we only balance the training samples and keeps the test samples as the original sample distribution in this paper. Considering that the accuracy is only more convincing when the data distribution is relatively balanced, we also introduce other metrics to measure the efficacy of the model. In the following chapters, we will give a detailed description of these contents.

3.2.1. Borderline-SMOTE (BLSM)

Given that deep learning models need large quantities of training data, we balance our training data by oversampling the minority class rather than undersampling the majority class. Also, to avoid overfitting caused by simple random oversampling, we choose BLSM algorithm to synthesize a series of linear interpolation samples.

Data augmentation can be regarded as expanding data by adding similar but different samples in the original training set (Zhang, Cisse, Dauphin, & Lopezpaz, 2018). On the basis of this definition, overcoming the imbalance problem in the number of training data can be regarded as local augmentation of minority class in training data. When realizing images classification, we usually obtained augmented data with horizontal reflection, rotation and scaling. Inspired by this idea, we plan to add virtual samples

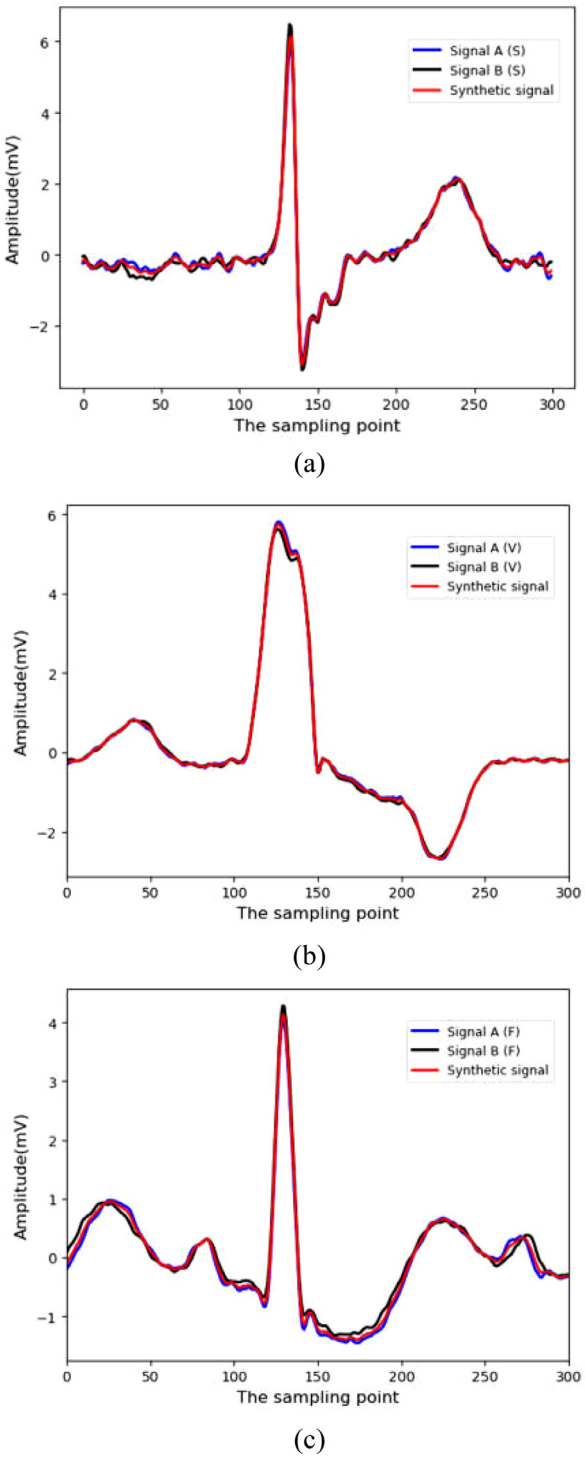


Fig. 3. Examples of synthetic signal (a) Example synthesized by same class S (b) Example synthesized by same class V (c) Example synthesized by same class F.

around the minority class. By learning a linear interpolation function in the 'blank area' around it, distribution of minority class in training data set will be enlarged, and complexity of space without data coverage will be reduced. This linear modeling reduces the inadaptability of predicting data beyond training samples, and makes the final model more stable in predicting data between training data. It is exactly consistent with the main idea of SMOTE (Chawla et al., 2002), i.e., filling the 'blank area' between original samples by linearly synthesizing minority sample with its k

Table 2
The changing number of training set in two grouping schemes.

Data div. scheme	Method	N	S	V	F	Total
Intra-patient scheme	Original	44,800	1345	3412	396	49,953
	BDSM	44,800	22,740	23,102	22,173	112,815
	CTFM	44,800	45,480	46,204	44,346	180,830
Inter-patient scheme	Original	45,492	865	3664	410	50,431
	BDSM	45,492	23,265	23,024	22,410	114,191
	CTFM	45,492	46,530	46,048	44,820	182,890

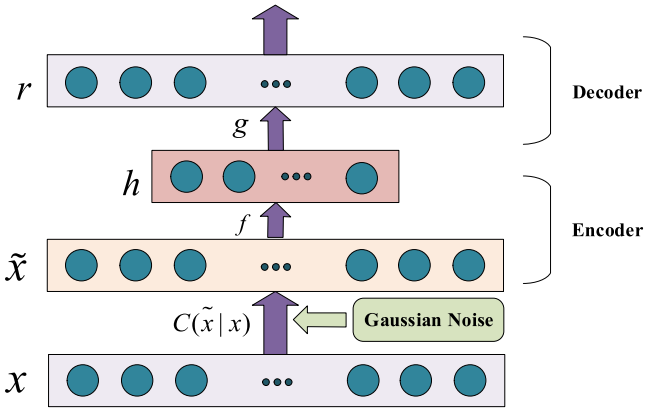


Fig. 4. CTFM DAE structure.

Table 3
The details for CNN structure with input length N of 180.

Layer name	No. of neurons	Kernel size for each feature map	Stride
Input Layer	180×1	–	–
Conv1	180×16	2	1
MaxPool1	180×16	2	2
Conv2	180×32	4	1
MaxPool2	180×32	2	2
Conv3	180×64	5	1
MaxPool3	180×64	2	2
Fully-Connected	50	–	–

nearest neighbors in feature space instead of data space. This enlarges the decision region and reduces the pertinence of classifier learning.

SMOTE resamples the training data by synthesizing each minority sample and its random selected neighbor. Since the number of original samples in each class located at the boundaries is small, the newly synthetic samples are mainly distributed in the non-boundary regions. However, the samples near the borderline, also referred as dangerous samples, are more likely to be misclassified than those far away from borderline, and are of more significance to classification tasks. Therefore, BLSM, an improved version of SMOTE, only oversamples the minority examples near the borderline, which is more in line with our expectations.

In BLSM, minority class examples which are easily misclassified will get more training. Minority examples on the borderline are first found, and then synthesized with their selected k nearest neighbors. BLSM algorithm is modified slightly and adapted to one-dimensional ECG time series data in this paper. The detailed process is shown as follows.

After oversampling with BLSM, the number of samples in minority class has increased significantly. Assuming that the amount of original samples in minority class is *num* and the number of samples in borderline samples set D is *dnum*, then the number of samples in M' reaches $(num + s \times dnum)$ in the end. In this algorithm, the values of s and m are determined by the amounts of samples that we need to synthesize. The values of s in S, V, F class with two grouping schemes are (11, 14), (11, 11), (17, 40) respectively, and the values of m are (30, 30), (50, 50), (60, 60). The value of k is set to 5 (Chawla et al., 2002; Han, Wang, & Mao, 2005). The synthetic ECG heartbeat patterns are shown in Fig. 3.

What needs to be stressed is that the number of samples varies greatly from one class to another. If a complete balance is achieved by BLSM, it may cause the boundary between different classes to become blurred, which makes it easier to misclassify. Therefore, we synthesize the minority class to make they reach only half of the number of majority class.

