DSTNN-X: ECG Arrhythmia Classification using Deep Spatio-Temporal Learning with Multi-Lead Contextual Aggregation

Yanchong Xie[1], Jiepeng Chen[1], Kai Zheng[2], Haoyang Huang[1], Zhanshang Nie[1], Minglong Zheng[3], Kai Huang[1], and Mingyue Cui[1,*]

[1]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, 400100, China
[2]School of Biomedical Engineering, Sun Yat-sen University, Guangzhou, 400100, China
[3]Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou, 400100, China

Corresponding author: cuimy5@mail.sysu.edu.cn

Abstract. Automated arrhythmia detection from electrocardiogram (ECG) signals is crucial for the early prevention and diagnosis of cardiovascular diseases. However, conventional ECG classification methods usually employ 1D convolutions to extract features from individual leads separately, neglecting the intrinsic correlations inherent in 12-lead ECGs. This paper proposes DSTNN-X, a deep neural network that performs spatio-temporal learning by effectively aggregating multi-lead contextual information for ECG arrhythmia classification. Specifically, a depth-stationary cross convolution (DSC-Conv) is introduced to synergistically integrate horizontal convolutions with vertical convolutions, effectively extracting temporal patterns and spatial dependencies of intra-group leads. Besides, a residual structure is designed, combining temporal convolution with cross-lead features to enhance inter-group lead correlations. Finally, multi-head attention and a linear layer are employed to map arrhythmia diagnosis labels. Experimental results on public multi-label ECG datasets PTB-XL and CPSC2018 demonstrate that the proposed lead-fusion network achieves consistent performance improvements across diverse arrhythmia classification tasks, outperforming current state-of-the-art models.

Keywords: Arrhythmia classification, Deep learning, ECG, Biomedical signal processing.

1. INTRODUCTION

Cardiovascular diseases (CVDs) are among the most significant global health challenges worldwide. In 2021, it was estimated that approximately 20 million individuals succumbed to CVD globally, accounting for nearly one-third of total global mortality [1]. Among cardiovascular disorders, cardiac arrhythmias represent one of the most prevalent electrophysiological disturbances, with malignant variants carrying a significant risk of sudden cardiac death. The 12-lead electrocardiogram (ECG) offers a non-invasive, real-time assessment of cardiac electrical activity, which is essential for the diagnosis of arrhythmias. However, due to the complexity of ECG signals, conventional ECG interpretation relies heavily on manual analysis by trained specialists, which is not only labor-intensive but also susceptible to interobserver variability and diagnostic inaccuracy. Thus, how to achieve accurate arrhythmia diagnosis has become a crucial issue.

With the rapid evolution of deep learning architectures, AI-assisted cardiac diagnosis has emerged as a new clinical paradigm. Deep neural networks can automatically extract meaningful features from complex physiological data, which makes their gradual application in ECG arrhythmia classification possible. For instance, Yao *et al.* [2] propose an ATI-CNN model for cardiac arrhythmia detection, combining convolutional neural networks (CNN), recurrent cells, and an attention module to learn contextual information from each lead. Yang *et al.* [3] introduce an MVMS-Net that employs a multi-scale CNN architecture for feature extraction and explores the knowledge distillation technique to reduce model complexity. However, those methods are usually unable to fully utilize the contextual information of the signal.

Achieving accurate arrhythmia diagnosis based on ECG signals is not easy. It mainly faces the following challenges: Firstly, as a complex dynamic representation of cardiac electrophysiological activity, ECG is characterized by its inherent non-stationarity and quasi-periodicity. These properties make it difficult to capture features that precisely reflect arrhythmias by morphological analysis with pre-set thresholds or machine learning methods utilizing fixed feature extraction techniques. Besides, the ECG waveform variations indicative of arrhythmias are

This work is supported in part by the College Students' Innovative Entrepreneurial Training Plan Program of Sun Yat-sen University, Grant No.67000-11250012.

subtle and highly susceptible to interference from physiological variability and technical noise, potentially leading to ambiguous or erroneous diagnoses. For example, in the diagnosis of early myocardial ischemia or non-ST-segment elevation myocardial infarction (NSTEMI), the deviation of ST-segment depression is only 0.05 mV to 0.1 mV. Therefore, an efficient diagnosis method for ECG arrhythmia classification is needed.

