A Dual-scale Network with Spatial-temporal Attention for 12-lead ECG Classification

Shuo Xiao[1, &], Yiting Xu[1,&], Chaogang Tang[1,*] and Zhenzhen Huang[1]
[1]School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, China
[e-mail: sxiao@cumt.edu.cn, ts21170104p31@cumt.edu.cn, cg.tang@foxmail.com,
huangzhenzhen@cumt.edu.cn]
[&]These authors contributed equally to this work
[*]Corresponding author: Chaogang Tang

Received March 5, 2023; accepted August 28, 2023; published September 30, 2023

Abstract

The electrocardiogram (ECG) signal is commonly used to screen and diagnose cardiovascular diseases. In recent years, deep neural networks have been regarded as an effective way for automatic ECG disease diagnosis. The convolutional neural network is widely used for ECG signal extraction because it can obtain different levels of information. However, most previous studies adopt single scale convolution filters to extract ECG signal features, ignoring the complementarity between ECG signal features of different scales. In the paper, we propose a dual-scale network with convolution filters of different sizes for 12-lead ECG classification. Our model can extract and fuse ECG signal features of different scales. In addition, different spatial and time periods of the feature map obtained from the 12-lead ECG may have different contributions to ECG classification. Therefore, we add a spatial-temporal attention to each scale sub-network to emphasize the representative local spatial and temporal features. Our approach is evaluated on PTB-XL dataset and achieves 0.9307, 0.8152, and 89.11 on macro-averaged ROC-AUC score, a maximum F1 score, and mean accuracy, respectively. The experiment results have proven that our approach outperforms the baselines.

Keywords: 12-lead ECG, deep neural networks, signal features, dual-scale network, spatial-temporal attention.

This work was supported by the National Natural Science Foundation of China (Nos. 62271486, 62071470).

http://doi.org/10.3837/tiis.2023.09.004 **ISSN : 1976-7277**

1. Introduction

Cardiovascular disease (CVD) seriously affects people's health and is the leading cause of death worldwide. According to statistics, cardiovascular deaths account for more than 30% of global deaths [1]. Electrocardiogram (ECG) is a one-dimensional medical signal obtained from the surface of human body, which reflect heart's electrical activities [2]. It is commonly employed for detecting and diagnosing various cardiac conditions, such as heart attack, myocardial ischemia, and arrhythmia. However, the visual inspection of ECG signal by a clinician or cardiologist is difficult and time-consuming. Moreover, it is prone to human error. Therefore, how to realize automatic classification of ECG signals to assist human diagnosis has become an important work. In recent years, intelligent medical care has become increasingly prominent. ECG based automatic detection of cardiovascular diseases can assist doctors in clinical operations.

The single-lead ECG signal is used for basic cardiac monitoring. Many studies have used single-lead ECG signal to classify ECG into different heartbeats and diseases. However, the single-lead ECG signal has the shortcomings of insufficient information and the diagnosis of many diseases requires information from different leads. Therefore, 12-lead ECG recordings are used to makes a diagnosis, which is the standard data for the cardiologist to diagnose the diseases. It provides a comprehensive view and detailed information of heart activities from different spatial angles, which overcomes the shortcomings of limited single-lead ECG signal. However, there are few studies on 12-lead ECG compared with single-lead ECG.

Deep learning has made great progress in the application of medical assistance and healthcare, such as drug development, medical image diagnosis and genomic analysis [3]. Meanwhile, it is employed to solve ECG classification problem. Deep learning methods can automatically extract features of ECG signal and realize ECG classification. The convolutional neural network (CNN) is the most commonly method to achieve ECG classification, which can obtain features of different levels. Many methods based on single scale CNN network with single scale convolution filters are proposed. However, each lead in the 12-lead ECG recording is extremely long sequence, which consists of many heartbeats. The single-scale network with single scale convolution filters is difficult to obtain the information of the long ECG sequence. Since convolutional filters of different sizes learn different scope of reception fields [4]. Compared with single scale convolutional filters, combining these features extracted by different size convolutional filters will provide better feature representation. Specifically, small scale convolution filters have small receptive field, which are appropriate for capturing local statistical information and amplitude information in the ECG signal [3]. Large scale convolutional filters have large receptive field, which is beneficial to extract many morphological features and interval information. In addition, the contribution of various spatial and time periods in the feature map from 12-lead ECG may differ in ECG classification. Many existing studies are not able to focus on important channels and time periods of the feature map simultaneously.

To solve the above problems, we propose a dual-scale network with spatial-temporal attention (STADSNet) for 12-lead ECG classification. The main contributions are summarized as below:

1. Firstly, we propose a dual-scale network with convolution filters of different sizes for 12-lead ECG signal feature extraction. The network includes two sub-networks to extract ECG features of different scales.
2. We add a spatial-temporal attention to each sub-network. The spatial-temporal attention can assign different weights to different channel information and time information, which

is conducive to obtaining representative local spatial-temporal features of ECG signal simultaneously.

3. Finally, we evaluate our model on PTB-XL dataset and achieve 0.9307, 0.8152, 89.11 on macro-averaged ROC-AUC, F1 score (Max), and Mean Accuracy, respectively.

The rest of this paper is structured as follows: The related works of ECG classification is described in Section 2. Section 3 presents our proposed method of ECG classification. Section 4 presents our experiment to verify our method. Finally, Section 5 summarizes this paper.

2. Related Works

Advances in artificial intelligence technology have enabled automatic ECG classification. In recent years, many methods of ECG signal classification have been proposed, which involves three main steps: ECG data preprocessing, extracting feature and ECG classification. Since the raw ECG signal has noises and varies in length. Therefore, in the preprocessing stage, ECG signal need to be denoised, removed baseline drift and processed as the same length. Feature extraction is the most important step. It is usually achieved by traditional ECG classification methods and ECG classification methods based on deep learning.