3.2.2. Context-feature module (CTFM)

After finishing oversampling using BLSM, a Context-Feature Module (CTFM) is proposed in this section to expand the samples obtained previously. The number of samples in minority class can be increased from N to 2 N by CTFM and all classes will reach balance then.

Compared with resampling methods in data level, few papers considered feature selection, which is known to select a subset of the original features. In classification problems, samples of different classes may overlap, which makes it difficult for classifiers to identify the boundaries between classes. When the distribution of training data is skewed, minority samples can be easily regarded as noise because of the small number. However, if irrelevant features are removed, boundaries between classes will be less ambiguous to some extent. Li, Guo, Liu, Li, and Li (2016) focus on feature selection and then combine it with resampling and ensemble leaning into an adaptive multiple classifiers system, finally solving the imbalance problem. A novel feature learning method using partial least square analysis is proposed in Bae and Yoon (2015) to learn unbiased results from imbalanced datasets.

Feature extraction, as another way to deal with dimensionality, converts the original features into new feature set. When solving imbalance problem, extracted features are often biased to predict the majority class samples that lead to poor performance on classification. Taking principal component analysis (PCA) as an example, (Braytee, Liu, & Kennedy, 2016) point out that PCA algorithm seeks the orthogonal feature extractors that maximize the total variance. As a result, the extracted features favor majority class because their number is larger than the minority class. Moreover, other feature extraction methods will encounter the same problem. Accordingly, improved feature extraction algorithms are proposed recently to adapt to imbalanced training data (Moepya, Akhoury, & Nelwamondo, 2015; Ng, Zeng, Zhang, Yeung, & Pedrycz, 2016).

Similarly, if we feed imbalanced data directly into DAE or CNN in this study, it is easy to mislead extractors to produce biased features. However, if we keep the ratio between majority and minority classes not that high, which is 2:1 here, the bias will be greatly reduced.

In the present study, we construct a CTFM module consisting of two parts: DAE-based feature extraction part (PART I) and QRS-based feature selection part (PART II). In PART I, we choose DAE as our extractor to extract sparse high-level features from all the heartbeats obtained in Section 3.2.1. On the one hand, the

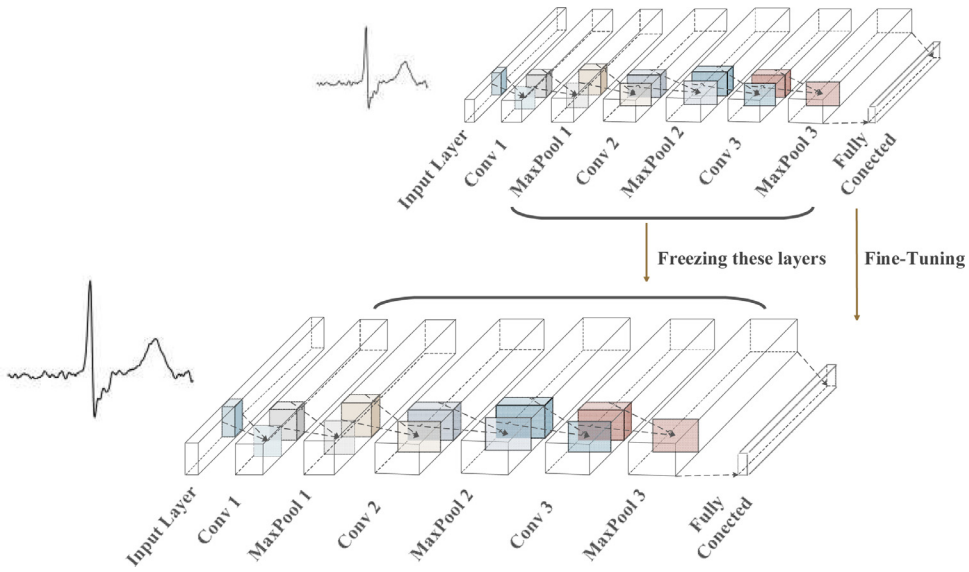


Fig. 5. The architecture of proposed CNN and training schematic diagram.

Table 4
The number of ECG beats used for training, validating and testing in the experiments.

Data div. scheme	Training set	Validation set	Testing set
Intra-patient	120,544	60,286	49,945
Inter-patient	121,927	60,963	49,467

Table 5
Confusion matrix of ECG heartbeat classification results for testing set with the inter-patient scheme.

Ground truth	Classification results			
	N	S	V	F
N	43,465	639	167	241
S	480	1173	134	35
V	163	25	2875	97
F	67	4	18	293

extracted features will not be heavily biased towards majority class, because the number of different classes is in a wide difference. On the other hand, dimensionality of the samples is reduced by DAE to the same size of N as the other part. In PART II, we focus on the selection of important features from minority samples obtained in the previous section, which can be regarded as adding another batch of samples in minority class. QRS complex, the most prominent part in a heartbeat, provides most significant information and reflects ventricular contraction (Šarlija et al., 2017). Therefore, the function of PART II is to intercept the feature area around QRS complex with the length of N from each heartbeat. The samples obtained by CTFM through aforementioned two parts are the samples finally fed to the classifier.

The DAE in PART I is showed in Fig. 4. DAE corrupts the input by randomly dropping-out some input values of randomly selected samples to enhance the robustness of the autoencoder (Wang, Zeng, Ng, & Li, 2017). The hyperparameters are set as follows: the learning rate, coefficient of Gauss noise is set at 0.001 and 0.01 respectively. Training is done in 1000 epochs with batch size of 200. The main purpose of this stage is to convert the original input signal into a shorter segment which can efficiently represent the original signal. Over-complete representation is more conducive to the subsequent operation, thus we only pre-train the DAE in a supervised manner. We do not need to reuse the pre-trained lower layers to create a network for the practical task, and we train it using labeled data. When the loss function gets stable, the features represented by encoder at this time are regarded as the input of the next stage.

In addition, in order to ensure the input of classifier with the same length of N, testing samples should also pass through a same DAE structure.

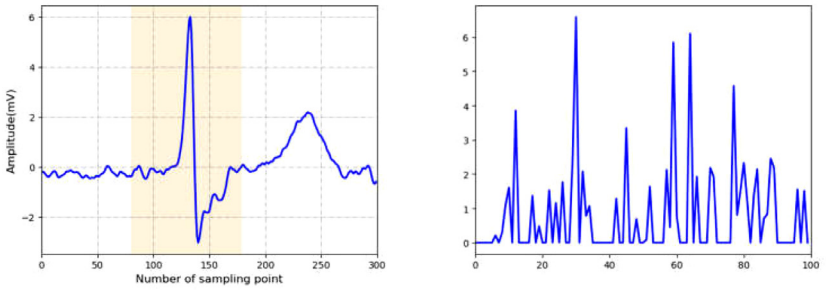
So far, we have used two methods in data level to balance the training data, and the numbers of different heartbeat types on MIT-BIH Arrhythmia Database are shown in Table 2.

3.2.3. Two-phase training (2PT)

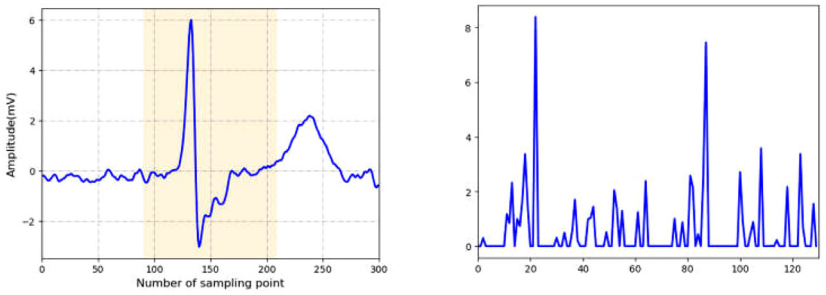
Two-Phase Training (2PT), first proposed in Havaei et al. (2017), is a hybrid method of data and algorithm level to deal with class imbalance problem, which has been shown the efficacy in Buda et al. (2017). In the first stage, we input the balanced data into CNN for training, then the neural network has the ability to distinguish different classes at the level of balanced data. In the second stage, we replace the input data with original unbalanced data, only fine-tuning output layer parameters while keeping those in previous layers unchanged. This way makes the final classification result more convincing.

