Biomedical Signal Processing and Control 62 (2020) 102091

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

CraftNet: A deep learning ensemble to diagnose cardiovascular diseases

Yong Li [a], Zihang He [a,*], Heng Wang [a], Bohan Li [b], Fengnan Li [a], Ying Gao [a], Xiang Ye [a]

[a] School of Electronic Engineering, Beijing University of Posts and Telecommunications, 10 Xitucheng Road, Haidian District, Beijing 100876, China
[b] School of Science, Beijing University of Posts and Telecommunications, 10 Xitucheng Road, Haidian District, Beijing 100876, China

ARTICLE INFO

Article history:
Received 9 March 2020
Received in revised form 29 June 2020
Accepted 19 July 2020
Available online 26 June 2020

Keywords:
Cardiovascular diseases
Deep learning
Handcraft features

ABSTRACT

The early diagnose of cardiovascular diseases (CVDs) is important and has attracted a lot of research attention. It can rescue more than 17 million people a year or alleviate their symptoms. Research interest has been devoted to the handcraft features and deep features for diagnosing CVDs from electrocardiograph (ECG). However, existing classifiers on handcraft features lacked the robust classification ability, while the deep neural networks are strongly affected by data imbalance. This paper proposed designing a simple architecture of deep neural network, CraftNet, for accurately recognizing the handcraft features. It assembled multiple child classifiers according to decision directed acyclic graph. The classifiers have a tailored structure for classifying the handcraft features, with a mixed loss function, named P-S loss, to optimize it. CraftNet has the advantages of both handcraft features and deep learning methods, i.e., it has a stronger classification ability and is less affected by data imbalance. The proposed CraftNet was tested on the public MIT-BIH dataset. Experimental results showed that it achieved the sensitivity 88.16%, 85.37%, 94.53%, and 88.92% for four categories, and increased the average sensitive accuracy from 86.82% to 89.25%, verifying the robust recognition ability of CraftNet.

1. Introduction

Cardiovascular diseases (CVDs) take life from more than 17 million people every year [22], and therefore it becomes an imperative task to early diagnose CVDs for a better recovery. To accomplish this task, one-lead portable electrocardiograph (ECG) devices such as wristbands and chest patches have been developed and widely utilized. ECG is also a non-invasive technique commonly employed by cardiologists in their clinic practice. It allows for people to monitor themselves in a real time manner, providing the signal carrying the information on whether CVDs are happening or not. Then, a natural problem arises on how to automatically identify the cardiac rhythm abnormalities, since arrhythmias may need immediate intervention. Because of the high mortality rate of CVDs, it has been attracting more and more research attention in the field of medical diagnosis.

An ECG arrhythmia classification approach typically consists of four stages, pre-processing, beat segmentation, feature extraction, and classification. The pre-processing stage mainly includes baseline removal, attenuating or denoising the high-frequency noise, and signal normalization. At the segmentation stage, the ECG signals are partitioned into shorter segments, e.g., at beat level, which can better describe the electrical activity. Then, features are devised and extracted out of each beat and input to classifiers for determining the type of arrhythmias. In literature, a lot of algorithms have been proposed to pre-process and segment ECG signals and achieved rather accurate result for the subsequent classification task. These algorithms include [3,35,19], and other algorithms [17,1,28,24,10,11,16,30,15,33,8,13].

Thus, current research attention on ECG classification is mainly paid to feature extraction and the classification. Feature extraction plays a key role in ECG signal detection and recognition as it directly determines the recognition accuracy of the classifiers. There are roughly two groups of methods, deep learning based methods and handcraft methods.

Deep neural networks typically work in an end-to-end way, i.e., learn the features from annotated samples and classify or detect the learned features in a single architecture. Unfortunately, although they excel at learning feature representation and have yielded a very competitive performance on a broad range of appli-

* Corresponding author.
E-mail addresses: yli@bupt.edu.cn (Y. Li), zhe@bupt.edu.cn (Z. He), Whzyf951620@163.com (H. Wang), bohan@bupt.edu.cn (B. Li), lifengnanbupt@outlook.com (F. Li), gaoying@bupt.edu.cn (Y. Gao), xye@bupt.edu.cn (X. Ye).

https://doi.org/10.1016/j.bspc.2020.102091
1746-8094/

cations, they suffer from the problem of data imbalance which often happens on ECG signals, since the number of normal data is dramatically larger than that of arrhythmic data.

Handcraft features are obtained by devising rules according to the human understanding of scenes and yielded a rather good performance on ECG classification [4]. They are usually fed to support vector machine (SVM) [27] classifiers for classifying ECG signals. The methods employing handcraft features and SVM classifiers have generated a competitive classification performance to the deep learning based methods. Nevertheless, the SVM can be substituted with a deep neural network composed only of a number of fully-connected layers equipped with proper normalizer. This indicates that such methods still have room for the further performance improvement.

In observance of the drawback of deep learning methods and handcraft methods, this paper proposes an approach that takes advantage of both methods. In particular, we designed an architecture of deep neural networks in replacement of the SVMs but fed the handcraft features to it. Main contributions include the following three aspects.

- This paper proposed utilizing handcraft features to alleviate the problem of data imbalance while employing specially designed neural networks as the classifier to improve the classification performance.
- Inspired by the principle of SVMs, a P-S loss (pseudo SVM loss) was designed to enhance the generalization ability.
- Different topologies were employed assembling multiple trained classifiers to effectively cope with data imbalance.

When adopting decision directed acyclic graph (DDAG) [25] topology, this paper performs considerably. The rest of the paper is arranged as follows. Section 2 reviews the development of recent detection algorithms. Section 3 details the MIT-BIH arrhythmia database [20] and the improvements of the CraftNet. Section 4 demonstrates the effectiveness of employing the CraftNet on the database. Section 5 discusses advantages and drawbacks of the CraftNet and Section 6 reviews the whole paper.

2. Related works

ECG records the electrical activity of heart and 12-lead ECG signals are often used for detecting arrhythmias. Among the 12 leads is the lead II which is the most commonly used one for rhythm strip [18]. It gives the important waves including P, Q, R, S and T that represent the defection caused by different reasons. Various types of arrhythmias can be detected based on the analysis of these waves from the lead II, and/or the single analysis of all 12 leads.

Since AlexNet [9] yielded the highest classification accuracy on ImageNet in 2012, deep neural networks have shown an intriguing ability to extract high level features from raw data, and achieved a satisfying performance in a wide range of fields including computer vision, early tumor identification [5], and face recognition, etc. Mar et al. [17] used a multi-layer perception (MLP) to extract features and classified them by linear discriminant analysis. Acharya et al. [1] used a 9 layers 1-dimensional neural network to classify MIT-BIH database which is classified into noisy and noisy-free parts. Sellami and Hwang [28] employed a batch-weighted loss function in 1-dimensional neural network to alleviate data imbalance to some extent. However, data imbalance has a large influence on the performance of neural networks classifying ECG signals. The number of normal data is often dramatically larger than that of illness data, which makes the neural networks extract features that tend to correctly classify the normal data at the cost of incorrectly classifying the illness data.

