HS-Vectors: Heart Sound Embeddings for Abnormal Heart Sound Detection Based on Time-Compressed and Frequency-Expanded TDNN With Dynamic Mask Encoder

Lihong Qiao, Yonghao Gao, Bin Xiao, Xiuli Bi, Weisheng Li, and Xinbo Gao

***Abstract*—In recent years, auxiliary diagnosis technology for cardiovascular disease based on abnormal heart sound detection has become a research hotspot. Heart sound signals are promising in the preliminary diagnosis of cardiovascular diseases. Previous studies have focused on capturing the local characteristics of heart sounds. In this paper, we investigate a method for mapping heart sound signals with complex patterns to fixed-length feature embedding called HS-Vectors for abnormal heart sound detection. To get the full embedding of the complex heart sound, HS-Vectors are obtained through the Time-Compressed and Frequency-Expanded Time-Delay Neural Network(TCFE-TDNN) and the Dynamic Masked-Attention (DMA) module. HS-Vectors extract and utilize the global and critical heart sound characteristics by masking out irrevelant information. Based on the TCFE-TDNN module, the heart sound signal within a certain time is projected into fixed-length embedding. Then, with a learnable mask attention matrix, DMA stats pooling aggregates multi-scale hidden features from different TCFE-TDNN layers and masks out irrelevant frame-level features. Experimental evaluations are performed on a 10-fold cross-validation task using the 2016 PhysioNet/CinC Challenge dataset and the new publicly available pediatric heart sound dataset we collected. Experimental results demonstrate that the proposed method excels the state-of-the-art models in abnormality detection.**

***Index Terms*—Abnormal heart sound detection, HS-vectors, Time-Compressed and Frequency-Expanded Time-Delay Neural Network, Dynamic Masked-Attention statistics pooling.**

Manuscript received 20 June 2022; revised 9 October 2022 and 22 November 2022; accepted 2 December 2022. Date of publication 8 December 2022; date of current version 7 March 2023. This work was supported in part by the National Key Research and Development Project under Grant 2019YFE0110800, in part by the National Natural Science Foundation of China under Grants 62276040 and 61976031, in part by the National Major Scientific Research Instrument Development Project of China under Grant 62027827, and in part by the Scientific and Technological Key Research Program of Chongqing Municipal Education Commission under Grant KJQN202200624. *(Corresponding author: Bin Xiao.)*

The authors are with the Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail: qiaolh@cqupt.edu.cn; s200231258@stu.cqupt.edu.cn; xiaobin@cqupt.edu.cn; bixl@cqupt.edu.cn; liws@cqupt.edu.cn; gaoxb@cqupt.edu.cn).

Digital Object Identifier 10.1109/JBHI.2022.3227585

I. Introduction

CARDIOVASCULAR diseases (CVDs) are the leading cause of death globally. According to statistics from the World Health Organization, the number of deaths due to CVDs represents 32% of all global deaths in 2019 [1]. Therefore, it becomes significant to investigate an effective CVDs detection method. As a non-invasive but also cost-effective procedure, cardiac auscultation is used for the early diagnosis of various heart diseases. However, effective cardiac auscultation depends on trained cardiologists, a resource that is insufficient particularly in low-income countries of the world [2]. Moreover, the accuracy of auscultation depends on the proficient clinical skills and extensive subjective experiences of the physicians. Therefore, to improve efficiency, a computer-based method for heart sound diagnosis highlights the necessity of assisting physicians in diagnosis.

In the past few decades, the traditional methods of Phonocardiogram (PCG) abnormality detection have been greatly extended. The process of PCG classification generally consists of three main steps: preprocessing, feature extraction, and classification. First, preprocessing performs operations such as filtering and segmentation. Filtering denoises the signal, and segmentation divides the PCG signal into four parts: S1, systole, S2 and diastole. Then, the physiological and pathological information about the heart is extracted [3], [4], [5], [6], [7], [8]. In the last step, a classifier is used to classify the PCG [9].