Traditional ECG classification methods extract features manually by cardiologists or using traditional feature extraction algorithms. Features include RR intervals, mentality features, time-frequency etc. The statistical methods or some feature selection algorithms are used to select the most representative features for training the classifier, such as support vector machine (SVM), decision tree (DT) and naive bayes etc. The classifiers classify ECG signal into different types of heartbeats and diseases. Ye et al. [5] extract morphological features and dynamic features of the two leads separately. Then the features of the two leads are fused. Finally, SVM is used to classify 16 heartbeats categories and the average accuracy reached 99.3%. Nasiri et al. [6] extract twenty-two features manually, then use SVM to achieve four types arrhythmias classification with 93% accuracy. Acharya et al. [7] first extract thirteen nonlinear features then use KNN and DT classifiers to classify five heartbeats types, with an average accuracy of 96.3%. In [8], nonlinear features and frequency domain features are extracted for Arrhythmia Classification. Finally, accuracy of 98.8% is obtained on MIT-BIH arrhythmia database. In general, traditional ECG classification methods are relatively faster in training and prediction. However, ECG signal contains many types of waveform and noise. The potential information in the original signal is ignored by manual feature extraction. In addition, the performance of traditional training classifiers is usually affected by data distribution. If the data distribution changes, the performance of the classifier can be significantly affected. The selection of model parameters becomes more difficult with the increase of feature dimension. Therefore, it is often difficult for traditional methods to deal with ECG signals effectively.

With deep learning technology making significant achievements in natural language processing and image classification. Many methods based on deep learning are proposed for ECG classification. Methods base on deep learning automatically extract features and realizes ECG classification. The CNN is an effective method, which can obtain features of different levels from ECG signals. In addition, ECG signal is 1-D signal. The 1-D CNN is used for feature extraction from the raw ECG signal. Hannun et al. [9] propose a model based on 34-layer CNN. The model is trained and tested using 91232 single-lead ECG recordings. Finally, F1 score achieved 0.837, which exceeded that of average cardiologists. Wang et al. [10] combine the attention module with CNN to effectively extract the features of different stages and realize 9 arrhythmias categories. It shows that features extracted by CNN at different stages are very important to classification results. Meanwhile, the attention module assigns a weight to the features extracted at different stages. In addition to [10], many other studies have also used attention modules, which improve the classification results. Many existing studies convert 1-D raw ECG signal into 1-D images. Then 2-D CNN is used to feature extraction. Wang et al. [11] use transform the ECG signal to the 2D-scalogram. Then 2-D CNN is used to extract time-frequency domain features, which combines RR interval features for heartbeats classification. The accuracy of 98.74% is achieved. Recurrent neural network (RNN) is an effective method for time series signal processing. Therefore, RNN is also utilized to solve ECG classification problem. In [12], authors propose LSTM-AE model to extract high-level features and then use SVM to classify five heartbeats. Finally, accuracy reaches 99.45%. Some studies combine CNN and LSTM for ECG signal classification. In [13], the researchers combine features obtained by CNN and bi-directional LSTM for heartbeats classification. The model inclues two network branches, which is CNN network and Bi-directional LSTM network. The accuracy is 99.42%. Chen et al. [14] uses a single beat and three beats from the single-lead ECG signal as two inputs. Then ECG signal information of different types was extracted by two network branches, which consists of SE-Residual blocks and bi-directional. Finally, the ECG signal is classified into four heartbeats categories by information fusion. The accuracy is 99.56%. Yao et al. [2] combine CNN and LSTM to fuse different types of features of ECG signal. To focus on important information, the attention module is added after LSTM. Finally, it completes the classification of 9 cardiac arrhythmias categories with accuracy of 81.2%. Recently, many studies have applied Transformer to ECG classification models have been used to ECG classification. Yan et al. [15] divide a long ECG signal into multiple heartbeats. Then use Transformer to get features of each heartbeats. Finally, the features of multiple heartbeats are fused to achieve the ECG classification.

The above models all adopt single scale network, ignoring the interaction of different scale features. The ECG signal is a time series with a long duration. Therefore, the extraction of features at different scales is crucial for ECG classification. In image classification field, multi-scale network is proposed and achieves good results. Therefore, some studies have applied multi-scale networks to the medical field. Zhang et al. [3] use multi-scale networks for ECG arrhythmia classification. The dilated convolution used in the network to achieve multi-scale feature extraction. Inspired by it, we propose a dual-scale network for ECG classification. At the same time, spatial-temporal attention is added to the dual-scale network. Since the network can pay attention to important spatial-temporal features.

3. Methods

3.1 Problem Definition

The input of the model is ECG recording, which is time-series signal. The original ECG recording is $x^{(i)} \in R^{l \times 12}, i \in [1, n]$, where l represents length of ECG signal, 12 represents 12-lead and n represents the quantity of ECG recordings. The corresponding reference label of $x^{(i)}$ is $y^{(i)} = \left(y_1^{(i)}, \ldots, y_c^{(i)}\right)$, where c represents the number of disease categories. The model needs to obtain representative features from ECG recording and output correct labels. The output of model is a sequence of labels $\hat{y}^{(i)} = \left[\hat{y}_1^{(i)}, \ldots, \hat{y}_c^{(i)}\right]$. If $x^{(i)}$ has the label $\hat{y}_j$, $j \in \{1, 2, \ldots, c\}$, then $\hat{y}_j = 1$ for the vector $\hat{y}^{(i)}$ otherwise $\hat{y}_j = 0$.