To solve the above problems, this paper proposes a novel 12-lead ECG arrhythmia classification method based on deep learning, called DSTNN-X, which can efficiently capture multi-lead contextual information. Initially, the 12-lead ECG signals are divided into four groups based on anatomy and electrophysiological structure. Then, a depth-stationary cross convolution (DSC-Conv) is introduced to capture spatial correlations and temporal dependencies of intra-group leads through horizontal and vertical convolutions. Besides, to enhance inter-group lead correlations, a bidimensional adaptive residual convolution (BARC) module is designed, which combines horizontal convolution with cross-lead feature aggregation. Finally, a multi-head attention is employed to aggregate the salient patterns across channels and map unified representations to the arrhythmia diagnosis label using a linear layer.

The proposed DSTNN-X is compared with other baseline methods such as LSTM [4], BiLSTM [5], fcn_wang [6], Xresnet1d101 [7], MobileNetV3 [8], ViT [9], InceptionTime [10], ATI-CNN [2], and MVMSNet [3]. Experimental results show that the method achieves 92.83% and 96.65% AUC (area under the receiver operating characteristic curve) for arrhythmia classification on the PTB-XL [11] and CPSC2018 [12] datasets respectively, outperforming all baseline methods. The main contributions are summarized as follows:

- This paper proposes a deep-learning-based spatio-temporal network called DSTNN-X with multi-lead contextual aggregation, which achieves efficient ECG arrhythmia classification.
- To capture spatial correlations and temporal dependencies in groups, a depth-stationary cross convolution (DSC-Conv) is designed to integrate horizontal convolutions with vertical convolutions.
- A residual hierarchical feature learning structure (BARC) combining horizontal convolution with cross-lead feature aggregation is introduced to enhance inter-group lead correlations.

The rest of this paper is organized as follows: Section I gives a brief introduction about ECG arrhythmia classification and Section II describes the related work. The details of the proposed method are presented in Section III. In Section IV, different experimental results are presented and Section V concludes the paper.

2. RELATED WORK

This section summarizes the related work about ECG arrhythmia classification. The current methods can be broadly categorized into two types: machine learning methods and deep learning methods.

2.1. Machine Learning Methods

Traditional machine learning methods usually extract ECG features such as HRV metrics, waveform morphology, and wavelet transform coefficients from signals, and input them into machine learning classifiers such as support vector machines (SVM), k-nearest neighbors (kNN), decision trees, and ensemble techniques for arrhythmia classification. For example, Karimifard *et al.* [13] introduce hermitian basis functions for feature extraction from ECG signals, coupled with a kNN classifier for arrhythmia classification. Pham *et al.* [14] explore the classification performance for different machine learning methods in 18 extracted nonlinear features from third-order cumulant images of ECG, demonstrating the superiority of the random forest classifier. While these methods have achieved certain classification results in specific arrhythmia classification tasks, their performance is fundamentally limited by the quality and distinctiveness of manually designed features, which restricts further improvement in classification accuracy.

2.2. Deep Learning Methods

In recent years, researchers try to use deep neural networks to automatically learn hierarchical features and temporal information from raw ECG signals for arrhythmia classification, such as fully convolutional networks (FCN) [6], ResNet [7], long short-term memory (LSTM) [4] and bidirectional long short-term memory (BLSTM) [5]. Yildirim *et al.* [15] propose a 16-layer deep 1D-CNN model, which combines convolutional feature extraction with max pooling downsampling for arrhythmia detection. Hou *et al.* [16] design a deep learning architecture that utilizes an LSTM autoencoder to extract features and subsequently employs a SVM to classify arrhythmia. However, the arrhythmia classification performance of the above methods is limited by issues such as local receptive fields and long-range dependencies.

Recently, attention-based architectures introduce a dynamic, context-aware weighting mechanism to capture long-range contextual relationships, making it more suitable for analyzing long sequential ECG signals. Yao *et*