However, the pattern of heart sound signals is complex due to different acquisition devices and environments. For traditional methods, the feature extractor mainly relies on artificial design and requires professional knowledge. Moreover, the features extracted by the designed function are often plain in pattern, which is not enough to characterize the complex pattern of the heart sound signals. These methods are usually aimed at a specific task and have poor generalization and robustness. In contrast, the DNN models are better at capturing the underlying relations in the PCG signal. The CNN-based methods [10], [11] achieve better results.

In this paper, we investigate a method for mapping heart sound signals with complex patterns to fixed-length global feature embeddings called HS-Vectors as a model for heart sound abnormality detection based on embedding representation.

The proposed HS-Vectors utilize Dynamic Masked-Attention (DMA) statistics pooling on the TCFE-TDNN to obtain the global heart sound characteristics by suppressing uninformative regions. First, the TCFE-TDNN module is proposed to improve the frequency resolution and compress the features in the time domain so that the network can observe the potential abnormal performance of each frame in the low-dimensional and high-dimensional frequency hidden feature space. Then, the DMA module masks out useless information and aggregates multi-scale hidden features from different TCFE-TDNN layers. HS-vectors aggregates global time-frequency feathers with good representation ability. Besides, HS-vectors does not employ segmentation in the frameworks, as the potentially incorrect segmentation of the input PCG signal may lead to poor performance. To our knowledge, this is the first time that a global vector integration method has been proposed to detect abnormal heart sounds.

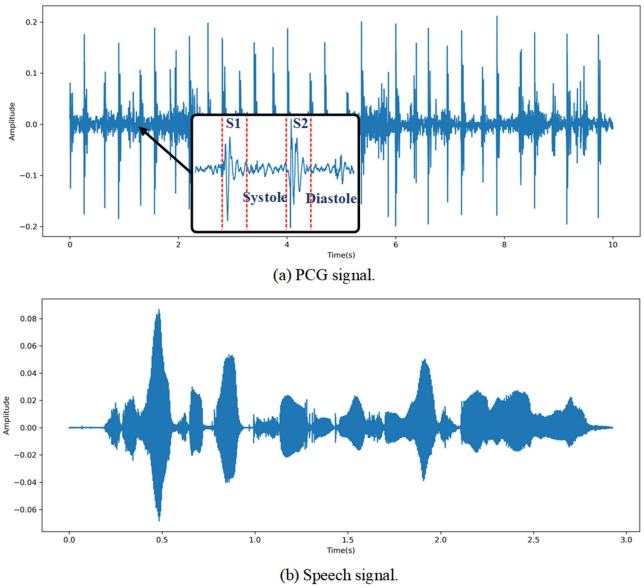
The main contributions of this paper are as follows:

- The proposed HS-Vectors consider the embedding of heart sound in Time-Compressed and Frequency-Expanded TDNN (TCFE-TDNN) module to project the heart sound signal within a certain time into a fixed-length signal characterizing embedding and enhance the feature discrimination.
- The proposed HS-Vectors apply Dynamic Masked-Attention statistics pooling on the TCFE-TDNN to obtain the stable key heart sound characteristics by suppressing uninformative regions.
- The proposed HS-Vectors have the complementary of mapping heart sound signals with complex patterns to fixed-length feature embedding, which improves the network performance according to global properties of the heart sound.

The rest of this paper is structured as follows. In Section II, we introduced a brief review of related works. Then, the proposed HS-Vectors for detection of abnormal heart sounds are introduced in detail in Section III. Section IV contains the implementation details of our evaluation experiment, results of our methods, and comparisons with several state-of-the-art PCG classification methods. Finally, we conclude this paper in Section V.