3.3. Feature extraction and classification

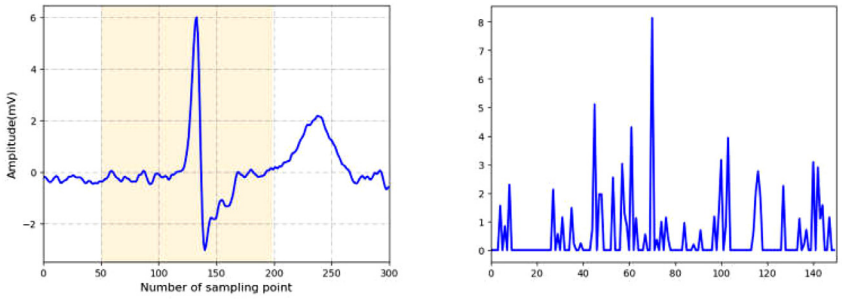
CNN is commonly used for 2-D image classification. In this paper, we modify the traditional structure and design a CNN suitable for processing 1-D ECG signal (Fig. 5). It consists of an input layer, three convolutional layers, three max-pooling layers, a fully-connected layer and an output layer. Each pooling layer follows the corresponding convolutional layer. The input layer, MaxPool1, MaxPool2 are convolved through Eq. (5) with kernels size of 2, 4, 5 and strides of 1, 1, 1 respectively. The kernels size and strides for layers Conv1, Conv2, Conv3 are all set at 2 and outputs of each layer follows Eq. (6) in Section 2.2. These layer in the model use full zero-padding to ensure input and output with the same size. Hence there will be no limitation on the number of convolutional layers that the network is able to contain. Finally, the extracted features are connected with 50 neurons in the fully-connected layer, and then a softmax function is used to classify the beats of N, S, V and F. The parameters above are all obtained through brute force technique. Table 3 gives the details of CNN.



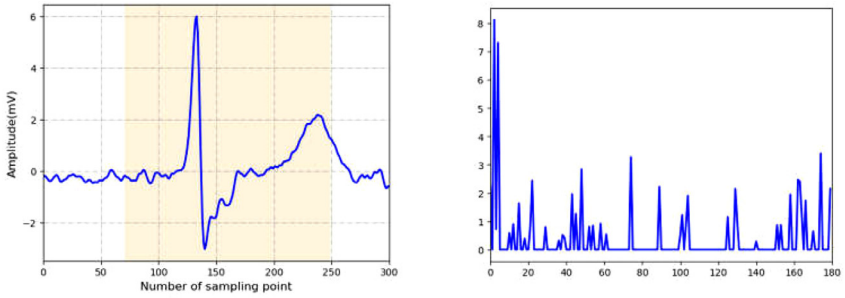
(a)Two parts of input with N=100



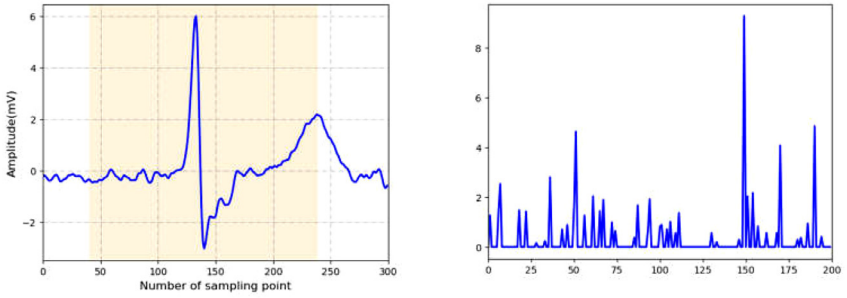
(b) Two parts of input with N=130



(c)Two parts of input with N=150



(d) Two parts of input with N=180



(e) Two parts of input with N=200

Fig. 6. Example of two parts of input in different length.

Table 6
Classification results in terms of SVEB and VEB using 22 testing records of MIT-BIH Arrhythmia Database with the inter-patient scheme.

N Size	SVEB						VEB						MAUC
	Acc	Sen	Spe	Ppr	F_m	GM	Acc	Sen	Spe	Ppr	F_m	GM	
100	94.7	44.3	97.1	51.8	47.8	65.6	97.9	72.6	97.2	81.2	76.6	84.0	92.2
130	95.2	56.7	97.4	**64.0**	60.1	74.3	98.4	87.9	98.6	89.3	88.6	93.1	93.7
150	96.8	61.2	**99.2**	**64.6**	62.9	78.0	**98.8**	90.4	**99.4**	88.6	89.5	**94.8**	**95.3**
180	**97.3**	**64.4**	**98.6**	63.7	**64.0**	**80.0**	**98.8**	**91.0**	**99.3**	**90.0**	**90.1**	**95.1**	**97.8**
200	**97.1**	**64.8**	98.5	62.1	**63.4**	**80.0**	**99.0**	**90.8**	98.9	**88.9**	**89.8**	**94.8**	94.1

The values in the table multiply the original values by 100.
Best two results are highlighted.

Table 7
With the intra-patient scheme, classification results in terms of SVEB and VEB obtained by different methods on the MIT-BIH.

Method	Features	Classifier	Accuracy
Gutiérrez-Gnecchi et al. (2017)	Wavelet	PNN	92.7
Tang and Shu (2014)	WT	QNN	92.8
Martis et al. (2013)	Pan-Tompkin + PCA	NN+LS-SVM	93.0
Acharya et al. (2017)	1D-CNN	Softmax	94.0
Li and Zhou (2016)	WPE+RR	Random Forest	94.6
Zadeh and Khazaee (2011)	CWT	SVM+GA	97.2
Li et al. (2017)	1D-CNN	Softmax	97.5
Proposed[1]	DAE+1D-CNN	Softmax	98.4

Proposed[1] are based on intra-patient scheme for SVEB and VEB detection.

Table 8
With the inter-patient scheme, classification results in terms of SVEB and VEB obtained by different methods on the MIT-BIH.

Method	SVEB				VEB				Overall Accuracy
	Acc	Sen	Spe	Ppr	Acc	Sen	Spe	Ppr	
De et al. (2004)	94.6	75.9	/	38.5	97.4	77.7	/	81.9	85.9
Zhang et al. (2014)	93.3	79.1	93.9	36.0	98.6	85.5	99.5	92.7	86.6
Kiranyaz et al. (2016)	96.4	64.6	98.6	62.1	98.6	95	98.1	89.5	96.6
Jiang and Kong (2007)	96.6	50.6	98.8	67.9	98.1	86.6	99.3	93.3	94.5
Ince et al. (2009)	96.1	62.1	98.5	56.7	97.6	83.4	98.1	87.4	93.6
Proposed[2] (MMNNS)	97.3	64.4	98.6	63.7	98.8	91.0	99.3	90.0	96.6

Proposed[2] are based on inter-patient scheme for SVEB and VEB detection.

Table 9
Different strategies for addressing imbalance with CNN.