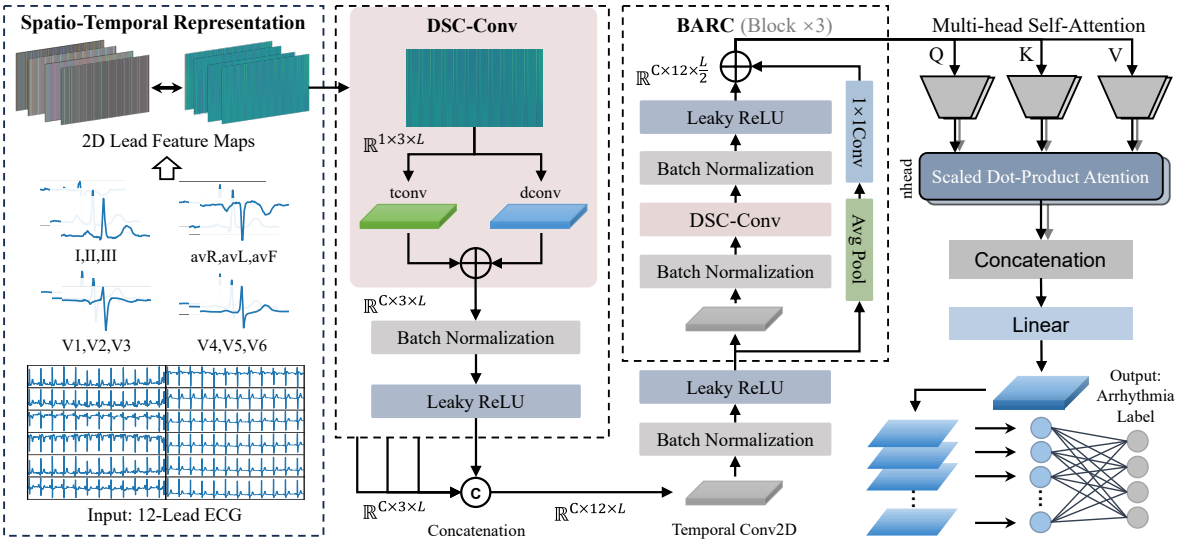


Fig. 1. The framework of DSTNN-X, which mainly consists of four parts: spatio-temporal representations of ECG leads, DSC-Conv blocks for extracting intra-group spatial features, BARC blocks for enhancing inter-group lead correlations, multi-head attention and a linear layer to map arrhythmia diagnosis labels.

al. [2] propose an attention-based time-incremental convolutional neural network (ATI-CNN), which combines convolutional modules for local feature extraction and attention mechanisms to achieve arrhythmia classification. Yang *et al.* [3] develop a multi-view multi-scale neural network (MVMS-Net), which segments 12-lead ECGs into multiple views and employs multi-scale convolutional blocks coupled with attention mechanisms for arrhythmia detection. Although these models can identify distinct long-term arrhythmia features for each lead, they still do not fully utilize the inherent correlations and spatio-temporal dependencies among leads, which limits classification performance.

3. METHODS

As shown in Fig. 1, a deep spatio-temporal neural network with multi-lead contextual aggregation called DSTNN-X is proposed for ECG arrhythmia classification. Firstly, the 12-lead ECG signals are divided into four groups: (I, II, III), (aVR, aVL, aVF), (V1, V2, V3), and (V4, V5, V6), and vertically concatenated along the lead axis to generate four aggregated feature maps. Then, the DSC-Conv is applied on the aggregated feature maps of each group, capturing intra-group spatial correlations and temporal dependencies through horizontal and vertical convolutions. To enhance inter-group lead correlations, all feature maps are concatenated along the lead axis and propose the BARC to enhance discriminative pattern extraction through horizontal convolution with residual connection. After that, the aggregated feature maps of each lead are flattened, and multi-head self-attention integrates the contextual information across all leads. Finally, adaptive max pooling downsamples the features to preserve the most salient patterns across channels, followed by a linear layer that maps the refined representations to the final arrhythmia diagnosis label.

3.1. Feature Embedding of 12-lead ECGs

In the feature embedding stage, 12-lead ECG signals are transformed into four 2D aggregated spatio-temporal feature maps. Specifically, given the original 12-lead ECG signal $X \in \mathbb{R}^{12\times L}$, the leads are first divided into four groups based on the similarity of signals determined by their electrophysiological characteristics and anatomical positions. The first two groups are limb leads, while the latter two are precordial leads. The first group comprises bipolar limb leads I, II, III from Einthoven's triangle, which generates frontal plane vectors and provides an electrical view of the heart along the superior-inferior axis, capturing global cardiac electrical orientation and detecting broad repolarization abnormalities. The second group consists of augmented unipolar leads aVR, aVL, aVF, which offer more focused perspectives with finer-grained spatial information, improving the localization of frontal plane events. The third group includes leads V1, V2, and V3, which capture the electrical activity from the right ventricular anterior wall and the interventricular septum. The remaining leads are the fourth group, measured from the left ventricular anterolateral wall, reflecting electrical activity influenced by left anterior descending and circumflex artery perfusion. Then, each group of ECG leads is transformed into a 2D feature map $X_i' \in \mathbb{R}^{1\times 3\times L}$ by

vertically concatenating the lead dimensions. Throughout the embedding process, the temporal dimensions remain the same, while the channel dimension is transformed into a specified latent dimension C to form embedded feature maps $X_i' \in \mathbb{R}^{C\times 3\times L}$.