To effectively obtain the representative features of ECG recording and achieve ECG classification, we propose a novel model architecture. It takes ECG recording as input and output a sequence of labels. The objective is to minimize the cross entropy between the corresponding reference label $y^{(i)}$ and the predict label $\hat{y}^{(i)}$. The objective loss function is calculated as:

$$Loss = -\sum_{i}\sum_{j} y_j^{(i)} \log\left(\sigma\left(\hat{y}_j^{(i)}\right)\right) + \left(1 - y_j^{(i)}\right) \log(1 - \sigma\left(\hat{y}_j^{(i)}\right)) \tag{1}$$

3.2 Model Architecture

Fig. 1 show our proposed model architecture, which includes backbone network and parallel dual-scale sub-networks, the information fusion layer and classifier layer. Firstly, backbone

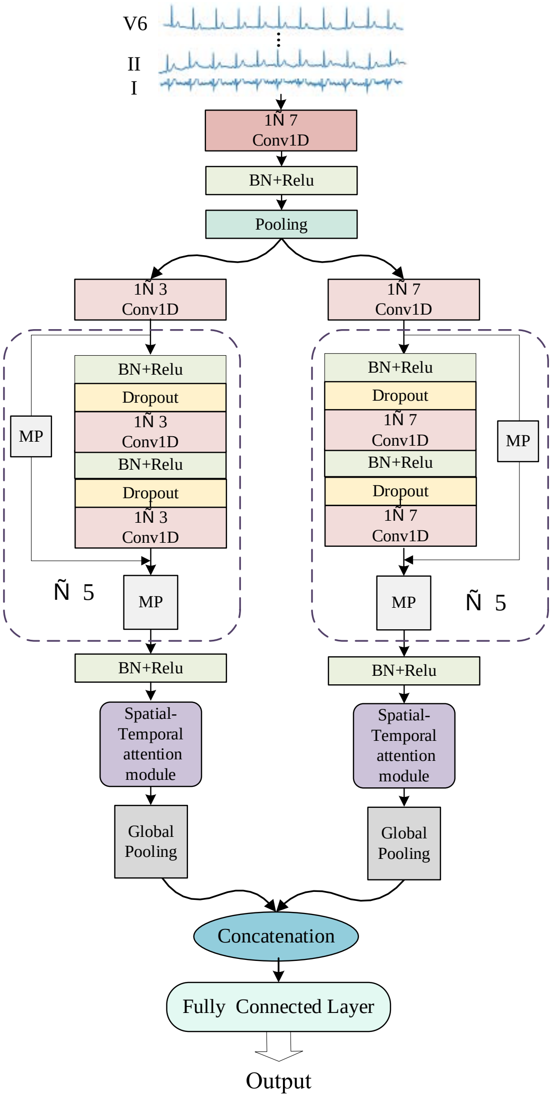

Fig. 1. Diagram of the proposed model architecture.

network is utilized to learn low-level ECG signal information, which makes the subsequent sub-networks get a better feature representation. Then, two parallel sub-networks with different scale filters are used to learn the high-level ECG signal information. To emphasize the representative local spatial and temporal features, we add a spatial-temporal attention to each sub-network. In addition, the global features of the ECG signal are also important for ECG classification. Therefore, the global pooling is used at each sub-network followed by spatial-temporal attention to squeeze the features dimension and obtained the global information. Finally, the various types of information of the two sub-networks is fused effectively using concatenation operation.

3.2.1 The Backbone Network

The backbone network is utilized to obtain low-level features effectively from the ECG signal. There are a 1-D convolutional layer and a pooling layer in the backbone network. 32 convolution kernels are used in 1-D convolutional layer. The size of kernel is 7. The size of stride is 1. Batch normalization (BN) can improve the training speed and help the model converge faster. Therefore, we add a BN layer after 1-D convolutional layer. A rectified linear units (ReLU) activation function is employed to introduce a nonlinearity, and its non-linear function is defined as $\sigma(x) = \max(0, x)$. It is added after BN layer. The pooling layer adopts max-pooling for downsampling. The filter of size is 3. The size of stride is 2. For input ECG recording $x^{(i)}, i \in \{1, 2, ..., m\}$, the output f_a of backbone network is defined as follows:

$$f_a = Net_a\left(x^{(i)}; \theta_a\right) \tag{2}$$

where f_a denotes the characteristic vector of raw input $x^{(i)}$. Net_a represents the backbone network and θ_a represents the network parameter.

3.2.2 The Parallel Dual-scale Sub-networks

Two parallel sub-networks receive the features obtained by the backbone network and get deeper feature representations. Each sub-network is consisted of a 1-D convolutional layer, 5 1-D convolutional blocks, a spatial-temporal attention and one global pooling layer. The convolution of two sub-networks uses different sizes convolutional filters, which enhance features diversity. The convolutional block adopts the shortcut connections. With the network depth increasing, the gradient will gradually disappear, which results in the weight of the previous network layer cannot be updated. The performance of the model will decline. The shortcut connections of the convolutional block improve backpropagation in deep neural networks and solve network optimization problems. The accuracy of model will improve. To merge the local spatial and temporal information, we use the spatial-temporal attention after each sub-network. Finally, global pooling layer squeezes the features dimension and obtains the global information. The output of the sub-networks is $f_{b_j}, j \in \{1, 2\}$.

$$f_{b_j} = Net_{b_j}\left(f_a; \theta_{b_j}\right) \tag{3}$$

1-D convolutional layer and 1-D convolutional block: The dual-scale sub-networks use 3 and 7 convolutional filters respectively. 1-D convolutional layer is followed by 1-D convolutional block. There are 2 convolutional layers, which is preceded by BN, ReLU activation function and dropout. The second convolutional layer is followed by max pooling. Max pooling can subsample and make the representation become approximately invariant. It reduces dimension of the input by half and obtain the maximum values in the local field. The filter of size is 2. The size of stride is 2. The shortcut connections also contain a max pooling, which is used to subsample and adjust size of feature. The filter of size is 2. The final convolutional block is given to a BN and ReLU activation function before spatial-temporal attention. We add dropout after ReLU activation function to reduce over overfitting and to accelerate the training procedure.

The spatial-temporal attention: The spatial-temporal attention is utilized to pay attention to representative local spatial-temporal features from ECG signal. The spatial attention can generate the weight of feature channels, which adaptively focuses on channels containing abnormal features and avoids irrelevant channels. The temporal attention is utilized to obtain relatively important time series information, which make the network to emphasize the essential time periods features.