Compared with the deep features learned by neural networks, handcraft features are probably less affected by data imbalance. In literature, a series of reliable handcraft features have been designed for ECG classification. De Chazal et al. [3] extracted a set of handcraft features from ECG signals including R-R, HOS, wavelet, Morphology, and LBP, with which they classified the ECG signals by a linear classifier. Zhang et al. [35] proposed a method that rejects useless features by support vector machines (SVMs) based on one versus one rule (OVO). Mondéjar-Guerra et al. [19] used features filtered designed by Zhang [35] and constructed ensemble of SVM to classify them. The classifier has also been improved. Park et al. [24] presented hierarchical SVM to lower the level of classification difficulty according to the similarity between categories. De Lannoy et al. [10] proposed weighted SVM to further address data imbalance. After that, De Lannoy et al. [11] again proposed weighted conditional random fields (CRF) to replace SVM and yielded a better result. Luz et al. [16] employed the optimum-path forest (OPF) classifier to classify different categories and compared it with SVM, MLP, and Bayesian.

Other methods mainly deal with three categories, N, S, and V, including [30,15,33,8,13]. These methods also used handcraft features and fed them in SVM or LD classifier.

Since a simple architecture of fully-connected layers can approximate or surpass SVM classifier in the ECG classification problem, the handcraft features can be classified by the architecture of artificial neural networks to further improve the performance. Such a network has a better ability to determine the contribution weight of features and the hyperplane between different categories. In light of the above methods, this paper proposes designing architecture of fully-connected neural networks as classifiers on handcraft features, named CraftNet. Also, this paper devises a novel P-S loss that can effectively enlarge the sample distance and enhance the generalization ability. Finally, an ensemble of the designed classifiers is constructed by DDAG and used for classifying four categories of ECG signals.

3. Material and methods

3.1. Material

This section details the Massachusetts Institute of Technology-BethIsrael Hospital (MIT-BIH) arrhythmia database [20] employed to train and test our classification models, and features extracted from the database for input.

3.1.1. Database

MIT-BIH is publicly available on Physionet [7]. It contains 48 30-min-long ECG records sampled from 47 patients at 360 Hz. Each record comprises information from two leads including MLII and one class in V1, V2, V4, and V5. Every heartbeat has P, Q, R, S, and T waves as shown in Fig. 1 and is annotated by Cardiologists with one of four categories including Normal beat (N), Supraventricular ectopic beat (S), Ventricular ectopic beat (V), Fusion beat (F), and Unknown beat (be discarded for practically non-existent). Since heartbeats of some records totally belong to N category and a large proportion of heartbeats of CVDs records belong to N category (intermittent abnormal), serious data imbalance exists in this database. Nearly 90% heartbeats belong to N category while the remaining 3%, 6%, and 1% heartbeats belong to S, V, and F category, respectively. As recommended by Association for the Advancement of Medical Instrumentation (AAMI), the MIT-BIH database is divided into two parts for training and testing [3]. Training part named DS1 includes patient records 101, 106, 108, 109, 112, 114, 115, 116, 118, 119,122, 124, 201, 203, 205, 207, 208, 209, 215, 220, 223, and 230, and testing part named DS2 includes patient records

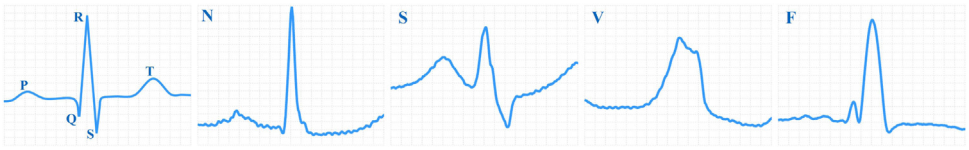

Fig. 1. The first graph represents the P, Q, R, S, and T complexes in a heartbeat. The next four graphs represent curves of N, S, V and F categories randomly sampled from the dataset, respectively.

Table 1
The distribution of the input vectors to the N, S, V and F categories in DS1 for training and DS2 for testing.

Categories	N	S	V	F	Total
DS1	45,839	944	3788	414	50,985
DS2	44,231	1837	3219	388	49,675
DS1+DS2	90,070	2781	7007	802	100,660

100, 103, 105, 111, 113, 117, 121, 123, 200, 202, 210, 212, 213, 213, 219, 221, 222, 228, 231, 232, 233, and 234. The two parts both contain total former four MIT-BIH categories. The number of heartbeats per category in DS1 and DS2 are detailed in Table 1.

The metrics employed for evaluating different methods on the database are listed in Fig. 2, including Positive productivity, sensitivity, and specificity that represent correct diagnosis rate, the diagnosis (recall) rate, and the misdiagnosis rate, respectively. They are of vital importance to evaluate the ECG recognition.

3.1.2. Handcraft features

R-R intervals (RR) [13,26], Wavelet [17,2], higher order statistics (HOS) [23,10], and morphological (Morph) [3,35] features have been widely utilized in the field of ECG recognition. This paper also adopted the four features and the feature extracted way in [19] for straightly comparison. The four features are extracted from the 180 points around heartbeats of MLII lead data (90 points before and after R-peak respectively), which rightly contains the P, Q, R, S, and T complexes in a heartbeat. They form a sum-up 45-D descriptor. Specifically,

R-R intervals. RR encodes four distance measures computed from the time between consequent beats. Pre-RR (the distance between actual heartbeat and the previous one), Post-RR (the distance between actual heartbeat and the next one), Local-RR (containing the average of 10 previous Pre-RRs), and Global-RR (containing the average of Pre-RR values produced in the last 20 min). The four distance measures and their normalized values make up an 8-D descriptor.

Wavelet transform. Wavelet transform is an ideal tool for time–frequency analysis and signal processing. The feature we use encodes data transformed from the Daubechies wavelet function (db1) with 3 levels of decomposition, forming a 23-D descriptor.

Higher order statistics. HOS (higher order statistics) here is the cumulants of the second, third, and fourth order. Each beat is divided into 5 intervals and their kurtosis and skewness values over each other are computed, forming a 10-D descriptor.

Morph. Morph contains amplitude information at four intervals in a heartbeat including the maximum in [0, 40] and [150, 180], and the minimum in [75, 85] and [95, 105].

3.2. Methods

Existing methods for recognizing ECG signals typically include three steps, extracting features from raw ECG signals, classifying the extracted features, and optimizing the parameters of the classifier through minimizing the loss function. In this work, we will continue to use the features of wavelet, HOS, and focus on the design of the classifier and loss function. Specifically, the features of varying dimensionality are firstly expanded and squeezed to be fed to the deep neural network. Then, a classifier of DDAG structure combining two-category child classifiers is constructed for accurate classification. Finally, a loss function is designed to further enhance the generalization ability of the structure.