3.2. Depth-Stationary Cross Convolution (DSC-Conv)

To efficiently capture intra-group spatial correlations and temporal dependencies, this module extracts information in the horizontal and vertical directions of aggregated spatio-temporal feature maps $O \in \mathbb{R}^{C\times H\times L}$, where C is the number of channels, H denotes the number of leads and L denotes the length of signals. Specifically, in the horizontal direction, ECG signals are modeled as the 1D representation $O_L \in \mathbb{R}^{1\times L}$ for each lead h, a 1D convolution operation is performed using a kernel $K_L \in \mathbb{R}^k$ as follows:

$$Y_L(h,l) = \sum_{i=0}^{k-1} K_L(i) \cdot O_L(h, l+i) \tag{1}$$

where $K_L(i)$ represents the kernel weight at position i, $O_L(h,l+i)$ is the value at position $l+i$ in h-th lead, $Y_L(h,l)$ denotes the convolution output in the horizontal direction at position (h,l). Besides, the padding and stride settings are adjusted for sufficient receptive field. For a kernel size k satisfying $k > 1$ and $k \equiv 1 (mod\ 2)$, this paper empirically sets the padding size $p = \frac{k-1}{2}$ applied to both ends of the input signal to extend its dimension, and stride $s = 1$ to determine the displacement of the kernel after each convolution operation. This setting maintains consistency between the output L' and input dimensions L, which can be described as:

$$L' = \lfloor \frac{L+2p-k}{s} \rfloor + 1 = L \tag{2}$$

Therefore, 1D convolution operation in the horizontal direction is further reformulated as:

$$Y_L(h,l) = \sum_{i=0}^{k-1} K_L(i) \cdot O_L(h, l+i-p) \tag{3}$$

Similar to the horizontal directions, ECG signals are also modeled as the 1D representation $O_H \in \mathbb{R}^{H\times 1}$ in the vertical direction, where H denotes the number of leads. However, different from the horizontal convolution, the 1D convolution with the kernel size of $\mathbb{R}^H$ is performed without a padding operation in the vertical direction. For each position $h \in [0, H-1]$, the convolution output at position (j,l) can be computed as:

$$Y_H(h,l) = \sum_{j=0}^{H-1} K_H(j) \cdot O_H(j,l) \tag{4}$$

where $K_H(j)$ represents the kernel weight at position j with $j \in [0, H-1]$ to propagate the aggregated features across all leads, $O_H(j,l)$ is the value in j-th lead at position l.

To integrate the information extracted in both horizontal and vertical directions, a depth-stationary cross-convolution (DSC-Conv) operation is introduced. The convolution kernel incorporates both horizontal and vertical dimensions with size $K \in \mathbb{R}^{k+H}$, as follows:

$$\begin{aligned} Y(h,l) =& Y_L(h,l) + Y_H(h,l) \\ =& \sum_{i=0}^{k-1} K_L(i) \cdot O_L(h, l+i-p) + \sum_{j=0}^{H-1} K_H(j) \cdot O_H(j,l) \end{aligned} \tag{5}$$

DSC-Conv operation enables the aggregation of features across all leads by integrating horizontal convolutions with vertical convolutions, while sliding only occurs along the horizontal dimension, which enhances sample-wise representative capability for intra-group leads.

3.3. Bidimensional Adaptive Residual Convolution (BARC)

Considering that the aggregate lead feature maps in DSC-Conv lack the information propagation between groups, this paper proposes a bidirectional adaptive residual convolution (BARC), as shown in Fig. 2. The BARC consists of two phases: in the first phase, temporal features of each ECG lead are extracted and aggregated, and in the second phase, the aggregated features are propagated across leads to incorporate comprehensive contextual information.