Algorithm	Strategy
PCNN	Pure CNN
SMCNN	SMOTE + CNN
BSCNN	BLSM + CNN
BSDC	BLSM + DAE-CNN
BSCP	BLSM + CNN + 2PT
MMNNS	BLSM + CTFM + CNN + 2PT

Previous operations dealt with imbalanced data also greatly expanded the amount of training data, which causes pressure on the accelerated convergence of the model. We use the following tricks to speed up the training and solve vanishing/exploding gradients problems. Xavier Initialization (Glorot & Bengio, 2010) is exploited to initialize the connection weights. Leaky ReLU, a variant of the ReLU function is used as the activation function of convolutional layers and fully-connected layers to avoid dying ReLUs. Referring to Géron (2017), setting $\alpha = 0.2$ in leaky ReLU results in better performance than $\alpha = 0.1$ that we commonly use. These two tricks can significantly overcome the vanishing/exploding gradients problems at the beginning of training. We then use a faster optimizer named Adam Optimizer instead of the regular Gradient Descent Optimizer and we adopt Exponential Decay Learning Rate to optimize the classification error with the base learning rate and learning rate decay rate set to 0.001 and 0.99 respectively.

In this paper, training of the algorithm is done in 800 epochs with a mini batch size of 200. Validation set is used to validate the model each 100 rounds of training. After training CNN with processed balanced data, the model is fine-tuned according to the method described in 3.2.3. The final performance of the testing set is measured by eight metrics defined below.

4. Experiments setup

4.1. Datasets

4.1.1. MIT-BIH arrhythmia database

MIT-BIH arrhythmia database consists of 48 records slightly longer than 30 min of two-channel ECG recordings with digitization rate of 360 Hz. The signals of the two channels were recorded by placing the electrodes at different angles on the chest. The upper signal is a modified limb lead II (MLII), and the lower signal is a modified lead V1, V2, or V5 (Moody & Mark, 2002). ECG data were collected from 47 subjects, including 25 men aged 32-89 and 22 women aged 23-89 (Records 201 and 202 are from the same male subject). More than 100,900 heart beats in the database have been annotated. Only one channel (channel MLII) for each recording is used for the classification task, excluding four records with poor quality (102,104,107,217).

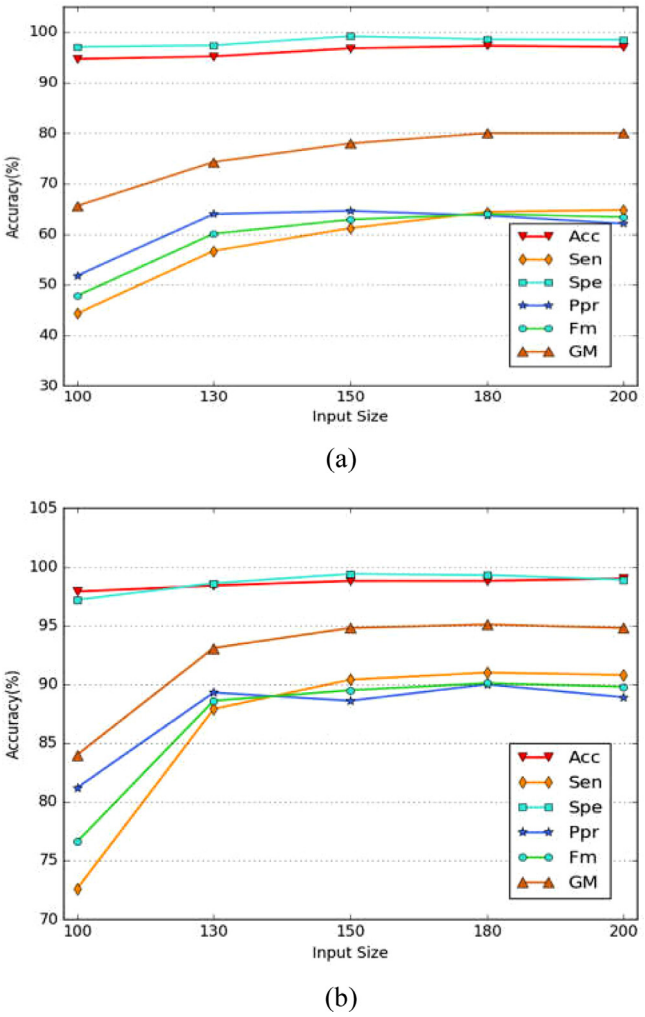


Fig. 7. Acc, Sen, Spe, Ppr, F_m and GM in different length of input for (a) SVEB and (b) VEB.

MIT-BIH Arrhythmia Database is the main database used in the experiment, including determining parameters of each modules in the system, selecting the appropriate data grouping scheme, and training and testing the model. Recommended by ANSI/AAMI EC57:1998/(R) 2008 standard, we should classify all samples into five classes: N, S, V, F, Q. However, considering that the number of samples of Q is too small after removing 4 paced records, which is only 12, we choose to realize classification on other 4 classes.

Two heartbeats grouping schemes are adopted on this database, called intra-patient and inter-patient. Intra-patient is a used by majority of scholars, in which training samples and testing set is randomly selected from all heartbeats. In this grouping method, heartbeats in training set and testing set are likely from the same patient. This may easily lead to biased results due to the heartbeats from same patient have strong correlation and dependence. In the inter-patient scheme, all records are divided into two groups with similar number and similar proportions of each class. Heartbeats in DS1 are all from records: 101, 106, 108, 109, 112, 114, 115, 116, 118, 119, 122, 124, 201, 203, 205, 207, 208, 209, 215, 220, 223 and 230 .Heartbeats in DS2 are all from records: 100, 103, 105, 111, 113, 117, 121, 123, 200, 202, 210, 212, 213, 214, 219, 221, 222, 228, 231, 232, 233 and 234. The heartbeats of two groups will not come from the same subjects. In the intra-patient scheme, 49,953 heartbeats (N44800, S1345, V3412, F396) are selected from the total heartbeats as the training set and the rest as the testing set. In the inter-patient scheme, we use DS1 as the training set and DS2 as the testing set. Processed training set is divided into training set and validation set in proportion of 2:1. The specific data partition is shown in the Table 4.

4.1.2. European ST-T database

European ST-T database (Taddei et al., 1992) consists of 90 annotated excerpts of ambulatory ECG recordings from 79 subjects. The subjects are 70 men aged 30 to 64, and 8 women aged 55 to 71. Each record lasts for two hours, and two signals contained are both sampled at 250 samples per second. Two cardiologists annotated each record beat by beat and for changes in ST segment and T-wave morphology, rhythm and signal quality.

4.1.3. MIT-BIH ST change database

MIT-BIH ST change database (Goldberger et al., 2000) includes 28 ECG recordings with different lengths. The annotation files contain only beat labels; they do not include ST change annotations, as in the European ST-T Database.

Only one channel of both two datasets is selected to conduct the experiments. The heartbeats obtained from these two datasets are fed into the model which we have already trained before. Test results are used for further comparison and evaluation from different aspects.

4.2. Evaluation metrics

To evaluate the classification performance, seven classification assessment metrics and two statistical measures are adopted in this study. The first seven metrics are defined by four representations of *TP*, *TN*, *FP*, *FN*, meaning true positive, true negative, false positive and false negative respectively.

(1) Accuracy
Accuracy (*Acc*) is a standard performance measure for classification, representing the proportion of test examples classified correctly. In learning imbalanced data, it will lead to highly misleading assessment if we regard it as the only evaluation metric, since majority class is overrepresented.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

(2) Sensitivity
Sensitivity (*Sen*), also called Recall, measures the integrity of positive samples being correctly classified.

$$Sen = \frac{TP}{TP + FN} \tag{8}$$

(3) Specificity
Specificity (*Spe*) means the percentage of correctly classified negative tuples.