Fig. 2 shows spatial attention mechanism. Global average-pooling and global max-pooling can reduce the dimension of the feature map and use different ways to separately aggregate input features of long sequence features of each channel. ECG signal anomalies can be classified into intermittent and continuous anomalies [16]. Global average-pooling is suitable for capturing anomalies occurring in long sequences by average operation. Global max-pooling can effectively capture intermittent anomalies in the ECG signal. The features compressed by global average-pooling and global max-pooling are fed into two shared fully connected layers. Then two channel weight vectors are generated by nonlinear operations. Two channel weight vectors are combined through element-wise addition. Then the spatial attention weights are obtained, which reflect the relative importance between channels. The relative important channels will be given the larger weights. The relative unimportant channels will be given the smaller weights. A sigmoid function is applied to scale the weights in 0-1. Finally, the spatial attention weight multiplies by the output of 1-D convolutional block layer, which help the network emphasize significant channel features. For input $S=[s_1,s_2,\cdots,s_c], s_i \in R^{1\times L}$, C represents the number of channels and L represents the temporal length of S.

$$z_{avg}=F_f\left(AvgPool(S)\right) \tag{4}$$

$$z_{\max}=F_f\left(MaxPool(S)\right) \tag{5}$$

$$z(z_1,...,z_C)=\sigma(z_{avg}+z_{\max}) \tag{6}$$

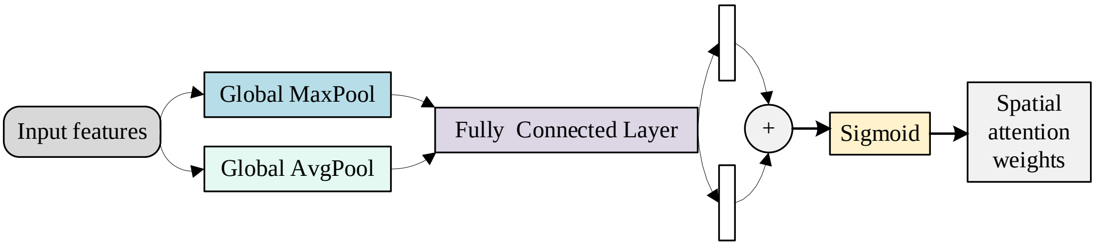

Fig. 2. Diagram of spatial attention mechanism.

where $AvgPool(\cdot)$ denotes global average pooling operation, $MaxPool(\cdot)$ denotes global maximum pooling operation, F_f denotes dense layers and non-linear operation and σ denotes sigmoid operation. z represents the contribution of different the channel. Finally, the spatial attention weight z is used to multiplies S to generate $\hat{S}$.

$$\hat{S} = z \otimes S = \left[z_1 s_1, z_2 s_2, \ldots, z_C s_C \right] \tag{7}$$

Fig. 3 shows temporal attention mechanism. The temporal attention mechanism is used to encode the relative importance between the time periods. Global average-pooling and global max-pooling adopt different methods to aggregate channel information at each time of the feature map, respectively. The features compressed by global average-pooling and global max-pooling are concatenated and fed into one convolution layer. Then temporal attention weights are generated, which reflect the relative importance between time periods. Then, A sigmoid function is applied to scale the weights in 0-1. Finally, the temporal attention weight multiplies by the output of spatial attention, which makes the network enhance important time features. For input $T = [t_1, t_2, \ldots, t_L], t_i \in R^{1\times C}$, L represents the length of time T.

$$n(n_1, \ldots, n_L) = F_c\left(\left[AvgPool(T); MaxPool(T)\right]\right) \tag{8}$$

where $AvgPool(\cdot)$ denotes global average pooling operation, $MaxPool(\cdot)$ denotes global maximum pooling operation and $F_c(T)$ denotes the convolution operation. The convolution kernel size is 7. n indicates the contribution of different features in the time periods. Finally, the temporal attention weight is utilized to multiplies T to get $\hat{T}$.

$$\hat{T} = n \otimes T = \left[\sigma(n_1) t_1, \sigma(n_2) t_2, \ldots, \sigma(n_L) t_L \right] \tag{9}$$

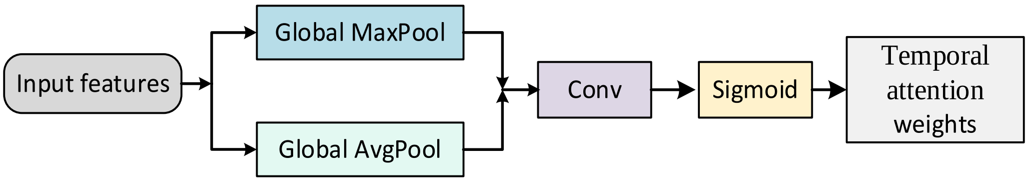

Fig. 3. Diagram of spatial attention mechanism.

The global pooling layer: The feature dimensions extracted from convolutional block of two sub-networks are different, which is not conducive to feature fusion by subsequent concatenation operations. In addition, global feature extraction from signal is important for ECG classification. Therefore, we add a global pooling layer after the spatial-temporal attention of each sub-network. Global pooling layer can compress the feature dimension and obtain the global features representation. Global pooling includes global max-pooling and global average-pooling. The former is more suitable for our model architecture and enables our model to achieve greater classification effect. Since we use global average-pooling. It is defined as follows:

$$g_{avg}\left(\hat{T}\right)=\frac{1}{L}\sum_{i}^{L}\hat{t}_i \tag{10}$$

where g_{avg} denotes global average-pooling operation.

3.3 The information fusion and classifier layer

For learning complementary information obtained by dual-scale sub-networks and obtaining robust features for classification. We fuse features extracted from dual-scale sub-networks by the concatenation operation. The final feature F can be obtained using the following calculation:

$$F = Cat(f_{b_1}, f_{b_2}) \tag{11}$$

where f_{b_1}, f_{b_2} denote the output of dual-scale sub-networks. Cat represents the concatenation operation.