In the first phase, horizontal convolution with kernel $K \in \mathbb{R}^{1\times k}$ is employed to extract temporal features from the input, yielding an intermediate feature map that aggregates information for each lead. This process can be

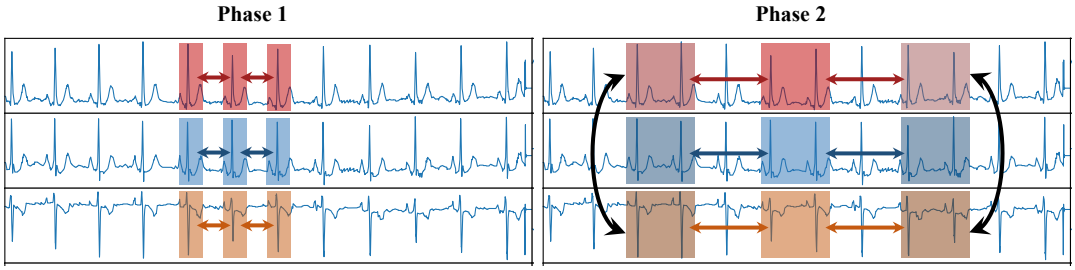


Fig. 2. The two phases of BARC, in which each line represents an individual lead. For phase 1, the horizontal convolution is employed for each lead. For phase 2, the DSC-Conv is applied to transmit relevant contextual information across leads.

regarded as downsampling in the temporal dimension. Subsequently, in the second phase, the DSC-Conv is applied to the intermediate feature map, which allows for transmitting relevant contextual information across leads and integrating it into the output via residual summation. Besides, an average pooling layer and an identity convolution layer are introduced to ensure dimensional compatibility in the residual path. In summary, the proposed DSTNN-X combines the DSC-Conv to extract intra-group features and BARC module to enhance inter-group correlations, achieving efficient multi-lead contextual aggregation and creating a rich, fused ECG representation.

3.4. Arrhythmia Classification

In this paper, a multi-head self-attention is used for channel feature fusion and a linear classification layer for arrhythmia label mapping. Specifically, given the feature maps $X \in \mathbb{R}^{C\times 12\times L'}$ from BARC, where L' denotes the downsampled horizontal dimension. The input time-frequency features are first processed from the feature maps. For each of the C channels, the $12 \times L'$ feature map is flattened, yielding a feature matrix $X_{\text{flat}} \in \mathbb{R}^{C\times D}$, where the feature dimension $D = 12L'$. To model the inter-channel dependencies, then employ the multi-head self-attention mechanism, as follows:

$$\text{head}_i = \text{softmax}\left(\frac{(X_{\text{flat}}W_i^Q)(X_{\text{flat}}W_i^K)^\top}{\sqrt{d_k'}}\right)(X_{\text{flat}}W_i^V) \tag{6}$$

$$X_{\text{att}} = \text{Concat}(\text{head}_1,\ldots,\text{head}_h)W^O \tag{7}$$

where $W_i^Q, W_i^K \in \mathbb{R}^{D\times d_k'}$ and $W_i^V \in \mathbb{R}^{D\times d_v'}$ are the learnable projection matrices for the i-th attention head. Here, h denotes the number of parallel heads, while d_k' and d_v' are the respective dimensions for the keys/queries and values. Finally, $W^O \in \mathbb{R}^{hd_v'\times D}$ is the output projection matrix that fuses the information from all heads to produce the final attention-aware feature representation $X_{\text{att}} \in \mathbb{R}^{C\times D}$.

Following the multi-head self-attention, adaptive max pooling is applied independently to each channel within X_{att}. For the c-th channel, its feature vector $x_{att,c} \in \mathbb{R}^{12\times L}$ is reduced to a single scalar value p_c:

$$p_c = \text{AdaptiveMaxPool}(x_{att,c}, \text{output\_size} = 1) \tag{8}$$

These values are then aggregated to construct a consolidated feature vector $P = [p_1, p_2, \ldots, p_C]^T \in \mathbb{R}^C$ to preserve the most salient patterns across channels. Finally, feature vector P is passed through a fully connected linear layer to map the arrhythmia classification labels N. The output logits $Y_{out} \in \mathbb{R}^N$ are computed as:

$$Y_{out} = W_{fc}P + b_{fc} \tag{9}$$

where $W_{fc} \in \mathbb{R}^{N\times C}$ is the weight matrix of the linear layer, $b_{fc} \in \mathbb{R}^N$ is the bias vector, and N represents the total number of target arrhythmia categories.

3.5. Learning

The proposed model is optimized using the cross entropy between the ground truth and predicted probability. The loss function l is defined as follows:

$$l = -\sum_{k=1}^{N} Y_{true,k}\log(\hat{P}_k) \tag{10}$$

where N denotes the total number of arrhythmia categories, k is the current category, the ground truth is $Y_{true,k} \in \{0,1\}^N$, $\hat{P}_k$ represents the model's confidence for the k-th arrhythmia category, obtained by applying the sigmoid activation function to its corresponding output.