3.2.1. Expanding and squeezing dimensions

Existing methods which use the neural network always extract features from raw ECG signals. However, serious data imbalance exists in the MIT-BIH database. The number of normal data is often dramatically larger than that of illness data, which makes the neural networks extract features that tend to correctly classify the normal data at the cost of incorrectly classifying the illness data. To alleviate this problem, handcraft features which are less affected by data imbalance are reliable substitutions.

The handcraft features are much less affected by data imbalance and are employed in this work. In [19], four features were used, 8-D RR, 23-D Wavelet, 10-D HOS, and 4-D Morph. To make best use of the four features, a straightforward way is to simply concatenate the four features outside the classifier, forming a 45-dimensional feature, and then inputs the concatenated feature to a classifier shown in Fig. 3(a).

Unfortunately, our experiment shows, the accuracy of classifying normal and abnormal categories drops from 90% to 60% by using the concatenated feature. A deep analysis reveals that the weight for each feature learned by the classifier is proportional to its dimensionality in case of concatenating them together. For example, wavelet which has 23 dimensions occupies almost half of the total concatenated feature. So, if wavelet does not contain useful information, a large part of the concatenated feature will be noise. This first-concatenated structure makes the feature of a larger dimensionality contributes more to the final classification decision, which is not desired since the semantic information contained in a feature may not be positively related to its dimensionality.

To address this problem, the topology first connects each kind of feature to a 128-dimensional full connected layer. This step not only expands the dimensions but also keeps the weight of each kind to the same start line and make the weight independent of the dimension. After that, all 128-dimensional features are concatenated and input to a 2048-dimensional layer. Then 2048-dimensional features are squeezing to 128-dimensional features and 2-dimensional features in order. The final probabilities are calculated through a Sigmoid layer. The 128 dimension is decided for the consideration of both performance and speed after experiments, and the 2048 dimension is also obtained by comparing with dimension 512, 1024, and 3072.

3.2.2. P-S loss

This section presents the P-S loss designed for the classifier that will be discussed in the next section. To clearly explain the designed loss function, we will briefly revisit the principle of SVM [27]. A binary SVM classifier typically increases the dimensionality of features in order to search a separating hyperplane in the high-dimensional space. In general, there exist a lot of such separating hyperplanes. The separating performance of a hyperplane is often

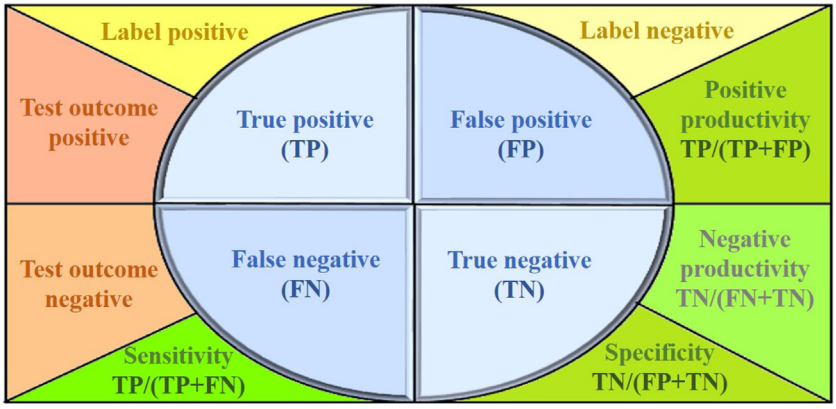

Fig. 2. Evaluation metrics for comparing different methods.

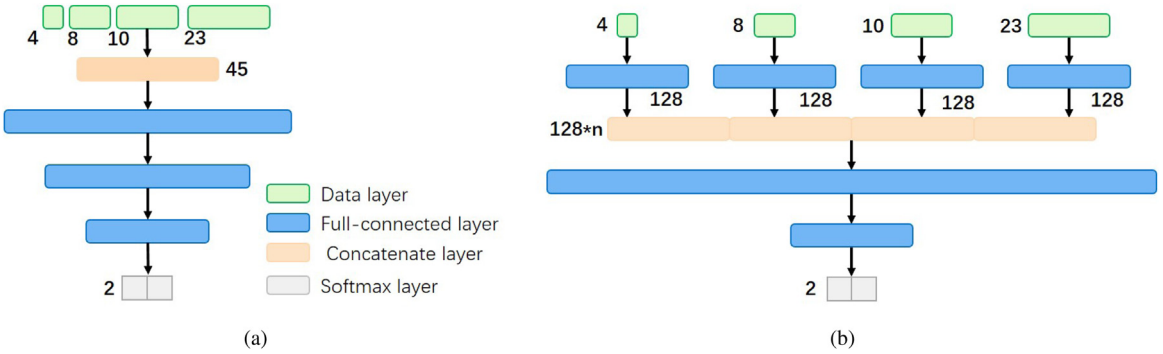

Fig. 3. (a) First concatenated the raw handcraft features outside the network and then input the concatenated one. (b) First input the raw handcraft features into layers of same dimensionality and then concatenate them inside the network. The CraftNet adopted the later pipline.

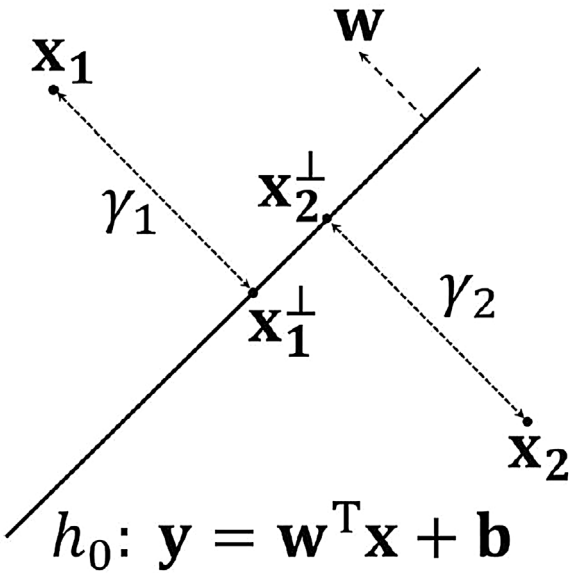

Fig. 4. Hyper plane of SVM. Here $\mathbf{w}$ denotes the weight and $\mathbf{b}$ denotes the bias. $\mathbf{x_i}$ is a sample and $\mathbf{x_i^\perp}$ is a point on Hyper plane h_0. γ_i means distance between $\mathbf{x_i^\perp}$ and $\mathbf{x_i}$.

characterized by the distance between the two samples belonging to different categories that are closest to it.