Finally, we use a fully connected layer as a classifier, which consists of c neurons. The output of classifier is mapped by a sigmoid activation function to get probability $p=\left(p_1,\ldots,p_c\right)$, where c represents the quantity of classes and $p_j, j\in\{1,2,\ldots,c\}$ represents the probability that $x^{(i)}$ has the label $\hat{y}_j$. We use θ to obtain $\hat{y}^{(i)}=\left[\hat{y}_1^{(i)},\ldots,\hat{y}_c{}^{(i)}\right]$, which is the final prediction. We set θ=0.5

$$\hat{y}_j=\begin{cases}1, & \text{if } \mathrm{p}_j\geq\theta\\ 0, & \text{otherwise}\end{cases} \tag{12}$$

4. Experiment

4.1 Data Source

The experimental ECG dataset is PTB-XL dataset [17]. This dataset includes 21837 12-lead ECG recordings, which come from 18885 patients with 52% male and 48% female. ECG recordings has 100 Hz sampling frequency and 500 Hz sampling frequency. The duration of the ECG recording is 10 seconds. The labels for ECG recording are manually annotated by ECG experts. It comprises 71 distinct ECG statements that are categorized into 44 diagnostic statements, 19 form statements, and 12 rhythm statements. The diagnostic statements are further divided into diagnostic_superclass and diagnostic_subclass. Every ECG recording may have multiple statements. In our experiments, we use the 100 Hz sampling frequency and 5 diagnostic_superclass labels, which consist of NORM, CD, MI, and STTC. The distribution of diagnostic_superclasses is showed in **Fig. 4**. **Fig. 5** shows visualization of an example of a 12-lead ECG recording, which contain MI and STTC.

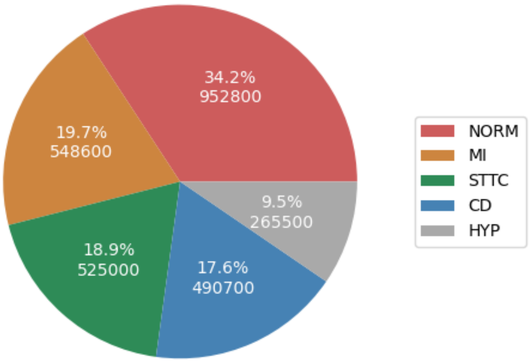

Fig. 4. The distribution of diagnostic_superclasses in PTB-XL dataset.

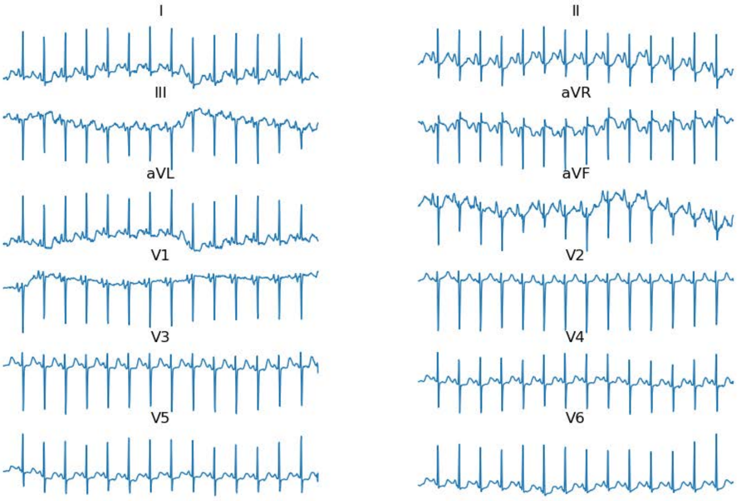

Fig. 5. An example of 12-lead ECG record.

4.2 Experimental Setup

All ECG recordings in PTB-XL dataset are standardized with a zero mean and unit variance. We split the ECG recording into ten folds. Every ECG recording belongs to one of 10 folds. We use 1 to 8 folds ECG recordings as train set. The remaining two folds is utilized to validate model and test model respectively. All experiments are run on a Nvidia GTX 2080Ti 24GB GPU machine. We used Cross-Entropy loss and Adam optimizer for training the model. The maximum number of epochs is 30. The batch size is 32. We applied a learning rate decay strategy in which we decreased the learning rate by a factor of 10 every 10 epochs. The initial learning rate was set to 0.01.

4.3 Evaluation Metrics

Since ECG recordings have more than one statement, which correspond to multiple labels. We evaluate performance of our proposed model using multiple label classification metric, including macro ROC-AUC, mean accuracy, and maximum F1 score [18]. For the j_{th} disease $y_j^{(i)}$, TP_j, FP_j, TN_j, FN_j are used to evaluate binary classification performance. ECG recordings with the j_{th} disease is positive, while ECG recordings without the j_{th} disease are negative. TP_j denotes the quantity of ECG recordings that belong to true positive, FP_j denotes the quantity of ECG recordings that belong to false positive, TN_j denotes the quantity of ECG recordings that belong to true negative, FN_j denotes the quantity of ECG recordings that belong to false negative. Based TP_j, FP_j, TN_j, FN_j, the F1 score and mean accuracy can be calculated as follows:

$$Precision = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|y^{(i)} \cap \hat{y}^{(i)}\right|}{\left|\hat{y}^{(i)}\right|} \tag{13}$$

$$Recall = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|y^{(i)} \cap \hat{y}^{(i)}\right|}{\left|y^{(i)}\right|} \tag{14}$$

$$F1\ score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{15}$$

$$Mean\ Accuracy = \frac{1}{C}\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}\left[\!\left[y_j^{(i)} = y^{(i)} \right]\!\right] \tag{16}$$