Table 1. Quality performance of different methods on PTB-XL dataset with various annotation categories, including 'all', 'diag.', 'sub-diag.', 'super-diag.', 'form', and 'rhythm'.

Method	all		diag.		sub-diag.		super-diag.		form		rhythm	
	AUC	SEN	AUC	SEN	AUC	SEN	AUC	SEN	AUC	SEN	AUC	SEN
LSTM [4]	91.42	69.06	89.30	64.32	92.07	68.21	91.58	76.99	75.60	39.77	90.81	86.32
BiLSTM [5]	89.92	68.30	90.33	64.40	91.77	68.60	92.75	78.02	85.00	50.45	94.13	89.25
fcn_wang [6]	89.92	69.86	91.65	65.62	92.18	68.49	92.02	78.39	83.54	50.14	88.38	88.44
resnet1d_wang [6]	92.25	70.66	93.51	66.11	**93.52**	69.10	92.77	77.61	87.08	54.43	93.12	89.82
Xresnet1d101 [7]	91.69	69.74	92.97	66.31	92.05	66.61	91.03	75.91	84.86	51.87	95.52	91.28
MobileNetV3 [8]	90.08	68.41	89.64	62.31	89.15	66.02	90.84	76.03	82.37	49.43	95.62	91.02
ViT [9]	77.77	53.25	81.14	46.13	81.20	46.98	82.11	60.55	64.42	22.10	82.66	80.43
InceptionTime [10]	92.07	71.27	93.30	67.40	93.41	70.71	92.91	79.02	87.33	53.94	93.18	90.44
ATI-CNN [2]	90.09	70.68	91.28	67.22	91.23	70.12	91.91	78.58	85.65	**58.11**	95.60	91.25
MVMSNet [3]	92.35	71.53	**93.85**	66.76	**93.52**	69.30	92.97	79.03	88.67	53.98	94.98	90.77
DSTNN-X	**92.83**	**73.49**	93.42	**69.95**	93.24	**71.92**	**93.26**	**80.95**	**88.75**	57.88	**96.35**	**91.44**

4. EXPERIMENTS

In this section, extensive experiments are conducted to demonstrate the superior performance of the proposed method on the 12-lead ECG benchmark PTB-XL and CPSC2018 dataset. The experimental results show that the DSTNN-X obtains comparable performance with the ground truth and outperforms other baseline methods.

4.1. Setup

4.1.1. Datasets

PTB-XL [11] is a widely used comprehensive 12-lead ECG database. It contains 21,799 clinical 100 Hz ECG recordings of 10s segment from 18,869 patients, in which 52 % are male and 48 % are female, covering ages from 0 to 95 years old. The ECG recordings are annotated and validated by up to two cardiologists and conform to the SCP-ECG standard [17]. CPSC2018 [12] dataset contains 6,877 clinical 12-lead ECG recordings from 6 to 60 seconds, annotated with 9 diagnostic classes. All data are resampled to 100 Hz and standardized the length to 10s by cropping or padding. The dataset is split into 10 groups, with the 9th as validation, the 10th as test, and the rest for training.

4.1.2. Evaluation Metrics

In this paper, the area under the receiver operating characteristic curve (AUC), sensitivity (SEN), the area under the precision-recall curve (PRC), accuracy (ACC), and F1 score are used to evaluate the quality performance. Specifically, AUC measures the model's ability to distinguish between classes across different threshold settings. SEN evaluates the proportion of true positives that are correctly identified, reflecting the model's capability to detect relevant instances. PRC is particularly useful for imbalanced datasets, as it evaluates the trade-off between precision and recall. ACC provides an overall measure of the proportion of correctly classified instances. F1 score balances precision and recall, offering a harmonic mean that is especially informative when the class distribution is uneven. All metrics range from 0 to 1, with higher values representing better performance.