II. Related Work

In the past few decades, the traditional methods of Phonocardiogram (PCG) abnormality detection have been greatly extended. The process of PCG classification generally consists of three main steps: preprocessing, feature extraction, and classification. First, preprocessing performs operations such as filtering and segmentation. Filtering is denoising the signal, and segmentation is dividing the PCG signal into four parts: S1, systole, S2 and diastole. For this step, Francesco Renna et al. [12] and Omer Deperlioglu et al. [13] respectively used a second-order 25 Hz–400 Hz Butterworth filter and elliptical digital filter to denoise the heart sound signal. Sun et al. [14] proposed an automatic heart sound segmentation method based on the Hilbert transform. The hidden semi-Markov model (HSMM)



(a) PCG signal.

(b) Speech signal.

Fig. 1. The waveforms of heart sound and speech. (a) PCG signal. (b) Speech signal (continuous English speech). The speaker information is spread throughout the speech signal. Similarly, the potential pathological information is spread throughout the heart sound signal.

method [15] was extended with logistic regression to achieve signal segmentation in a noisy environment. Furthermore, there are some segmentation methods based on deep learning [16], [17], [18]. In the second step, the physiological and pathological information about the heart is extracted. There are a variety of heart sound feature extraction methods, which involve features that mainly include temporal features [3], [4], [19], frequency features [6], [20], wavelet transform [5], [21], [22] and MFCCs features [7], [8]. MFCCs have been shown to be an effective individual identification feature [23]. Therefore, MFCCs are used as features for abnormal heart sound detection in this paper. In the last step, a classifier is used to classify the PCG. Classifiers based on machine learning like Support Vector Machines (SVM) [9], k-Nearest Neighbor(k-NN) [24], and Artificial Neural Network(ANN) [25] were employed frequently.

Recently, Deep Learning Neural Network (DNN) has been used for abnormal heart sound detection due to its powerful feature representation ability. In the case of temporal or frequency features, Thomae et al. [26] built an end-to-end deep neural network that directly extracts hidden features in the temporal domain. Besides, Ryu et al. [10] constructed a CNN model for classifying segmented PCGs. Recently, Humayun et al. [11] designed a time-convolutional (tConv) unit to learn hidden features from temporal features. Devjyoti Chakraborty et al. [6] trained a 2D-CNN with the spectrum of the cardiac cycle. However, it is difficult for a single feature to represent complex heart sound patterns, which limits the classification capabilities of these methods.

To improve the classification performance, some models are constructed based on the time-frequency features of heart sounds and classify the segmented PCG. For instance, Rubin et al. [27]