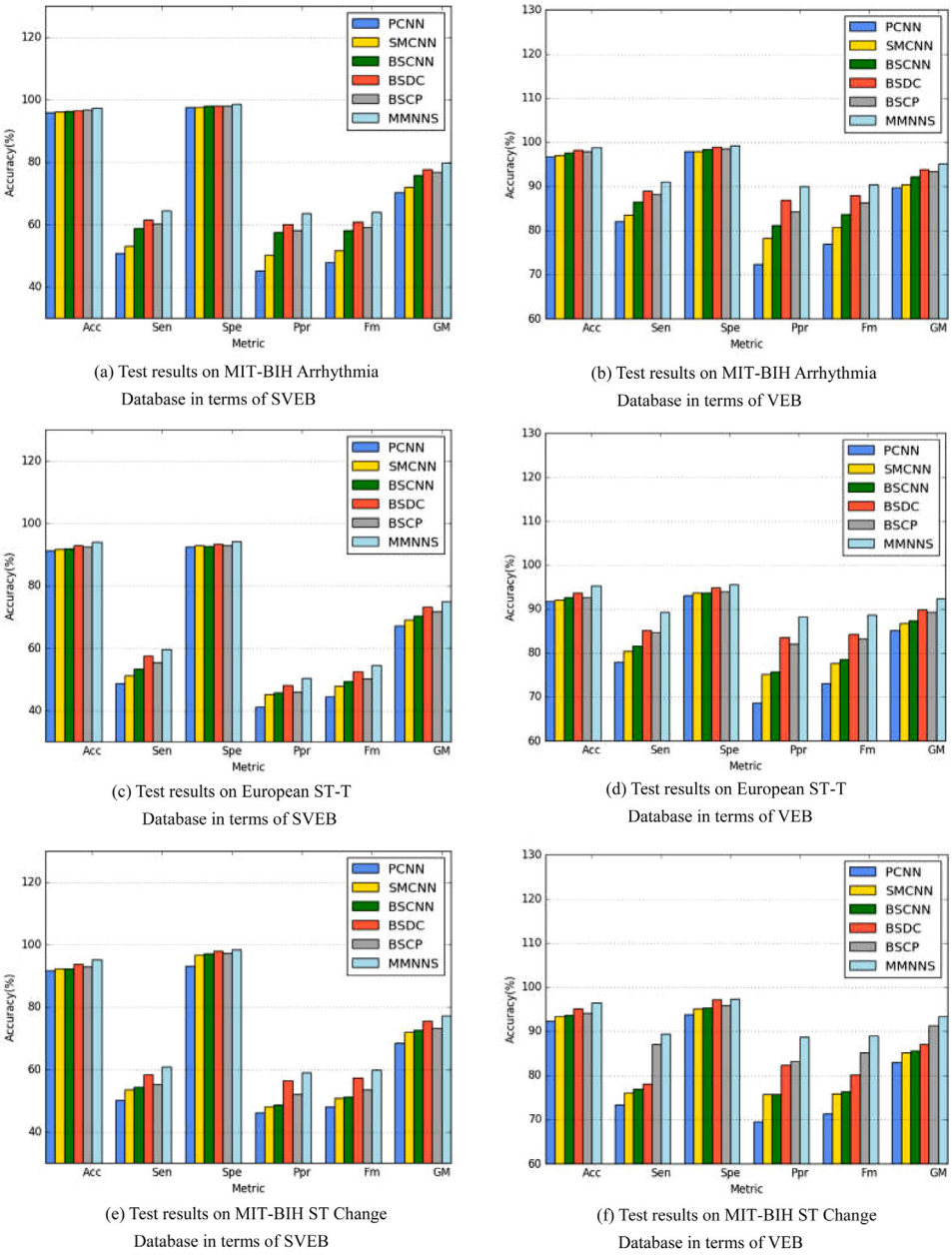
$$Spe = \frac{TN}{TN + FP} \tag{9}$$

(4) Positive predictive rate
Positive predictive rate (*Ppr*), also called Precision, means the proportion of positive samples classified correctly, being sensitive to data distribution.

$$Ppr = \frac{TP}{TP + FP} \tag{10}$$

(5) F-measure
F-measure (F_m) incorporates *Sen* and *Ppr* to express their tradeoff. Parameter β is used to adjust the relative importance of recall *Sen* and *Ppr*, which is usually set to 1.

$$F_m = \frac{(1 + \beta^2) \times Sen \times Ppr}{\beta^2 \times Sen + Ppr} \tag{11}$$



(a) Test results on MIT-BIH Arrhythmia Database in terms of SVEB

(b) Test results on MIT-BIH Arrhythmia Database in terms of VEB

(c) Test results on European ST-T Database in terms of SVEB

(d) Test results on European ST-T Database in terms of VEB

(e) Test results on MIT-BIH ST Change Database in terms of SVEB

(f) Test results on MIT-BIH ST Change Database in terms of VEB

Fig. 8. Comparisons between six algorithms on SVEB and VEB.

(6) G-mean

G-mean (*GM*) can evaluate the degree of inductive bias according to the positive accuracy and negative accuracy, which is widely used in imbalanced classification problems.

$$GM = \sqrt{\frac{TN}{TN+FP} \times \frac{TP}{TP+FN}} \tag{12}$$

(7) Multiclass AUC

Area under the receiver operating characteristic curve (AUC) is a sound measure of discrimination and has been widely used as an evaluation metric for classifiers, especially in imbalanced learning. We use a multi-class modification of basic version of ROC. The multi-class AUC (*MAUC*) is calculated by taking the average of AUCs obtained independently for each class for the binary classification task of distinguishing a given class from all the other (Buda et al., 2017).

(8) Statistical measures

Wilcoxon signed-rank test (Wilcoxon, 1945) and sign test (Dixon & Mood, 1946) are two non-parametric tests adopted in this study to explore the differences between each two algorithms by calculating the p-value.

5. Experiments and analysis

5.1. Experiment 1: Determination of the input feature size

To determine the suitable length of input features in the CTFM and the parameters of final classification model, we perform experiments on five different input lengths ($N = 100, 130, 150, 180, 200$) around QRS complex intercepted from the original heartbeat.

Fig. 6 takes S class as an example to show two parts of the input for different length. The colored region in the left subfigures are the selected features with N sampling points processed by

QRS-based feature selection part (PART II). Those in the right subfigures are the features extracted from DAE-based extraction part (PART I).

Taking $N = 180$ as an example, Table 5 shows the confusion matrix of the classification results on the testing set in inter-patient scheme. Table 5 also indicates that the number of samples in N class is the largest, and the number of misclassified samples is the largest correspondingly. In N beats, 639 are misclassified into class S, 167 are misclassified into class V and 241 are misclassified into class F. In class S, V and F, the number of samples that are wrongly classified into N class is the largest relatively.

To compare the classification performances of five different input lengths comprehensively, we detect the SVEB and VEB classes according to AAMI standard. Table 6 highlights the best two results for each metric. The trends of each metric with different length in the detection of SVEB and VEB are plotted in Fig. 7. When detecting SVEB class, the values of Acc and Spe are generally higher than those of Sen, Ppr, F_m and GM which are consistent with the detection on VEB. It's easy to see that when input length is set to 180 and 200, the best detection results are achieved on SVEB and VEB. This may be explained that larger feature areas contribute more to determine the class of heartbeats.

For the SVEB class, the Acc, Spe, Ppr and F_m with input length N of 180 are all higher than those with $N = 200$. For the VEB class, the gross ACC was 99.0% with $N = 200$, a little higher than those with $N = 180$, which is 98.8%. However, the values of other five metrics are slightly lower than those with $N = 180$. It indicates that larger feature areas may not always achieve good performance, and the location of feature area also plays an important role. We also compare MAUC scores in different input length, and the results in Table 6 indicate that the highest score is achieved at $N = 180$. In a word, the input length of 180 sampling points is the respectively best. Hence, we use it as the final input in this paper.