4.1.3. Implementation Details

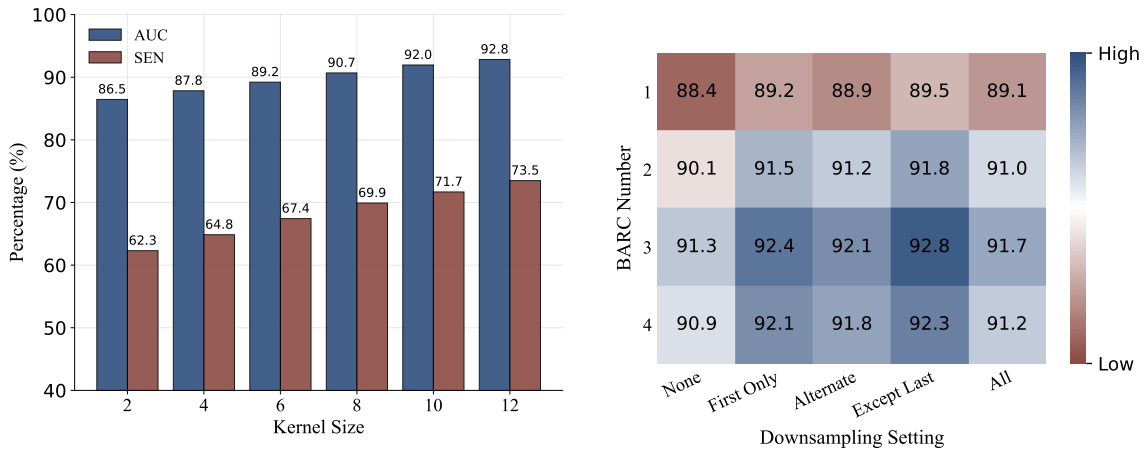
The proposed method is implemented in PyTorch and trained/tested on an NVIDIA Tesla V100 GPU (32GB Memory). For a fair comparison, the implementation follows the same experimental setup as MVMSNet [3]. The training parameters of all networks in the experiments are set the same: the batch size is set to 64, the fixed learning rate is set to 1e-3, the optimizer is the adaptive moment estimation (Adam) optimizer, and the total number of training epochs is 100.

4.2. Experimental Results

To verify the effectiveness of the method, this section compare DSTNN-X with other baseline methods such as LSTM [4], BiLSTM [5], fcn_wang [6], Xresnet1d101 [7], MobileNetV3 [8], ViT [9], InceptionTime [10], ATI-CNN [2], and MVMSNet [3], as shown in Table 1 and Table 2. From the table, it can be observed that DSTNN-X

Table 2. Quality performance of different methods on the CPSC2018 dataset

Method	AUC	SEN	PRC	ACC	F1
LSTM [4]	89.78	58.43	65.24	93.46	53.88
BiLSTM [5]	93.14	73.26	74.27	94.75	69.06
fcn_wang [6]	93.29	71.44	72.85	94.30	68.51
resnet1d_wang [6]	95.43	79.94	83.59	95.93	78.81
Xresnet1d101 [7]	95.85	80.89	83.19	96.03	79.41
MobileNetV3 [8]	95.44	80.23	81.67	95.41	76.55
ViT [9]	81.35	36.34	43.61	90.41	30.03
InceptionTime [10]	95.95	82.12	84.17	96.01	80.03
ATI-CNN [2]	95.43	82.85	80.90	95.87	78.63
MVMSNet [3]	95.82	82.34	84.01	96.08	79.89
DSTNN-X	**96.65**	**84.52**	**85.10**	**96.20**	**80.96**



(a) The kernel size of DSC-Conv

(b) Settings of BARC

Fig. 3. Ablation study about the kernel size of DSC-Conv and different settings of BARC on the PTB-XL dataset.

outperforms all baseline methods for arrhythmia classification performance, as expected. Specifically, for the PTB-XL dataset, the proposed method achieves the highest AUC (92.83%) and SEN (73.49%) on average, while for the CPSC2018 dataset, it delivers outstanding AUC (96.65%), SEN (84.52%), PRC (85.10%), ACC (96.20%) and F1 (80.96%). This is attributed to the grouping of 12-lead ECG data and the design of DSC-Conv for extracting intra-group features and BARC module for enhancing inter-group correlations, respectively, which effectively captures the multi-lead contextual information. Overall, the experimental results demonstrate the effectiveness and practicality of the proposed method.

4.3. Ablation Studies

4.3.1. Kernel Size of DSC-Conv

As shown in Fig. 3(a), this section conducts an ablation study about the kernel size of DSC-Conv on the PTB-XL dataset. From the figure, it can be observed that the model's classification performance gradually improves as the kernel size increases, where the AUC increases from 86.5% to 92.8% and the SEN increases from 62.3% to 73.5%. The reason is that a larger kernel allows DSC-Conv to capture features across more leads, effectively modeling the inter-lead relationships relevant to the disease. Experimental results show that when the kernel size is set to 12, the model achieves the best classification performance.