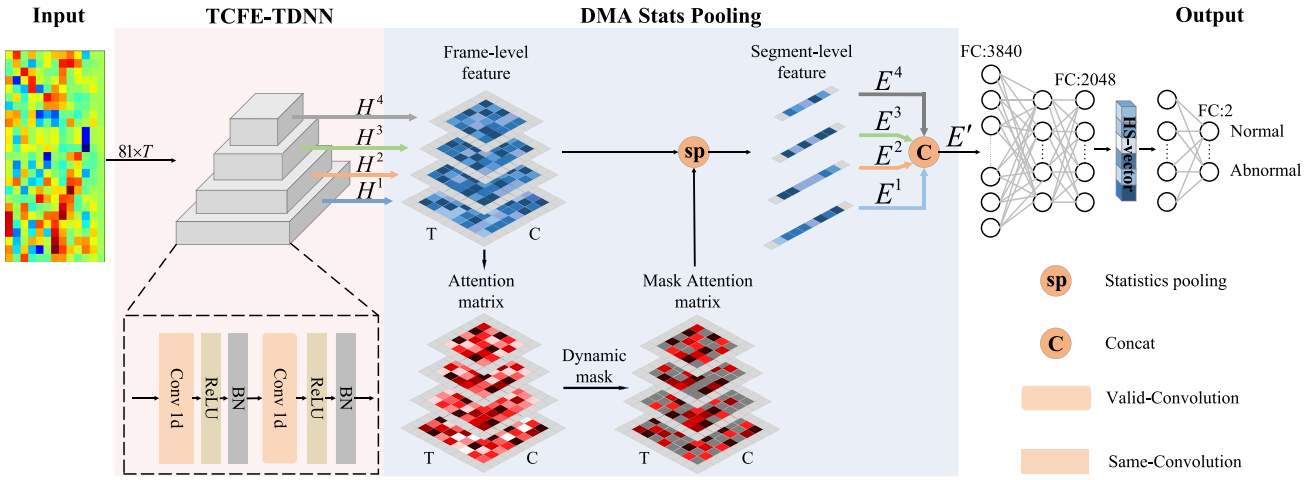


Fig. 2. Illustration of the network structure in our proposed HS-Vectors. Our architecture takes heart sound MFCCs as input and uses TCFE-TDNN to extract frame-level features at different resolutions in the low-dimensional and high-dimensional frequency hidden feature space. Then, DMA stats pooling is used to aggregate frame-level features into segment-level features, and a fully connected network is utilized to encode these features into a fixed-dimensional vector representation (called: HS-vector). Finally, the HS-vector is fed into a fully connected network with dropout to classify PCG.

extracted MFCCs features from the segmented PCG and fed them to the CNN for abnormal heart sound detection. Potes et al. [28] proposed a model integrating AdaBoost-abstain and CNN. Recently, Deng et al. [29] constructed CRNN, which combines CNN and RNN to improve feature extraction capabilities. In recent studies, some non-segmented methods have been proposed in abnormal heart sound detection [23], [30], [31]. However, a large number of time-frequency models mainly learn the local features of heart sounds, and there are few studies on the global feature representation of heart sounds. The challenge in abnormal heart sound detection is to estimate the global characteristics of heart sounds. This problem is similar to the speaker verification (SV) task. The waveforms of heart sound and speech are shown in Fig. 1. The speaker information is assumed to be spread throughout the speech signal, and an embedded representation of the entire speech is used in SV. Among them, X-vectors is a DNN model with superior performance, and its variant models have been widely used in SV [32], [33], [34]. Although the CNN-based heart sound methods obtained acceptable results, most of them either blindly increase the filters and the number of convolution layers to pursue a slight performance improvement at the cost of increasing the computation burden. Besides, most of them didn't fully consider the characteristics of heart sound in designing the network structure, which also limits the improvement of classification performance.

III. Proposed Method

For abnormal heart sound detection, how to holistically represent the periodic changes of long-term signals is the key problem. In previous studies, the feature extractor was mainly based on convolutional networks to capture local hidden relationships of heart sounds, and a few RNN-based models only incorporated contextual information into local features. Inspired by X-vectors, which is a Time-Delay Neural Network (TDNN) that applies statistics pooling to project variable-length utterances into fixed-length speaker characterization embeddings, we define HS-vector as the feature embedding of the heart sound signal within a certain time and design the HS-Vectors for abnormal heart sound detection. As shown in Fig. 2.

First, the one-dimensionally transformed MFCC features are fed to the Time-compressed and Frequency-expanded TDNN (TCFE-TDNN) module, which adopts a variable-scale frequency hidden feature extraction strategy to diversify the frame-level features so that the network can observe the potential abnormal performance of each frame in the low-dimensional and high-dimensional frequency hidden feature space. TCFE-TDNN focuses on extracting frame-level features with different time-frequency resolutions, but these features only reflect local variations of the heart sound signal. Therefore, this study combines TCFE-TDNN with statistical pooling to achieve global heart sound signal representation by dynamic masked-attention statistical pooling (DMA stats pooling). This module is proposed to aggregate multi-scale frame-level features from different TCFE-TDNN layers through adaptive statistical methods so that the output of the segment-level feature by the module contains the time-frequency feature details of heart sounds. The two-layer fully connected (FC) network followed by DMA stats pooling encodes the fragment-level features as a HS-vector. Finally, a fully connected network with dropout is used to perform the heart sound classification task. This architecture enables the model to aggregate different features to obtain a HS-vector with good representation ability and mask out the irrelevant frames. Besides, our method does not employ segmentation due to the potentially incorrect segmentation of the input PCG signal. To our knowledge, this is the first time a whole vector embedding method has been applied to the task of abnormal heart sound detection.