Fig. 4 illustrates a hyperplane $h_0 : y_0 = \mathbf{w}^T\mathbf{x} + b$ and the two samples $\mathbf{x_1}$ and $\mathbf{x_2}$ that are closest to it. Let the distance of $\mathbf{x_1}$ and $\mathbf{x_2}$ to h_0 be γ_1 and γ_2, then the separating ability of h_0 is just defined as $\gamma = \gamma_1 + \gamma_2$. Note that $\mathbf{x_i}$, $i = 1, 2$, can simply be written as

$$\mathbf{x_i} = \mathbf{x_i^\perp} + (-1)^{i-1}\gamma_i \frac{\mathbf{w}}{||\mathbf{w}||}, \tag{1}$$

Where $x_i^\perp$ is the perpendicular intersection of $\mathbf{x_i}$ to h_0. Without loss of generality, we can assume $y_0 = 0$, i.e., the hyperplane is represented as $\mathbf{w^T x + b = 0}$, then multiplying $\mathbf{w^T}$ and adding $\mathbf{b}$ on both sides of (1) gives

$$w^T x_i + b = w^T x_i^\perp + b + \gamma_i \cdot \frac{w \cdot w^T}{||w||} = w^T x_i^\perp + b + \gamma_i \cdot \frac{||w||^2}{||w||}$$

$$= (-1)^{i-1}\gamma_i \cdot ||w||. \tag{2}$$

Thus,

$$\gamma_i = (-1)^{i-1}\frac{w^T x_i + b}{||w||} \tag{3}$$

So the separating ability is calculated as,

$$\gamma = \gamma_1 + \gamma_2 = \frac{w^T \mathbf{x_1} + b - (w^T \mathbf{x_2} + b)}{||w||} \tag{4}$$

For the two-category classification problem, the hyperplane parameters $\mathbf{w^T}$ and $\mathbf{b}$ can be learned by the training process such that the decision rule is set as follows. If $\mathbf{w^T x + b} \geq 1$ then $\mathbf{x}$ will be classified to C_1 and if $\mathbf{w^T x + b} \leq -1$ then it will be classified to C_2. This further indicates that in (4), $\gamma = \frac{2}{||\mathbf{w}||}$. Then, a straightforward way to improve the robustness of the hyperplane (i.e., increasing γ) is to decrease the norm of $\mathbf{w}$.

In observance of this, we proposed a mixed loss function, P-S loss, that can help better train the neural networks. Specifically, it is composed of two items, the commonly used cross entropy and the weight decreasing item. Formally,

$$L_{P-S} = L_{WCE} + \alpha \cdot L_2 \tag{5}$$

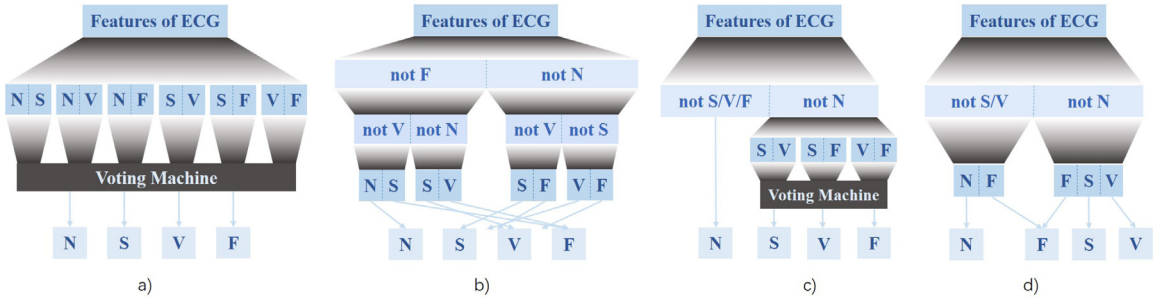

Fig. 5. Four different ensembles of child classifiers, vote, DDAG, one versus three (OVT), and one versus three with filters (OVTF) from left to right.

where the 2-norm $L_2(w) = ||w||_2 = \left(\sum_i^{N_f} w_i^2\right)^{1/2}$ with $N_f = 3072$ being the dimensionality of the feature. Its gradient can be easily calculated as $\nabla_w L(w) = 2\frac{w}{||w||}$. Since the gradient is proportional to the weight itself, one can imagine that the larger the weight magnitude, the faster the weight will decrease. Also, $L_{WCE} = \frac{N_{all}}{N_c} \cdot L_{CE}$ where N_{all} denotes the total number of samples and N_c denotes the number of c category samples.

The use of the L_2 norm in (5) allows for the weight to decrease faster, and thus improves the classification performance. Note that, it can also be explained as a regularizer term, preventing the weight explosion.

3.2.3. Decision directed acyclic graph model

With the proposed P-S loss, one can simply employ deep neural networks to conduct a four-category classification structure. But our experiments and reported results [28] both indicate that this classification scheme suffers from data imbalance. For MIT-BIH database, the N category is overwhelmingly more than S, V, and F, therefore, the average accuracy[1] is relatively low. To further reduce the effect of data imbalance, an alternative scheme is to cast the four-category problem to multiple two-category classifications. Four options are illustrated in Fig. 5.

Voting structure. Fig. 5(a) shows a voting structure that contains 6 two-category classifiers, and every classifier outputs probabilities for the two categories. Take the N beat for an example, let C_{NS}, C_{NV}, and C_{NF} denote the three classifiers dealing with the N beat, and let $\mathcal{P}_{NS}(x = N)$, $\mathcal{P}_{NV}(x = N)$, and $\mathcal{P}_{NF}^{N}(x = N)$ denote the probability of C_{NS}, C_{NV}, and C_{NF} determining it to be the N beat. The first scheme is to directly define

$$P(\mathbf{x} = N) = \mathcal{P}_{NS}(\mathbf{x} = N) \cdot \mathcal{P}_{NV}(\mathbf{x} = N) \cdot \mathcal{P}_{NF}(\mathbf{x} = N) \tag{6}$$

Similarly, $P(\mathbf{x} = S)$, $P(\mathbf{x} = V)$, and $P(\mathbf{x} = F)$ can be computed with the three classifiers dealing with them, and $\mathbf{x}$ will be classified to be the category having the largest probability. The other scheme is essentially a voting scheme, i.e., counting the number of classifiers that tend to determine $\mathbf{x}$ to a certain category, e.g., the N beat,

$$P(x = N) = |\mathcal{P}_{\mathbf{NS}}(x = N) > 1/2| + |\mathcal{P}_{NV}(x = N) > 1/2| + |\mathcal{P}_{NF}(x = N) > 1/2| \tag{7}$$

And likewise, the $P(\mathbf{x} = S)$, $P(\mathbf{x} = V)$, and $P(\mathbf{x} = F)$ can be computed for classifying $\mathbf{x}$.