4.3.2. Settings of BARC

As shown in Fig. 3(b), the quality performance of different settings (BARC number and downsampling setting) of the BARC are evaluated on the PTB-XL dataset, in which "None" means that no downsampling is performed,

"First Only" indicates that downsampling is applied exclusively to the first layer, "Alternate" represents that downsampling is applied only on odd layers, "Except Last" indicates that downsampling is applied to all layers except the final one, and "All" means that downsampling is applied to all layers. From the figure, it can be seen that the classification performance of the proposed DSTNN-X increases as the BARC number rises, due to the enhanced inter-group feature extraction capability with a deeper architecture. Another observation is that the downsampling in the BARC improves performance by broadening the contextual view for leads to identify arrhythmia changes, while the particular strategies exhibit limited effects on the final results if a sufficient extent of feature condensation and contextual scaling is achieved.

5. CONCLUSIONS

This paper proposes a novel neural network architecture with multi-lead contextual aggregation called DSTNN-X for ECG arrhythmia classification. DSTNN-X combines the DSC-Conv to extract intra-group features and the BARC module to enhance inter-group correlations, which creates a rich, fused 12-lead ECG representation. Experimental results on both open source PTB-XL and CPSC2018 datasets show that the proposed method outperforms other baseline methods, efficiently improving arrhythmia classification performance. In the future, the application will be explored in other physiological signal processing tasks and further improve clinical interpretability.

References

1. S. S. Martin, A. W. Aday, N. B. Allen, Z. I. Almarzooq, C. A. Anderson, P. Arora, C. L. Avery, C. M. Baker-Smith, N. Bansal, A. Z. Beaton, *et al.*, "2025 heart disease and stroke statistics: A report of us and global data from the american heart association," *Circulation*, 2025.
2. Q. Yao, R. Wang, X. Fan, J. Liu, and Y. Li, "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.
3. S. Yang, C. Lian, Z. Zeng, B. Xu, J. Zang, and Z. Zhang, "A multi-view multi-scale neural network for multi-label ecg classification," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 7, no. 3, pp. 648–660, 2023.
4. S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
5. S. Zhang, D. Zheng, X. Hu, and M. Yang, "Bidirectional long short-term memory networks for relation classification," in *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation* (H. Zhao, ed.), (Shanghai, China), pp. 73–78, PACLIC, October 2015. SemEval-2010 Task-8.
6. Z. Wang, W. Yan, and T. Oates, "Time series classification from scratch with deep neural networks: A strong baseline," in *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 1578–1585, 2017.
7. T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of tricks for image classification with convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 558–567, 2019.
8. A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, and Q. Le, "Searching for mobilenetv3," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1314–1324, 2019.
9. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
10. H. Ismail Fawaz, B. Lucas, G. Forestier, C. Pelletier, D. F. Schmidt, J. Weber, G. I. Webb, L. Idoumghar, P.-A. Muller, and F. Petitjean, "Inceptiontime: Finding alexnet for time series classification," *Data Mining and Knowledge Discovery*, vol. 34, no. 6, pp. 1936–1962, 2020.
11. P. Wagner, N. Strodthoff, R.-D. Bousseljot, D. Kreiseler, F. I. Lunze, W. Samek, and T. Schaeffter, "Ptb-xl, a large publicly available electrocardiography dataset," *Scientific data*, vol. 7, no. 1, pp. 1–15, 2020.
12. F. Liu, C. Liu, L. Zhao, X. Zhang, X. Wu, X. Xu, Y. Liu, C. Ma, S. Wei, Z. He, *et al.*, "An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection," *Journal of Medical Imaging and Health Informatics*, vol. 8, no. 7, pp. 1368–1373, 2018.
13. S. Karimifard, A. Ahmadian, M. Khoshnevisan, and M. S. Nambakhsh, "Morphological heart arrhythmia detection using hermitian basis functions and knn classifier," in *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 1367–1370, IEEE, 2006.
14. T.-H. Pham, V. Sree, J. Mapes, S. Dua, O. S. Lih, J. E. Koh, E. J. Ciaccio, and U. R. Acharya, "A novel machine learning framework for automated detection of arrhythmias in ecg segments," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–18, 2021.
15. Ö. Yıldırım, P. Pławiak, R.-S. Tan, and U. R. Acharya, "Arrhythmia detection using deep convolutional neural network with long duration ecg signals," *Computers in biology and medicine*, vol. 102, pp. 411–420, 2018.
16. 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, 2019.
17. P. Rubel, J. Fayn, P. W. Macfarlane, D. Pani, A. Schlögl, and A. Värri, "The history and challenges of scp-ecg: The standard communication protocol for computer-assisted electrocardiography," *Hearts*, vol. 2, no. 3, pp. 384–409, 2021.