The proposed HS-vector could be derived by the TCFE-TDNN module and DMA stats pooling, respectively. Thus, we first implemented the TCFE-TDNN to get the low-dimensional

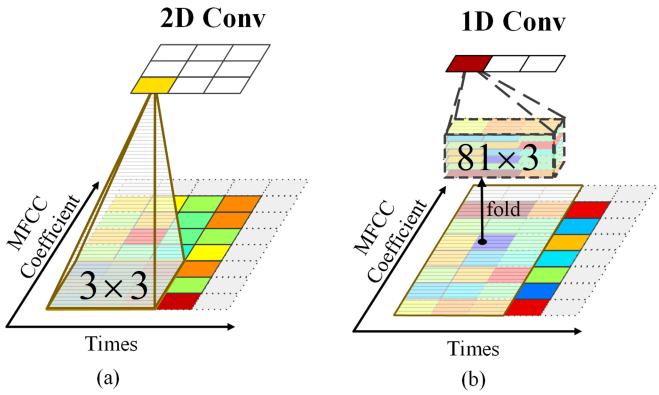


Fig. 3. The receptive fields of different convolutions methods of the heart sound MFCCs. (a) Receptive field of 2D convolution. (b) Receptive field of 1D convolution.

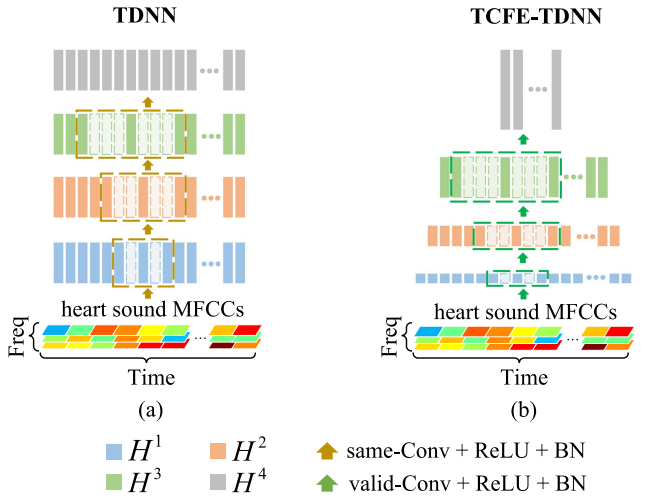


Fig. 4. Schematic diagram of frequency features variation in two different network structures. (a) The hidden features variation in TDNN. (b) The hidden features variation in TCFE-TDNN.

and high-dimensional frequency features of PCG signals in each frame.

A. Time-Compressed and Frequency-Expanded TDNN Module

It is a difficult problem to get the potential abnormal performance of each frame in the low-dimensional and high-dimensional frequency hidden feature space. We introduce a mechanism that can automatically increase the frequency resolution and compress the features in the time domain of the PCG. To derive HS-Vectors, we introduce the Time-Compressed and Frequency-Expanded TDNN (TCFE-TDNN) module, which doubles the frequency features by increasing the number of convolution filters and compresses temporal features because of valid-convolution layer by layer. The heart sound MFCCs are regarded as a one-dimensional time series of multi-frequency channels and perform frame-level hidden feature extraction on it using a valid convolution-based network.