5.2. Experiment 2: Comparison with the state-of-the-arts on two data grouping schemes

In this section, we conduct experiments on both intra-patient and inter-patient scheme. Intra-patient, as a popular grouping scheme, randomly selects the corresponding number of beats from all hearts as training set and testing set. It has achieved a high classification accuracy in early literature. As is shown in Table 7, compared with the manual feature extraction and traditional machine learning classifier by Martis, Acharya, Mandana, Ray, and Chakraborty (2013), the classification accuracy of Acharya et al. (2017) and Li and Zhou (2016) using CNN has been significantly improved. The accuracy in this paper with intra-patient scheme is 98.4%, remaining top when compared with the related literature. It demonstrates that a series of methods proposed in this paper are reliable. However, since the heartbeats in testing set are likely from the same subject as the heartbeats in training set, strong correlation between the beats of the same subject can easily lead to biased results. That means the final classification result seems to be high, but in fact there remains a deviation.

From Tables 7 and 8, we know that the accuracy of intra-pattern scheme is superior to that of inter-pattern scheme, which is consistent with the previous description. Nevertheless, when it comes to real-life situation, the inter-patient scheme clearly distinguishing each object is more fair and persuasive. All the test beats used in the experiments in Table 8 with inter-patient scheme come from entirely new individuals that differ from the training data. Among them, De et al. (2004) and Zhang, Dong, Luo, Choi, and Wu (2014) adopted the same data grouping method as this paper, using beats of 22 records in DS2 as testing set to show the detection results on SVEB, VEB and the overall accuracy. Jiang and Kong (2007); Kiranyaz, Ince, and Gabbouj (2016) and Ince, Kiranyaz, and Gabbouj (2009) used 24 new records relative to the training set as testing set data to verify the effectiveness of models. As is seen from Table 8, the proposed method yields the best classification result regardless of the overall accuracy of the model or detection for SVEB and VEB separately.

5.3. Experiment 3: Comparison with different strategies for addressing imbalance problem

To show the efficacy of MMNNS, we test the other two datasets on the basis of the previous experiments, and compared it with other five algorithms, including several classical methods dealing with imbalanced learning with CNN. The brief descriptions of the strategy for these algorithms are shown in Table 9.

The first algorithm named PCNN aims to feed the original imbalanced data directly into a pure CNN model without any process. Algorithm SMCNN and BSCNN preprocess the original data by oversampling them using SMOTE and BLSM respectively. On the basis of BSCNN, the classifier CNN is replaced by a cascaded structure of DAE and CNN in algorithm BSDC, which feeds oversampled training data to a DAE at first and then to a CNN. The fifth algorithm is BSCP, which integrates 2PT to BSCNN. Unlike BSCNN, which only fine-tunes the model using balanced training data after oversampling, BSCP introduces an extra fine-tuning phase using the original imbalanced data. The last one is our method: MMNNS.

For the purpose of comparison, we present the test results of each method on SVEB and VEB separately with respect to Acc, Sen, Spe, Ppr, F_m and GM, together with overall accuracy and MAUC score in Tables 10 –12. The variation trends of these metrics in different algorithms on different datasets are shown in Fig. 8. As illustrated in the tables and figures, the performances in all metrics on MIT-BIH Arrhythmia Dataset are obviously better than those on other dataset. This is because our model is trained exactly on MIT-BIH Arrhythmia Dataset. However, no matter which dataset we test on, we can see that the five algorithms show significant improvements, compared with feeding imbalanced original data into CNN directly. The SMCNN and the BSCNN have similar effects, and the latter one is slightly better than the former one because it focuses on the borderline samples that are easier to be misclassified. Algorithm BSDC constructs a cascaded structure combining DAE and CNN, extracting high-level features of heartbeats sequentially. The accuracy of each metric is just a little lower than MMNNS while higher than others. On the basis of BSCNN, BSCP involves a fine-tuning phase using imbalanced data, which improves the final performance somewhat. The algorithm MMNNS achieves the highest overall accuracy and MAUC score, and the performances of the detection on SVEB and VEB are almost always better than other methods.

In the last step, Wilcoxon signed-rank test and sign test are implemented over average accuracy, sensitivity, specificity, positive predictive rate, F-measure, G-mean and MAUC between each two algorithms. We consider MMNNS as the control method, and calculate the p-value to see if there are significant differences between such two algorithms. The smaller the p-value is, the more significant the difference between the two algorithms is. As shown in Table 13, there remain significant differences between MMNNS and the other five algorithms, with p-value lower than 0.05. This makes it clear that the MMNNS significantly outperforms other methods.

6. Conclusion and future work

The heartbeats of ECG signals in the existing datasets are usually imbalanced among different classes, which brings great difficulties in classifying them accurately. However, most scholars mainly pay attention to enhance the overall accuracy of the system

Table 10
Classification results on MIT-BIH arrhythmia database.

Algorithm	SVEB						VEB						Overall Acc	MAUC
	Acc	Sen	Spe	Ppr	F_m	GM	Acc	Sen	Spe	Ppr	F_m	GM		
PCNN	95.9	50.8	97.6	45.1	47.8	70.4	96.8	82.1	97.9	72.4	77.0	89.7	92.5	73.8
SMCNN	96.1	53.2	97.6	50.3	51.7	72.1	97.0	83.6	97.9	78.2	80.8	90.5	93.4	82.6
BSCNN	96.4	58.7	97.9	57.6	58.1	75.8	97.6	86.4	98.3	81.2	83.7	92.2	94.1	86.3
BSDC	96.5	61.6	97.9	60.1	60.8	77.7	98.2	88.9	99.0	86.9	87.9	93.8	95.7	92.1
BSCP	96.8	60.2	98.0	58.2	59.2	76.8	97.9	88.3	98.5	84.3	86.3	93.3	94.5	87.4
MMNNS	97.3	64.4	98.6	63.7	64.0	79.7	98.8	91.0	99.3	90.0	90.5	95.1	96.6	97.8

The values in the table multiply the original values by 100.

Table 11
Classification results on European ST-T database.

Algorithm	SVEB						VEB						Overall Acc	MAUC
	Acc	Sen	Spe	Ppr	F_m	GM	Acc	Sen	Spe	Ppr	F_m	GM		
PCNN	91.2	48.7	92.6	41.1	44.6	67.2	91.8	78.0	93.1	68.7	73.1	85.2	88.4	68.2
SMCNN	91.7	51.2	92.9	45.2	48.0	69.0	92.1	80.5	93.6	75.1	77.7	86.8	90.1	79.9
BSCNN	92.0	53.3	92.8	45.9	49.3	70.3	92.6	81.7	93.6	75.7	78.6	87.4	91.8	82.4
BSDC	92.9	57.6	93.3	48.2	52.5	73.3	93.6	85.1	94.8	83.6	84.3	89.8	93.1	88.3
BSCP	92.6	55.4	93.0	46.1	50.3	71.8	92.7	84.7	93.9	82.0	83.3	89.2	92.4	86.0
MMNNS	94.1	59.7	94.2	50.4	54.7	75.0	95.3	89.2	95.6	88.3	88.7	92.3	93.7	94.1

The values in the table multiply the original values by 100.

Table 12
Classification results on MIT-BIH ST change database.

Algorithm	SVEB						VEB						Overall Acc	MAUC
	Acc	Sen	Spe	Ppr	F_m	GM	Acc	Sen	Spe	Ppr	F_m	GM		
PCNN	91.8	50.3	93.2	46.3	48.2	68.5	92.3	73.4	93.8	69.6	71.4	83.0	90.2	71.5
SMCNN	92.3	53.6	96.8	48.2	50.8	72.0	93.4	76.1	95.2	75.7	75.9	85.1	91.4	81.2
BSCNN	92.4	54.3	97.1	48.7	51.3	72.6	93.6	76.9	95.3	75.8	76.3	85.6	91.9	81.4
BSDC	93.8	58.3	97.9	56.4	57.3	75.5	95.2	78.1	97.2	82.3	80.1	87.1	93.7	89.6
BSCP	93.0	55.2	97.3	52.1	53.6	73.3	94.1	87.0	95.9	83.2	85.1	91.3	92.8	85.3
MMNNS	95.2	60.8	98.4	59.1	59.9	77.3	96.5	89.4	97.4	88.6	89.0	93.3	94.3	93.2

The values in the table multiply the original values by 100.