Decision directed acyclic graph. Fig. 5(b) illustrates a second DDAG structure. For an input sample $\mathbf{x}$, the first step is to feed it to the N-F classifier. If it is classified as the N beat, it is considered to belong to a wider category NSV beat (i.e., belonging to any beat of N, S, and V), otherwise it is considered to belong to the FSV beat. The second step is to feed $\mathbf{x}$ to classifiers according to the classification result of the first step. Specifically, If $\mathbf{x}$ is classified as NSF beat at the first step, it will be fed to the N-V classifier, to determine it to belong to N beat (NS) or V beat (VS).

Fusion structures. Fig. 5(c) and (d) shows the other two classification structures. Both structures classify an input sample to normal or abnormal category by i.e., the initial classifier, followed by subsequent finer classification of the abnormal category. In Fig. 5(d), the finer classification of the normal category differs from Fig. 5(c) in that the F category will be further screened out by a N-F classifier.

This work employs the DDAG [36,31] that utilizes the six two-category models. Each model is trained on two categories from N, S, V, and F without grouping them. There were four categories, therefore $\binom{4}{2} = 6$ models were needed. The advantage of DDAG is that the other two categories will not influence the classifiers that are not trained on them. For example, V does not have an influence on the classifier for N and S, while the influence will dampen the performance of the Vote structure as shown in Fig. 5(a).

For a particular model, varying combinations of the features were exploited to train it, and the best-performing trained model was kept. By this mean, the best 6 models trained on different feature combinations were obtained by utilizing the P-S loss. The best version first inputs the sample into the N-F classifier. Then the part divided to N is input into the N-V classifier and the part divided to F is input into the S-V classifier. Finally, the part divided to N is input to the N-S classifier, the part divided to V is input to the V-S classifier, the part divided to S is input to the F-V classifier, and the part divided to V is input to the S-F classifier. The results of the last step are the final output.

4. Results

The proposed CraftNet is tested on the public Massachusetts Institute of Technology-BethIsrael Hospital (MIT-BIH) arrhythmia database [20] and compared with the state-of-the-art methods. The database MIT-BIH has been widely used for investigating the recognition performance of different methods, e.g., [19,28,35].

4.1. Comparison with state-of-the-art

This section presents the experimental results of DDAG structure and compares it with the state-of-the-art methods which provided the confusion matrix. For CraftNet, we adopted the 5-fold cross validation method to yield the standard values and grouped the best 2-categories classifier in each fold to produce the final CraftNet model. The recognition result is listed in Table 2.

The specific standard values of CraftNet in order are: 99.22 ± 0.03, 84.90 ± 3.26, and 94.72 ± 0.38 for category N; 40.00 ± 1.64, 85.27 ± 2.14, and 94.43 ± 0.42 for category S; 94.87 ± 1.74, 94.08 ± 0.45, and 99.65 ± 0.13 for category V; 8.45 ± 2.44, 89.43 ± 1.29, and 91.60 ± 2.68 for category F. The proposed method yields the highest average sensitivity 89.25% and specificity 95.79%. And it yields the second highest average posi-

[1] Here accuracy means recall.

Table 2
The recognition result of different methods. 'Pos' denotes positive productivity, 'Sen' denotes sensitivity, 'Spe' denotes Specificity, 'Avg' denotes the average value, and 'Std' denotes the stand value. The last column gives the proportion of right recognized samples over all samples. The best result is blackened.

	N (%)			S (%)			V (%)			F (%)			Avg (%)			Acc (%)
	Pos	Sen	Spe	Pos	Sen	Spe	Pos	Sen	Spe	Pos	Sen	Spe	Pos	Sen	Spe	Num
Chazal (2004) [3]	99.17	87.06	93.95	38.53	75.98	93.33	81.67	80.31	98.79	8.57	89.43	92.46	56.98	83.19	95.13	86.24
Mar (2011) [17]	99.12	89.94	**95.53**	33.53	83.22	93.67	75.89	86.75	98.09	16.57	61.08	97.57	56.28	80.18	95.71	88.99
Li (2016) [12]	**99.76**	94.67	–	0.16	20.00	–	89.78	94.20	–	0.52	50.00	–	47.56	64.72	–	–
Zhang (2014) [35]	98.98	88.94	92.84	35.98	79.06	93.95	92.75	85.48	99.54	13.73	**93.81**	95.36	60.36	86.82	95.42	88.34
Mondéjar-Guerra (2018) [19]	98.20	**95.94**	86.34	**49.75**	78.10	**96.61**	93.79	**94.75**	99.57	23.65	12.37	**99.69**	**66.35**	70.29	95.55	**94.47**
Sellami (2018) [28]	98.81	88.52	91.31	30.44	82.04	92.80	72.22	92.05	97.55	**26.58**	68.30	98.51	57.01	82.72	95.04	88.35
CraftNet (standard value)	99.22	84.90	94.72	40.00	**85.27**	94.43	**94.87**	94.08	**99.65**	8.45	89.43	91.60	60.64	**88.31**	95.10	88.69
CraftNet	99.18	88.16	94.34	41.64	**85.37**	94.85	**95.63**	94.53	**99.70**	10.89	88.92	94.28	61.84	**89.25**	**95.79**	89.24

Table 3
The combinations of four features and the performance of two-category classifiers taking them as input.

RR	Wavelet	HOS	Morph	N/S (%)	N/V (%)	N/F (%)	S/V (%)	S/F (%)	V/F (%)
•	•			92.43/81.90	94.43/94.16	85.89/75.71	97.17/95.59	97.80/93.29	91.39/31.95
•		•		**94.27/88.09**	98.16/96.67	92.07/87.62	98.58/87.14	98.58/95.10	94.09/92.01
•			•	92.37/87.90	94.99/95.12	73.83/83.50	85.41/87.14	97.17/96.39	95.83/96.64
	•	•		91.58/12.68	92.73/94.22	84.58/61.60	91.02/97.17	93.41/81.19	94.13/88.14
	•		•	83.27/76.15	92.63/94.07	58.22/59.28	90.05/96.71	90.39/90.72	89.66/85.57
		•	•	76.33/79.51	98.47/93.14	89.61/70.88	95.61/96.49	97.41/92.27	93.85/82.73
•	•	•		86.01/85.02	96.97/95.86	79.49/78.09	96.73/96.36	98.04/94.07	92.26/75.77
•	•		•	86.13/80.53	95.07/96.21	87.00/80.15	97.80/95.65	97.85/96.39	95.83/90.20
•		•	•	92.95/87.70	**98.20/98.01**	**94.05/88.40**	98.68/86.61	**98.39/94.07**	**97.24/96.39**
	•	•	•	60.90/41.12	93.19/94.19	79.53/61.85	91.90/96.14	92.87/91.23	87.88/86.59
•	•	•	•	90.04/81.80	96.85/95.84	92.30/67.01	**97.70/97.04**	96.18/90.46	95.71/93.30

tive product, 61.84%, which is only lower than 66.35% achieved by Mondéjar-Guerraa et al. [19]. However, [19] only provides 70.29% average sensitivity since it mis-classified most of the F beats [19]. Of all the 16 metrics in Table 2, CraftNet yields 5 highest scores and 7 second highest scores, comparable to [19] that yields 7 highest and 3 second highest scores, but outperforms other methods. Mondéjar-Guerraa et al. [19] mis-classified most of the F beats and hence increased the sensitivity of N beats (95.94%), which is much higher than other methods. This can also explain why [19] has a much higher total accuracy (94.47%) as N beats are mostly classified. Also, the proposed method gives a higher sensitivity (85.37%) than [19] as S beats are more mis-classified as N beats [19], showing the robustness of CraftNet identifying illnesses since S and V beats occupy about 93% abnormal beats.