4.4 Results and Discussion

To evaluate of our proposed model, we compare our proposed model with other reference models Mousavi et al. [19], ECGNet [4], Zhang et al. [20], Wang et al. [21]. Mousavi et al. [19] propose a sequence-to-sequence model for ECG classification, which includes three convolutional layers and the bidirectional RNN. The model proposed in the literature [4] combines CNN and LSTM. Features are obtained separately using CNN and LSTM. Finally, the features are fused for ECG classification. Zhang et al. [20] propose a CNN model with residual blocks. Wang et al. [21] propose a multiscale model. The backbone network uses the traditional 34-layer residual network. **Table 1** shows comparison results. Our model outperforms other four models. Specifically, our model achieves 0.9307 on macro ROC-AUC, which outperforms the other models about 6.53%, 2.06%, 1.78%, 1.33% respectively. Our model achieves 89.11 on Mean Accuracy, which outperforms the other models about 4.92, 1.76, 1.63, 1.39, respectively. Our model achieves 0.8152 on F1 score (Max), which outperforms the other models 8.37%, 4.4%, 3.04%, 2.56%, respectively. **Fig. 6** shows a performance of class-wise ROC-AUC with other four models. Our model outperforms others models in ROC-AUC for CD, HYP, MI, NORM. **Fig. 7** shows a performance of class-wise accuracy with other four models. The accuracy of our models in CD, MI, NORM and STTC is superior to all other models.

Table 1. Comparison results of our model with related works

	Macro ROC-AUC	Mean Accuracy	F1 score (Max)
Mousavi et al. [19]	0.8654	84.19	0.7315
ECGNet [4]	0.9101	87.35	0.7712
Zhang et al. [20]	0.9129	87.48	0.7848
Wang et al. [21]	0.9174	87.72	0.7896
STADSNet	**0.9307**	**89.11**	**0.8152**

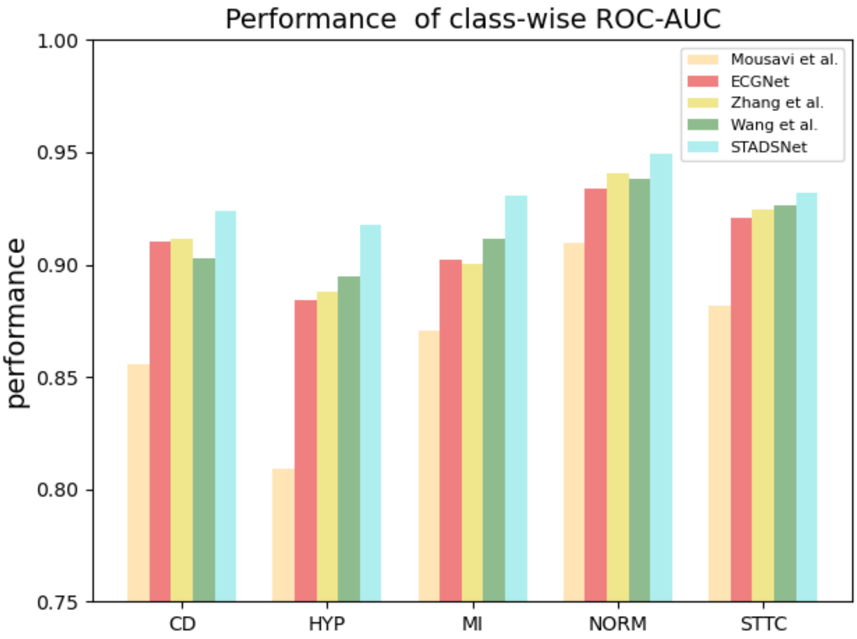

Fig. 6. Performance of class-wise ROC-AUC

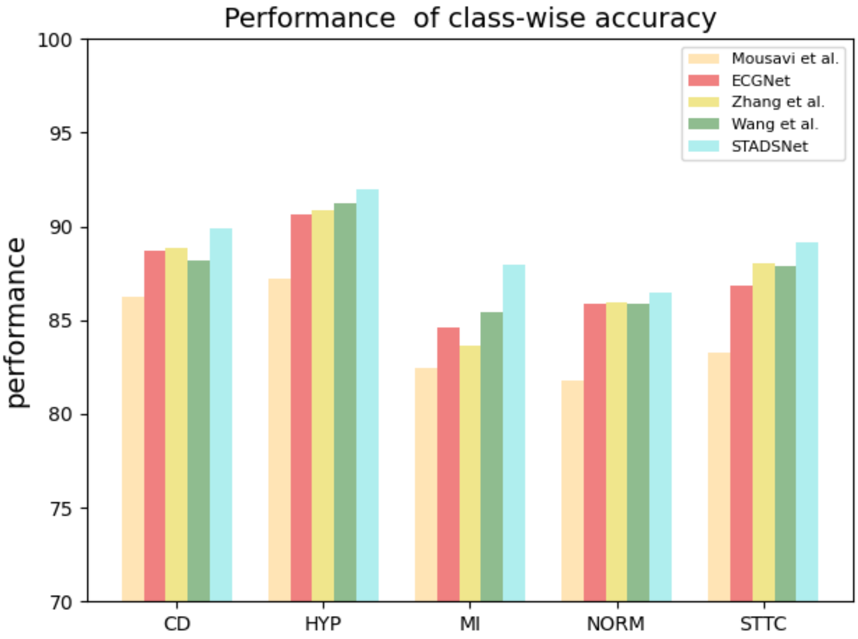

Fig. 7. Performance of class-wise accuracy

Table 2 and Table 3 show ROC-AUC and accuracy of model trained on single-lead ECG, respectively. The results show that using 12-lead ECG achieves the better performance compared to using single-lead ECG. When model only use single-lead ECG, the macro ROC-AUC of the model decreases by 6.38%-13.47% compared with the 12-lead ECG. The mean accuracy of the model trained on the single-lead ECG decreases by 4.19-7.95. When only using single-lead ECG, the model has the best performance in lead aVR, V5 and V6. The macro ROC-AUC achieves 0.8679, 0.8639 and 0.8669, respectively. The mean accuracy achieves 84.78, 84.92 and 84.88, respectively. The ROC-AUC of STTC achieves 0.9101, 0.9190 and 0.9125, respectively. The accuracy of HYP achieves 89.27, 90.57 and 90.15, respectively. When only using single-lead aVR, the ROC-AUC of STTC achieves 0.9101. This indicates that lead aVR, V5 and V6 are very important for the diagnosis of some diseases. When only using single-lead lead III, the model has the worst performance. Specifically, macro ROC-AUC is 0.7960 and mean accuracy is 81.16.