Table 13
Wilcoxon signed-rank test and sign test of different algorithms (taking MMNNS as the control method).

Algorithm	Wilcoxon signed-rank test p-value	Sign test p-value	Hypothesis ($\alpha = 0.05$)
PCNN	0.0003	0.0000	Rejected for MMNNS
SMCNN	0.0010	0.0002	Rejected for MMNNS
BSCNN	0.0047	0.0017	Rejected for MMNNS
BSDC	0.0287	0.0023	Rejected for MMNNS
BSCP	0.0046	0.0006	Rejected for MMNNS

losing sight of the fact that other metrics are also needed to consider at the same time to make it more reliable. In addition, even though some scholars have already taken some measures to balance the original data, they only adopt simple oversampling methods, not considering the specific characteristics of the ECG itself.

In this paper, we propose a novel multi-module neural network system named MMNNS for ECG heartbeats classification. We focus on the imbalance problem of heartbeats which is not well addressed before. A system consisting of four submodules is constructed, and a series of balancing measures which combine resampling, data feature, and algorithm together are taken on the original data. DAE and CNN are used as feature extractors of ECG heartbeats, from which the feature representations are fed into softmax regression. We then realize the classification on four classes of N, S, V and F according to AAMI standard. We train our model on MIT-BIH Arrhythmia Database using inter-patient scheme especially, and fine-tuning the model by 2PT algorithm. Then, we validate the model on other two datasets with five classical methods for addressing imbalance problem. Finally, seven classification assessment metrics and two statistical measures are used to evaluate the classification results. Our designed system achieves an accuracy of 97.3%, a sensitivity, specificity and positive predictive rate of 64.4%, 98.6% and 63.7%, a F-measure of 64% and G-mean of 79.7% for the SVEB class. The VEB accuracy is 98.8%, sensitivity, specificity, positive predictive rate, F-measure and G-mean are 91.0%, 99.3%, 90.0%, 90.5%, 95.1% respectively. The overall accuracy and MAUC score are 96.6% and 0.978. Such results are superior over most state-of-the-arts. Specifically, MMNNS proposed in this paper not only provides a solution for heartbeats classification, but also shows a way for other time series classification even image classification.

In the future works, other types of neural networks could be taken into account to enhance the classification performance. Besides, more attention should be paid to the structure and nature of minority classes to gain a better insight into the source of learning difficulties. Furthermore, the approach proposed in this paper could be extended to solve the imbalance problem in other related fields. It is possible to evaluate this approach on more datasets and make more comprehensive comparisons with other approaches.

Credit Author Statement

Ph.D Jing Jiang is in charge of the whole section of the paper, and has completed the design of the algorithm, the implementation of the algorithm, the writing and revising of the original draft. Ph.D Huaifeng Zhang is mainly responsible for the test of the algorithm, the validation of the algorithm and visualization

of experimental results. Ph.D Dechang Pi is responsible for the review and has provided the fund support. Ph.D Chenglong Dai is responsible for the review.

Acknowledgements

The research work is supported by National Natural Science Foundation of China (U1433116), the Fundamental Research Funds for the Central Universities (NP2017208), and the Foundation of Graduate Innovation Center in NUAA (kfjj20181605).

References

Acharya, U. R., Fujita, H., Sudarshan, V. K., Sree, V. S., Eugene, L. W. J., & Ghista, D. N. (2015). An integrated index for detection of sudden cardiac death using discrete wavelet transform and nonlinear features. *Knowledge-Based Systems, 83*, 149–158.

Acharya, U. R., Oh, S. L., Hagiwara, Y., Tan, J. H., Adam, M., & Gertych, A. (2017). A deep convolutional neural network model to classify heartbeats. *Computers in Biology & Medicine, 89*, 389–396.

Bae, S. H., & Yoon, K. J. (2015). Polyp detection via imbalanced learning and discriminative feature learning. *IEEE Transactions on Medical Imaging, 34*, 2379–2393.

Braytee, A., Liu, W., & Kennedy, P. (2016). A cost-sensitive learning strategy for feature extraction from imbalanced data. In *International Conference on Neural Information Processing* (pp. 78–86). Kyoto, Japan.

Buda, M., Maki, A., & Mazurowski, M. A. (2017). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks, 106*, 249–259.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357.

Chawla, N. V., Lazarevic, A., Hall, L. O., & Bowyer, K. W. (2003). SMOTEBoost: Improving prediction of the minority class in boosting. In *European Conference on Principles and Practice of Knowledge Discovery in Databases* (pp. 107–119). Cavtat-Dubrovnik, Croatia.

Chui, K. T., Tsang, K. F., Chi, H. R., Ling, B. W. K., & Wu, C. K. (2016). An accurate ECG-based transportation safety drowsiness detection scheme. *IEEE Transactions on Industrial Informatics, 12*, 1438–1452.

De, C. P., O'Dwyer, M., & Reilly, R. B. (2004). Automatic classification of heartbeats using ECG morphology and heartbeat interval features. *IEEE Transactions on Bio-medical Engineering, 51*, 1196–1206.

Dixon, W. J., & Mood, A. M. (1946). The statistical sign test. *Publications of the American Statistical Association, 41*, 557–566.

Elkan, C. (2001). The foundations of cost-sensitive learning. In *Seventeenth International Joint Conference on Artificial Intelligence* (pp. 973–978). Seattle, Washington, USA.

Géron, A. (2017). *Hands-on machine learning with Scikit-learn and Tensorflow O'Reilly*. Media.

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *Journal of Machine Learning Research, 9*, 249–256.

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., & Peng, C. K. (2000). Circulation. *PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals, 101*, 215–220.

Goodfellow, Ian, B. , Y., & Courville, Aaron (2016). *Deep Learning*. MIT Press.

Golrizkhatami, Z., & Acan, A. (2018). ECG classification using three-level fusion of different feature descriptors. *Expert Systems with Applications, 114*, 54–64.

Guo, H., Li, Y., Shang, J., Gu, M., Huang, Y., & Gong, B. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications, 73*, 220–239.

Guo, H., & Viktor, H. L. (2004). Learning from imbalanced data sets with boosting and data generation: The DataBoost-IM approach. *Acm Sigkdd Explorations Newsletter, 6*, 30–39.

Gutiérrez-Gnecchi, J. A., Morfin-Magaña, R., Lorias-Espinoza, D., Reyes-Archundia, E., Méndez-Patiño, A., & Castañeda-Miranda, R. (2017). DSP-based arrhythmia classification using wavelet transform and probabilistic neural network. *Biomedical Signal Processing & Control, 32*, 44–56.

Han, H., Wang, W. Y., & Mao, B. H. (2005). Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. *Lecture Notes in Computer Science, 3644*, 878–887.

Haritha, C., Ganesan, M., & Sumesh, E. P. (2016). A survey on modern trends in ECG noise removal techniques. In *International Conference on Circuit, Power and Computing Technologies* (pp. 1–7). Nagercoil, India.

Havaei, M., Davy, A., Wardefarley, D., Biard, A., Courville, A., & Bengio, Y. et al. (2017). Brain tumor segmentation with deep neural networks. *Medical Image Analysis, 35*, 18–31.

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*, 1415–1426.

Isin, A., & Ozdalili, S. (2017). Cardiac arrhythmia detection using deep learning. *Procedia Computer Science, 120*, 268–275.

Jiang, W., & Kong, S. G. (2007). Block-based neural networks for personalized ECG signal classification. *IEEE Transactions on Neural Networks, 18*, 1750–1761.

Jo, T., & Japkowicz, N. (2004). Class imbalances versus small disjuncts. *ACM SIGKDD Explorations Newsletter, 6*, 40–49.