Benefiting from the DDAG structure and the proposed P-S loss, the CraftNet yields the best average recognition performance on all four beats. This is especially meaningful for F beats which have much fewer samples than other beats, indicating the CraftNet has a stronger ability of dealing with severe data imbalance that often happens on medical data.

4.2. Comparison for different combinations of features

This section investigates the performance of the combinations of features when fed to two-category classifiers. Inspired by Mondéjar-Guerraa et al. [19], we consider various combinations of RR, Wavelet, HOS, and Morph. The classification results are listed in Table 3.

For each two-category model, the combination of the features providing the best classification performance is used as the reference one. Table 3 gives the recognition performance for varying combinations. From Table 3 we can see that some features are only useful for certain classifiers. For example, for the N-S classifier (the classifier for N and S), RR plays an indispensable role and HOS also acts as a supporting feature; Morph and Wavelet may be counterproductive, e.g., when RR, HOS, and Wavelet are used together, the accuracy drops from 91.18% to 85.52%. For the N-V classifier, the combination of RR, Morph, and HOS perform best. If Wavelet is combined, the accuracy drops by 1.76%, saying that it may not be meaningful for improving the N-V classifier. Similar phenomenons can be observed on the N-F, S-F, and V-F classifier. When Wavelet feature is used, the accuracy drops by 11.57%, 2.91%, and 2.31%, respectively. But for the S-V classifier, the combination of all the four features outperforms other combinations.

In this work, the best combination is chosen for each classifier, and the trained classifiers are to be used in the ensemble.

4.3. Comparison for different structures

This section presents the classification result by four ensemble structures shown in Fig. 5.

The experimental results are listed in Table 4. The performance of the two voting schemes given in Eqs. (6) and (7) is comparable to each other. There are 16 metrics in Table 4, 'Pos', 'Sen', and 'Spe' for each beat; their average and standard values over all types of beats; and the total average accuracy. The DDAG structure has the best comprehensive classification ability: yielding 7 highest scores on 16 metrics. It achieves the best average sensitivity and specificity 89.43% and 95.84%. The OVT structure yields a recognition result close to [19], which recognizes the N category very well but at the cost of low detection rate of F beat (only 28.87% sensitivity). The OVTF structure partly attacks the problem with OVT, i.e., "salvage" the F beat but at the cost of low detection rate of N beat.

Among these structures, DDAG owns a more reasonable structure. Also, it has many types depending on the sequence of child classifiers. This character makes it easy to find a more reasonable result from all types. Table 5 lists the top six results of different type DDAG. The order of models in Table 5 follows the sequence from left to right, top to bottom in Fig. 6. The results shows that different sequences of child classifiers in DDAG have some slight influences. This happens because the top child classifier directly decides the accuracy of early classified two categories. The accuracy of these categories decreases in the form of a multiplier as the number of corresponding classifiers.

Table 4
The results of different ensembles on four categories.

	N (%)			S (%)			V (%)			F (%)			Avg (%)			Acc (%)
	Pos	Sen	Spe	Pos	Sen	Spe	Pos	Sen	Spe	Pos	Sen	Spe	Pos	Sen	Spe	Num
Vote(mul)	99.13	88.61	93.92	42.13	86.24	94.90	93.51	**94.35**	99.55	11.49	85.31	94.83	61.56	88.63	95.80	88.86
Vote(add)	99.13	88.60	93.94	42.18	86.34	94.91	93.50	94.32	99.55	11.46	85.31	94.81	61.57	88.64	95.80	88.86
DDAG	**99.20**	88.22	**94.43**	41.16	**87.66**	94.61	**94.28**	94.22	**99.60**	11.54	87.63	94.72	61.55	**89.43**	**95.84**	88.58
OVT	98.60	**93.20**	89.68	**43.66**	85.22	**95.27**	90.04	94.04	99.28	**15.93**	28.87	**98.80**	**62.06**	75.33	95.76	**92.43**
OVTF	99.10	87.64	93.83	**43.66**	85.22	**95.27**	90.04	94.04	99.28	10.12	**88.40**	93.82	60.73	88.82	95.55	87.96

Table 5
The results of different orders of child classifiers in DDAG.

	N (%)			S (%)			V (%)			F (%)			Avg (%)			Acc (%)
	Pos	Sen	Spe	Pos	Sen	Spe	Pos	Sen	Spe	Pos	Sen	Spe	Pos	Sen	Spe	Num
1	99.20	88.22	94.43	41.16	87.66	94.61	94.28	94.22	99.60	11.54	87.63	94.72	61.55	89.43	95.84	88.58
2	99.20	88.22	94.43	41.11	87.66	94.60	94.91	93.88	99.65	11.47	87.89	94.66	61.67	89.41	95.84	88.56
3	99.20	88.22	94.43	40.91	87.66	94.55	95.42	93.73	99.69	11.47	87.89	94.66	61.75	89.37	95.83	88.55
4	99.20	88.22	94.43	41.11	87.66	94.60	95.42	93.73	99.69	11.39	87.89	94.62	61.78	89.37	95.83	88.55
5	99.16	88.32	94.19	41.17	87.66	94.61	95.42	93.73	99.69	11.56	87.63	94.72	61.83	89.33	95.80	88.64
6	99.20	88.22	94.43	41.05	87.66	94.58	94.28	94.22	99.60	11.49	86.86	94.73	61.50	89.24	95.84	88.58

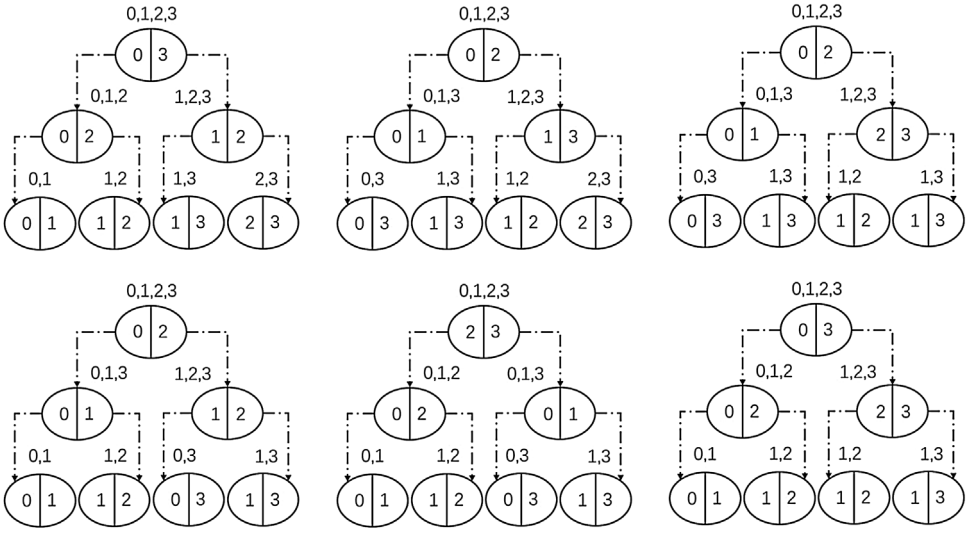

Fig. 6. Different sequence of child classifiers in DDAGS. 0, 1, 2, and 3 denote N, S, V, and F category respectively. X/X denotes the child classifier.