Table 2. Performance of ROC-AUC trained on single-lead ECG

	CD	HYP	MI	NORM	STTC	Macro ROC-AUC
I	0.8221	0.8163	0.7908	0.8881	0.8767	0.8388
II	0.8757	0.8005	0.8281	0.9023	0.8836	0.8580
III	0.8147	0.7645	0.8228	0.8442	0.7339	0.7960
aVR	0.8700	0.8318	0.8202	0.9074	0.9101	**0.8679**
aVL	0.8208	0.7913	0.8014	0.8719	0.8046	0.8180
aVF	0.8464	0.7848	0.8305	0.8743	0.8088	0.8290
V1	0.8687	0.8150	0.8010	0.8511	0.7851	0.8242
V2	0.8417	0.7584	0.8013	0.8305	0.7952	0.8054
V3	0.8242	0.7673	0.8186	0.8700	0.8392	0.8238
V4	0.8220	0.8247	0.7968	0.8945	0.8946	0.8465
V5	0.8340	0.8545	0.8029	0.9092	0.9190	**0.8639**
V6	0.8506	0.8493	0.8133	0.9087	0.9125	**0.8669**
All	**0.9240**	**0.9177**	**0.9309**	**0.9492**	**0.9319**	**0.9307**

Table 3. Performance of accuracy trained on single-lead ECG

	CD	HYP	MI	NORM	STTC	Mean Accuracy
I	84.56	89.18	78.41	80.12	83.63	83.18
II	86.92	88.26	80.07	81.65	84.88	84.36
III	84.37	88.40	80.58	75.13	77.30	81.16
aVR	85.71	89.27	80.31	82.29	86.32	**84.78**
aVL	85.71	89.00	79.20	77.76	79.38	82.21
aVF	84.28	88.26	81.78	78.78	80.54	82.73
V1	86.22	88.58	79.75	75.03	78.32	81.58
V2	85.30	88.21	83.36	74.34	80.17	82.27
V3	84.88	88.12	83.26	77.67	81.00	82.99
V4	83.87	88.95	80.21	80.40	85.02	83.69
V5	84.93	90.57	79.84	81.92	87.33	**84.92**
V6	85.58	90.15	79.33	82.62	86.73	**84.88**
All	**89.92**	**91.96**	**87.98**	**86.50**	**89.18**	**89.11**

The performance of the network is influenced by the size of the convolutional kernel. To analyze the impact of different convolution kernel sizes, we set different convolution kernel sizes for two sub-networks. The convolution kernels of two subnetworks are selected in 3, 5, and 7, respectively. As can be seen from the **Table 4**, when the kernel size of the two sub-networks is 3 and 7 respectively, model achieves the best Macro ROC-AUC performance. When the kernel size of the two sub-networks is 3, 5, respectively, the performance of the model drops. The worst performance is obtained when the kernel size of the two sub-networks is 5 and 7. In addition, the performance of the model is worse when two sub-networks use the same convolutional kernel size of (3, 7) compared to using different kernel sizes.

Table 4. Comparison of different convolution kernel size models

	STDSNet35	STDSNet37	STDSNet57	STDSNet33	STDSNet77
Macro ROC-AUC	0.9287	**0.9307**	0.9258	0.9274	0.9262
F1 score (Max)	**0.8154**	0.8152	0.8109	0.8129	0.8102
Mean Accuracy	**89.26**	89.11	89.13	89.10	88.98

We conducted ablation experiments to analyze the impact of dual-scale network structure and spatial-temporal attention. We compared the classification results of our model, our model without spatial-temporal attention, and single-scale network. The single-scale network includes a backbone network, a sing-scale sub-network, a spatial-temporal attention, and the global pooling layer. The convolution kernel size of sub-network is 3. **Table 5** shows the ablation results. We can observe that the dual-scale model outperforms the single-scale model. Specifically, the macro ROC-AUC increases from 0.9276 to 0.9307, F1 score (Max) increases from 0.8082 to 0.8152 and mean Accuracy increases from 89.05 and 89.11, respectively. The performance of the model can also be enhanced by adding spatial-temporal attention. Specifically, macro ROC-AUC increases from 0.9272 to 0.9307, F1 score (Max) increases from 0.8133 to 0.8152 and mean accuracy increases from 89.05 and 89.01, respectively.

Table 5. Ablation experiments results of proposed model

	Macro ROC-AUC	F1 score (Max)	Mean Accuracy
w/o spatial-temporal attention	0.9276	0.8082	89.05
Single-scale network	0.9272	0.8133	89.01
Ours	**0.9307**	**0.8152**	**89.11**

5. Conclusion

In this paper, we propose a dual-scale network with spatial-temporal attention for ECG classification. Different from the previous single-scale network, we utilize a dual-scale network for feature extraction from ECG signal. We also employ a spatial-temporal attention to learn the appropriate weights for different channels and temporal feature, which facilitates focusing on representative local spatial and temporal features. Finally, global pooling is also used to further obtain global features. We use PTB-XL dataset to evaluate the proposed model, which achieves 0.9307 on macro ROC-AUC, 0.8152 on F1 score (Max) and 89.11 on mean accuracy. It outperforms several existing models. To evaluate the impact of the proposed improvements, we also performed an ablation study. Ablation experimental results show that the two improvements proposed in this paper can improve the performance of the model.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (Nos. 62271486, 62071470). Chaogang Tang is the corresponding author.