Kaplan Berkaya, S., Uysal, A. K., Sora Gunal, E., Ergin, S., Gunal, S., & Gulmezoglu, M. B. (2018). A survey on ECG analysis. *Biomedical Signal Processing and Control, 43*, 216–235.

Khan, S. H., Hayat, M., Bennamoun, M., Sohel, F. A., & Togneri, R. (2018). Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems, 29*, 3573–3587.

Khandoker, A. H., Palaniswami, M., & Karmakar, C. K. (2009). Support vector machines for automated recognition of obstructive sleep apnea syndrome from ECG recordings. *IEEE Transactions on Information Technology in Biomedicine, 13*, 37–48.

Kiranyaz, S., Ince, T., & Gabbouj, M. (2016). Real-time patient-specific ECG classification by 1-D convolutional neural networks. *IEEE Transactions on Biomedical Engineering, 63*, 664–675.

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*, 436–444.

Lee, H. J., & Cho, S. (2006). The novelty detection approach for different degrees of class imbalance. In *13th International Conference on Neural Information Processing* (pp. 21–30). Hongkong, China.

Li, D., Zhang, J., Zhang, Q., & Wei, X. (2017). Classification of ECG signals based on 1D convolution neural network. In *IEEE International Conference on E-Health Networking, Applications and Services* (pp. 1–6). Dalian, China.

Li, Q., Rajagopalan, C., & Clifford, G. D. (2014). Ventricular fibrillation and tachycardia classification using a machine learning approach. *IEEE Transactions on Bio-medical Engineering, 61*, 1607–1613.

Li, T., & Zhou, M. (2016). ECG classification using wavelet packet entropy and random forests. *Entropy, 18*, 285.

Li, Y., Guo, H., Liu, X., Li, Y., & Li, J. (2016). Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data. *Knowledge-Based Systems, 94*, 88–104.

Liu, X. Y., Wu, J., & Zhou, Z. H. (2009). Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems Man & Cybernetics Part B, 39*, 539–550.

Maldonado, S., & López, J. (2014). Imbalanced data classification using second-order cone programming support vector machines. *Pattern Recognition, 47*, 2070–2079.

Maldonado, S., López, J., & Vairetti, C. (2019). An alternative SMOTE oversampling strategy for high-dimensional datasets. *Applied Soft Computing, 76*, 380–389.

Maratea, A., Petrosino, A., & Manzo, M. (2014). Adjusted F-measure and kernel scaling for imbalanced data learning. *Information Sciences, 257*, 331–341.

Martis, R. J., Acharya, U. R., Mandana, K. M., Ray, A. K., & Chakraborty, C. (2013). Cardiac decision making using higher order spectra. *Biomedical Signal Processing & Control, 8*, 193–203.

Martis, R. J., Acharya, U. R., Prasad, H., Chua, C. K., & Lim, C. M. (2013). Automated detection of atrial fibrillation using Bayesian paradigm. *Knowledge-Based Systems, 54*, 269–275.

Moepya, S. O., Akhoury, S. S., & Nelwamondo, F. V. (2015). Applying cost-sensitive classification for financial fraud detection under high class-imbalance. In *IEEE International Conference on Data Mining Workshop* (pp. 183–192). Shenzhen, China.

Moody, G. B., & Mark, R. G. (2002). The impact of the MIT-BIH arrhythmia database. *IEEE Engineering in Medicine & Biology Magazine, 20*, 45–50.

Ng, W. W. Y., Zeng, G., Zhang, J., Yeung, D. S., & Pedrycz, W. (2016). Dual autoencoders features for imbalance classification problem. *Pattern Recognition, 60*, 875–889.

Pan, J., & Tompkins, W. J. (2007). A real-time QRS detection algorithm. *IEEE Transactions on Biomedical Engineering, 32*, 230–236.

Pławiak, P. (2018). Novel methodology of cardiac health recognition based on ECG signals and evolutionary-neural system. *Expert Systems with Applications, 92*, 334–349.

Rahhal, M. M. A., Bazi, Y., Alhichri, H., Alajlan, N., Melgani, F., & Yager, R. R. (2016). Deep learning approach for active classification of electrocardiogram signals. *Information Sciences, 345*, 340–354.

Raj, S., & Ray, K. C. (2018). Sparse representation of ECG signals for automated recognition of cardiac arrhythmias. *Expert Systems with Applications, 105*, 49–64.

Raj, V., Magg, S., & Wermter, S. (2016). Towards effective classification of imbalanced data with convolutional neural networks. In *7th IAPR TC3 Workshop on Artificial Neural Networks in Pattern Recognition* (pp. 150–162). Ulm, Germany.

Šarlija, M., Jurišić, F., & Popović, S. (2017). A convolutional neural network based approach to QRS detection. In *International Symposium on Image and Signal Processing and Analysis* (pp. 121–125). Ljubljana, Slovenia.

Shen, L., Lin, Z., & Huang, Q. (2016). Relay backpropagation for effective learning of deep convolutional neural networks. In *European Conference on Computer Vision* (pp. 467–482). Amsterdam, The Netherlands.

Sun, Y., Kamel, M. S., Wong, A. K. C., & Wang, Y. (2007). Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition, 40*, 3358–3378.

Taddei, A., Distante, G., Emdin, M., Pisani, P., Moody, G. B., & Zeelenberg, C. (1992). The European ST-T database: Standard for evaluating systems for the analysis of ST-T changes in ambulatory electrocardiography. *European Heart Journal, 13*, 1164–1172.

Tang, X., & Shu, L. (2014). Classification of electrocardiogram signals with RS and quantum networks neural. *International Journal of Multimedia & Ubiquitous Engineering, 9*, 363–372.

Wang, H., Naghavi, M., Allen, C., Barber, R. M., & Bhutta, Z. A. (2015). Global, regional, and national age–sex specific all-cause and cause-specific mortality for 240 causes of death, 1990–2013: A systematic analysis for the Global Burden of Disease Study 2013. *The Lancet, 385*, 117–171.

Wang, S., Liu, W., Wu, J., Cao, L., Meng, Q., & Kennedy, P. J. (2016). Training deep neural networks on imbalanced data sets. In *International Joint Conference on Neural Networks* (pp. 4368–4374). Vancouver, Canada.

Wang, T., Zeng, G., Ng, W. W. Y., & Li, J. (2017). Dual denoising autoencoder features for imbalance classification problems. In *IEEE International Conference on Internet of Things* (pp. 312–317). Exeter, UK.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin, 1*, 80–83.

Yücelbaş, Ş., Yücelbaş, C., Tezel, G., Özşen, S., Küççüktürk, S., & Yosunkaya, Ş. (2017). Pre-determination of OSA degree using morphological features of the ECG signal. *Expert Systems with Applications, 81*, 79–87.

Zadeh, A. E., & Khazaee, A. (2011). High efficient system for automatic classification of the electrocardiogram beats. *Annals of Biomedical Engineering, 39*, 996–1011.

Zhai, X., & Tin, C. (2018). Automated ECG classification using dual heartbeat coupling based on convolutional neural network. *IEEE Access, 6*, 27465–27492.

Zhang, C., Gao, W., Song, J., & Jiang, J. (2016). An imbalanced data classification algorithm of improved autoencoder neural network. In *Eighth International Conference on Advanced Computational Intelligence* (pp. 95–99). ChangMai, Thailand.

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopezpaz, D. (2018). mixup: Beyond empirical risk minimization. *International Conference on Learning Representations* Vancouver, Canada.

Zhang, Z., Dong, J., Luo, X., Choi, K. S., & Wu, X. (2014). Heartbeat classification using disease-specific feature selection. *Computers in Biology & Medicine, 46*, 79–89.

Zhou, Z. H., & Liu, X. Y. (2006). Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge & Data Engineering, 18*, 63–77.