5. Discussion

The advantage of the proposed method is that it employs an architecture of neural network to classify the ECG signal by utilizing the handcraft features. The devised architecture of the neural network is able to make best of the handcraft features for robust classification, and the proposed P-S loss further enhances its generalization ability. The use of handcraft features can effectively cope with data imbalance, which is a commonly encountered problem with the deep neural networks when trained on MIT-BIH. In addition, like Focal loss [14], an inverse weight was assigned to the type of beat that has abundant samples, e.g., N beats, which further alleviates the influence of data imbalance. The ensembles of varying structures were explored to find the structure that provides the best classification performance. Experimental results showed that the proposed method had the comprehensive and precise classification ability, yielding a high score on all sensitivity metrics.

Some methods designed only for N, S, and V are shown in Table 6. [13,6,21] all yield accuracy higher than 90%. However, the difficulty of three-category classification task is different from that of four-

Table 6
The recognition result of three-category methods. 'Pos' denotes positive productivity, 'Sen' denotes sensitivity, and 'Spe' denotes Specificity. The last column gives the proportion of right recognized samples over all samples.

	N(%)			S(%)			V(%)			Acc(%)
	Pos	Sen	Spe	Pos	Sen	Spe	Pos	Sen	Spe	Num
Song (2005) [29]	78.0	83.9	–	27.0	48.3	–	80.8	38.7	–	76.3
Yu and Chou (2008) [34]	78.3	79.2	–	1.8	5.9	–	83.9	66.4	–	75.2
Ye (2010) [32]	80.2	78.2	–	3.2	10.3	–	50.2	48.5	–	75.2
Lin and Yang (2014) [13]	91.0	99.0	–	81.0	31.0	–	86.0	73.0	–	93.0
Garcia (2017) [6]	94.0	98.0	82.6	62.0	53.0	97.9	87.3	59.4	95.9	92.4
Mousavi (2019) [21]	99.7	99.6	96.1	88.9	92.6	99.7	99.9	99.5	99.9	99.5

category classification task. Especially when class F is so rare. As mentioned in contributions, the proposed method contributes to alleviate the data imbalance of CVDs database. Unlike other methods that provide a higher recognition performance at the cost of wrongly classifying F beats, the proposed method provides a pretty high recognition accuracy for every type of beat.

The drawback of the proposed method lies in that it is not an end-to-end framework because the handcraft features can only be extracted beforehand. The training process also needs much attention since the mixed loss function always pushes the weight parameters out of the local extrema and hence makes the converging more difficult.

6. Conclusion

This paper proposed an approach to classify ECG signals. Inspired by the principle of SVM, an architecture of the neural network, CraftNet, was designed to recognize ECG signals by utilizing the handcraft features. This architecture effectively coped with data imbalance problem which is often encountered by the end-to-end deep neural networks. The designed loss function for the architecture further enhances the classification performance and improves the generalization ability. The results showed that the proposed approach had a stronger classification ability than the state-of-the-art methods and was less influenced by data imbalance that exists in MIT-BIH database.

Authors' contribution

Y. Li conceived the experiment(s), Z. He, H. Wang, B. Li, N. Li and Y. Gao conducted the experiment(s), Z. He and Y. Li analysed the results. All authors reviewed the manuscript.

Acknowledgements

The authors thank the editor and anonymous reviewers for their helpful comments and valuable suggestions. This work was supported by the National Natural Science Foundation of China (Grant No. 61620106002), the Beijing Municipal Natural Science Foundation (No. 4172024), Beijing University of Posts and Telecommunications (Nos. 2013XD-04, 2015XD-02), the Beijing Key Laboratory of Work Safety and Intelligent Monitoring Foundation.

The study does not involve human or animal subjects.

Declaration of Competing Interest

The authors report no declarations of interest.