References

[1] WHOFactSheet, 2021 [Online]. Available: http://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[2] Q. Yao, R. Wang, X. Fan, et al., "Multi-class arrhythmia detection from 12-lead varied-length ECG using attention-based time-incremental convolutional neural network," *Information Fusion*, vol. 53, pp. 174-182, 2020. Article (CrossRef Link)

[3] R. Wang, J. Fan, and Y. Li, "Deep Multi-Scale Fusion Neural Network for Multi-Class Arrhythmia Detection," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 9, pp. 2461-2472, 2020. Article (CrossRef Link)

[4] B. Murugesan, V. Ravichandran, K. Ram, et al., "ECGNet: Deep Network for Arrhythmia Classification," in *Proc. of 2018 IEEE International Symposium on Medical Measurements and Applications*, pp. 1-6, 2018. Article (CrossRef Link)

[5] C. Ye, B. V.K. Vijaya Kumar, and M.T. Coimbra, "Heartbeat Classification Using Morphological and Dynamic Features of ECG Signals," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 10, pp. 2930-2941, 2012. Article (CrossRef Link)

[6] J. A. Nasiri, M. Naghibzadeh, H. S. Yazdi and B. Naghibzadeh, "ECG Arrhythmia Classification with Support Vector Machines and Genetic Algorithm," in *Proc. of 2009 Third UKSim European Symposium on Computer Modeling and Simulation*, pp. 187-192, 2009. Article (CrossRef Link)

[7] U. R. Acharya, H. Fujita, M Adam, et al., "Automated characterization of arrhythmias using nonlinear features from tachycardia ECG beats," in *Proc. of 2016 IEEE international conference on systems, man, and cybernetics (SMC)*, pp. 000533-000538, 2016. Article (CrossRef Link)

[8] Li. H, D. Yuan, Y. Wang, et al., "Arrhythmia classification based on multi-domain feature extraction for an ECG recognition system," *Sensors*, vol. 16, no. 10, pp. 1744, 2016. Article (CrossRef Link)

[9] A.Y. Hannun, P. Rajpurkar, M. Haghpanahi, et al., "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature medicine*, vol. 25, no. 1, pp. 65-69, 2019. Article (CrossRef Link)

[10] R. Wang, Q. Yao, X. Fan, et al., "Multi-class arrhythmia detection based on neural network with multi-stage features fusion," in *Proc. of 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pp. 4082-4087, 2019. Article (CrossRef Link)

[11] T. Wang, C. Lu, Y. Sun et al., "Automatic ECG classification using continuous wavelet transform and convolutional neural network," *Entropy*, vol. 23, no. 1, pp. 119, 2021. Article (CrossRef Link)

[12] B. Hou, J. Yang, P. Wang and R. Yan, "LSTM-Based Auto-Encoder Model for ECG Arrhythmias Classification," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 4, pp. 1232-1240, 2020. Article (CrossRef Link)

[13] G. Ma, X. Wang, and J. Yu, "ECG signal classification algorithm based on fusion features," *Journal of Physics: Conference Series*, vol. 1207, no. 1, pp. 012003, 2019. Article (CrossRef Link)

[14] A. Chen, F. Wang, et al., "Multi-information fusion neural networks for arrhythmia automatic detection," *Computer methods and programs in biomedicine*, vol. 193, pp. 105479, 2020. Article (CrossRef Link)

[15] X. Li, C Li, Y. Wei, et al., "BaT: Beat-aligned Transformer for Electrocardiogram Classification," in *Proc. of 2021 IEEE International Conference on Data Mining (ICDM)*, pp. 320-329, 2021. Article (CrossRef Link)

[16] J. Zhang, A. Liu, M. Gao, et al., "ECG-based multi-class arrhythmia detection using spatio-temporal attention-based convolutional recurrent neural network," *Artificial Intelligence in Medicine*, vol. 106, pp. 101856, 2020. Article (CrossRef Link)

[17] P. Wagner, N. Strodthoff, R. D. Bousseljot, et al., "PTB-XL, a large publicly available electrocardiography dataset," *Sci Data*, vol. 7, no. 1, pp. 154, 2020. Article (CrossRef Link)

[18] L. Reddy, V. Talwar, S. Alle, et al., "IMLE-Net: An Interpretable Multi-level Multi-channel Model for ECG Classification," in *Proc. of 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 1068-1074, 2021. Article (CrossRef Link)

[19] S. Mousavi, and F. Afghah, “Inter- and Intra-Patient ECG Heartbeat Classification for Arrhythmia Detection: A Sequence to Sequence Deep Learning Approach,” in *Proc. of ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing*, pp. 1308–1312, 2019. Article (CrossRef Link)

[20] D. Zhang, S. Yang, X. Yuan, and P. Zhang, “Interpretable deep learning for automatic diagnosis of 12-lead electrocardiogram,” *iScience*, vol. 24, no. 4, pp.102373, 2021. Article (CrossRef Link)

[21] S. Wang, R. Li, X. Wang, et al., “Multiscale residual network based on channel spatial attention mechanism for multilabel ECG classification,” *Journal of Healthcare Engineering*, p. 6630643, 2021. Article (CrossRef Link)

Shuo Xiao received Ph.D. degree from the School of Electronic and Information Engineering, Beijing Jiaotong University. He works in China University of Mining and Technology since 2010, where he is now an Associate Professor. His research interest includes the internet of things, intelligent information processing and artificial intelligence.

Yiting Xu received her B.S. degree from the Kunming University of Science and Technology. She is currently a graduate student at the School of Computer Science and Technology, China University of Mining and Technology. Her research interests include medical signal processing and deep learning.

Chaogang Tang received his B.S. degree from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, and Ph.D. degree from the School of Information Science and Technology, University of Science and Technology of China, Hefei, China, and the Department of Computer Science, City University of Hong Kong, under a joint Ph.D. Program, in 2012. He is now with the China University of Mining and Technology. His research interests include mobile cloud computing, fog computing, Internet of Things, big data.

Zhenzhen Huang received Ph.D. degree from the School of Computer Science and Technology, China University of Mining and Technology. Her main research interests are personalized recommendation, intelligent information processing and artificial intelligence.