References

[1] U.R. Acharya, S.L. Oh, Y. Hagiwara, J.H. Tan, M. Adam, A. Gertych, A deep convolutional neural network model to classify heartbeats, Comput. Biol. Med. 89 (2017) 389–396, http://dx.doi.org/10.1016/j.compbiomed.2017.08.022.
[2] A. Al-Fahoum, I. Howitt, Combined wavelet transformation and radial basisneural networks for classifying life-threatening cardiac arrhythmias, Med. Biol. Eng. Comput. 37 (5) (1999) 566–573, http://dx.doi.org/10.1007/BF02513350.
[3] P. de Chazal, M. O'Dwyer, R. Reilly, Automatic classification of heartbeats using ECG morphology andheartbeat interval features, IEEE Trans. Biomed. Eng. 51 (7) (2004) 1196–1206, http://dx.doi.org/10.1109/TBME.2004.827359.
[4] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 237–297, http://dx.doi.org/10.1023/A:1022627411411.
[5] A. Esteva, B. Kuprel, R.A. Novoa, J.M. Ko, S.M. Swetter, H.M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (2017) 115–118, http://dx.doi.org/10.1038/nature21056.
[6] G. Garcia, G. Moreira, D. Menotti, E. Luz, Inter-patient ECG heartbeat classification with temporal VCG optimized by PSO, Sci. Rep. 7 (1) (2017), 10543, http://dx.doi.org/10.1038/s41598-017-09837-3.
[7] A.L. Goldberger, L.A.N. Amaral, L. Glass, J.M. Hausdorff, P.C. Ivanov, R.G. Mark, Physiobank, physiotoolkit, and physionet: components of a newresearch resource for complex physiologic signals, Circulation 101 (23) (2000) e215–e220, http://dx.doi.org/10.1161/01.CIR.101.23.e215.
[8] H. Huang, J. Liu, Q. Zhu, R. Wang, G. Hu, A new hierarchicalmethod for inter-patient heartbeat classification usingrandom projections and RR intervals, Biomed. Eng. Online 13 (2014) 1–26, http://dx.doi.org/10.1186/1475-925X-13-90.
[9] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, International Conference on Neural Information Processing Systems, vol. 25(2) (2012), http://dx.doi.org/10.1145/3065386.
[10] G. de Lannoy, D. Francois, J. Delbeke, M. Verleysen, Weighted svms and feature relevance assessment insupervised heart beat classification, Biomed. Eng. Syst. Technol. 40 (9) (2010) 212–223, http://dx.doi.org/10.1007/978.3.642.18472.7.17.
[11] G. de Lannoy, D. Francois, J. Delbeke, M. Verleysen, Weighted conditional random fields for supervisedinterpatient heartbeat classification, IEEE Trans. Biomed. Eng. 59 (1) (2012) 241–247, http://dx.doi.org/10.1109/TBME.2011.2171037.
[12] T. Li, M. Zhou, Ecg classification usingwavelet packet entropy and random forests, Entropy 18 (8) (2016), http://dx.doi.org/10.1109/IEMBS.2011.6091236.
[13] C.C. Lin, C.M. Yang, Heartbeat classification usingnormalized rr intervals and morphological features, Math. Problem Eng. (2014) 1–11, http://dx.doi.org/10.1109/IS3C.2014.175.
[14] Y.T. Lin, P. Goyal, R. Girshick, K. He, P. Dollar, Focal loss for dense object detection., IEEE Trans. Pattern Anal. Mach. Intell. PP (2017) 2999–3007.
[15] M. Llamedo, J. Martinez, Heartbeat classification usingfeature selection driven by database generalization criteria, IEEE Trans. Biomed. Eng. 58 (3) (2011) 616–625, http://dx.doi.org/10.1109/TBME.2010.2068048.
[16] E. Luz, T. Nunes, V.D. Albuquerque, J. Papa, D. Menotti, ECG arrhythmia classification based on optimum-path forest, Expert Syst. Appl. 40 (9) (2012) 3561–3573, http://dx.doi.org/10.1016/j.eswa.2012.12.063.
[17] T. Mar, S. Zaunseder, J. Martnez, M. Llamedo, R. Poll, Optimization of ECG classification by means of feature selection, IEEE Trans. Biomed. Eng. 58 (8) (2011) 2168–2177, http://dx.doi.org/10.1109/TBME.2011.2113395.
[18] S. Meek, F. Morris, Introduction. I – leads, rate, rhythm, and cardiac axis, BMJ 324 (7334) (2002) 415–418, http://dx.doi.org/10.1136/bmj.324.7334.415.
[19] V. Mondéjar-Guerraa, J. Novo, J. Rouco, M. Penedo, M. Ortega, Heartbeat classification fusing temporal and morphologicalinformation of ECGs via ensemble of classifiers, Biomed. Signal Process. Control 47 (2019) 41–48, http://dx.doi.org/10.1016/j.bspc.2018.08.007.
[20] G. Moody, R. Mark, The impact of the MIT-BIH arrhythmia database, IEEE Eng. Med. Biol. Mag. 20 (3) (2001) 45–50, http://dx.doi.org/10.1109/51.932724.
[21] S. Mousavi, F. Afghah, Inter- and intra- patient ecg heartbeat classification for arrhythmia detection: a sequence to sequence deep learning approach., International Conference on Acoustics, Speech, and Signal Processing (2019) 1308–1312.
[22] Organization, W.H., The Health Topic of Cardiovascular Diseases. https://www.who.int/health-topics/cardiovascular-diseases/ (Accessed 4 April 2020).
[23] S. Osowski, T. Linh, ECG beat recognition using fuzzy hybrid neuralnetwork, IEEE Trans. Biomed. Eng. 48 (11) (2001) 1265–1271, http://dx.doi.org/10.1109/10.959322.
[24] K. Park, B. Cho, D. Lee, S. Song, J. Lee, Y. Chee, I.Y. Kim, S. Kim, Hierarchical support vector machine based heartbeat classification using higher order statistics and hermite basis function, Comput. Cardiol. (2008) 229–232, http://dx.doi.org/10.1109/CIC.2008.4749019.
[25] J.C. Platt, N. Cristianini, J. Shawe-Taylor, Large margin dags for multiclass classification, in: Advances in Neural Information Processing Systems, 2000, pp. 547–553.
[26] S. da, E. Luz, W. Schwartz, G. Cmara-Chvez, D. Menotti, ECG-based heartbeat classification for arrhythmia detection: a survey, Comput. Methods Prog. Biomed. 127 (Suppl. C) (2016) 144–164, http://dx.doi.org/10.1016/j.cmpb.2015.12.008.
[27] C. Saunders, M.O. Stitson, J. Weston, R. Holloway, L. Bottou, B. Scholkopf, A. Smola, Support vector machine, Comput. Sci. 1 (4) (2002) 1–28, http://dx.doi.org/10.1007/978-3-642-27733-7-299-3.
[28] A. Sellami, H. Hwang, A robust deep convolutional neural network with batch-weighted loss for heartbeat classification, Expert Syst. Appl. 122 (2019) 75–84, http://dx.doi.org/10.1016/j.eswa.2018.12.037.
[29] H.M. Song, J. Lee, P.S. Cho, J. Lee, K.S. Yoo, Support vector machine based arrhythmia classification using reduced features, Int. J. Control Autom. Syst. 3 (4) (2005) 571–579, http://dx.doi.org/10.1007/s00170-004-2187-3.
[30] M. Soria, J. Martinez, Analysis of multidomain features for ECG classification, Comput. Cardiol. (2009) 561–564.
[31] H. Xu, X. Bie, F. Hui, T. Ye, Multiclass SVM active learning algorithm based on decision directed acyclic graph and one versus one, Cluster Comput. (2018) 1–11, http://dx.doi.org/10.1007/s10586-018-1951-3.
[32] C. Ye, T.M. Coimbra, K.V.B. Kumar, Arrhythmia detection and classification using morphological and dynamic features of ECG signals, Conference proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference (2010) 1918–1921, http://dx.doi.org/10.1109/IEMBS.2010.5627645.

[33] C. Ye, B. Kumar, M. Coimbra, Combining generalmulti-class and specific two-class classifiers for improvedcustomized ECG heartbeat classification., International Conference on Pattern Recognition (ICPR) (2012) 2428–2431.
[34] 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, http://dx.doi.org/10.1016/j.eswa.2007.05.006.
[35] Z. Zhang, J. Dong, X. Luo, K. Choi, X. Wu, Heartbeat classification using disease-specific feature selection, Comput. Biol. Med. 46 (Suppl. C) (2014) 79–89, http://dx.doi.org/10.1016/j.compbiomed.2013.11.019.
[36] L. Zhou, Q. Wang, H. Fujita, One versus one multi-class classification fusion using optimizing decision directed acyclic graph for predicting listing status of companies, Inform. Fusion 36 (2016) 80–89, http://dx.doi.org/10.1016/j.inffus.2016.11.009.