1) One-Dimensional Transformation of MFCC: First, we treat the heart sound MFCCs as a one-dimensional time series of multi-frequency channels and extract their hidden features through 1D convolution with a larger frequency receptive field. In recent studies [23], [29], the heart sound MFCCs was used as the main heart sound feature research objects due to its excellent representation ability. Traditionally, the heart sound MFCCs are regarded as a 2D image and extract hidden features through 2D convolution, which causes the model to capture features in a non-existent spatial domain, as shown in Fig. 3(a). In SV, people treat the MFCCs as a one-dimensional time series of multi-frequency channels and extract its hidden features through 1D convolution with large kernel attention, as shown in Fig. 3(b). It can be known that 1D convolution has a larger frequency receptive field, which builds up large receptive fields. We treat the heart sound MFCCs as a one-dimensional time series of multi-frequency channels and extract its hidden features. The strategy based on 1D convolution makes it easier for the model to observe the details of the PCG signal over time.

2) The Structure of TCFE-TDNN Module: TDNN is a multi-layer one-dimensional convolutional neural network structure, which expands the receptive field of the convolution kernel by increasing the dilation rate layer by layer, as shown in Fig. 4(a). Each TDNN layer performs feature extraction through a fixed number of convolution filters to obtain a frame-level feature with a fixed dimension. We define the outputs of layers 3 and 4 as high-level features and the outputs of layers 1 and 2 as low-level features. In the extensive research of CNNs [35], the output of shallow layers focuses on the representation of details and the output of deep layers has more complex global information. Frequency is one of the important information reflecting the characteristics of heart sounds. In TDNN, the high-level features are restricted to be represented in the same scale space as the low-level features, which is not conducive to the complete representation of complex features and makes it difficult to observe more potential frequency details from the high-level hidden features.

To solve the above problems, we proposed the Time-Compressed and Frequency-Expanded TDNN (TCFE-TDNN) module as shown in Fig. 4(b). To obtain more details of potential features, the TCFE-TDNN doubles the frequency features by increasing the number of convolution filters layer by layer and compresses temporal features, which adopts a variable hidden feature extraction strategy to diversify the hidden features. In TCFE-TDNN, the frequency dimension of the heart sound features is used as the channel dimension of the input features. The 1D convolution filter uses all the frequency elements of the temporal feature frame in the receptive field to obtain a linear representation of the frequency features. This linear representation is an implicit frequency feature in the heart sound signal. We increase the number of filters layer by layer so that the network can obtain more details of the implicit frequency features layer by layer. This process can be considered an extension of the heart sound frequency features. The variable hidden feature extraction strategy uses valid convolution to reduce the time domain dimension and increase the frequency dimension of the heart sound features layer by layer. The feature variations in

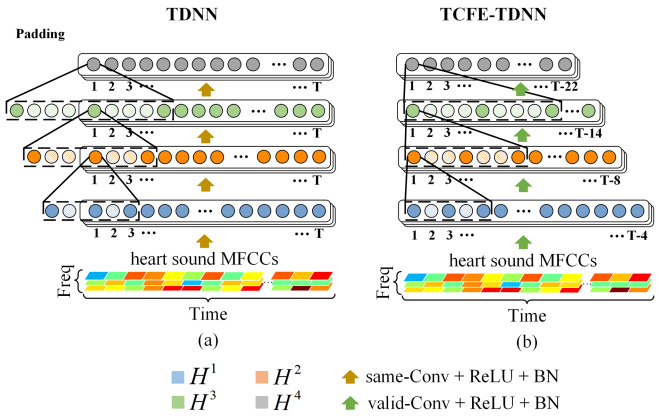


Fig. 5. Schematic diagram of temporal features variation in two different network structures. (a) The hidden features variation in TDNN. (b) The hidden features variation in TCFE-TDNN.

the time domain are shown in Fig. 5. Compared with TDNN, TDFE-TDNN uses valid convolution without padding operation, which makes the temporal dimension of features decrease layer by layer. The temporal features are compressed by doing these. In addition, the elements of TCFE-TDNN output features are obtained from the real input features of the same receptive field due to valid convolution without padding operation. The temporal feature frames from the same layer of TCFE-TDNN have the same real receptive field. In contrast, the output features from different layers contain different numbers of temporal feature frames and different frequency features, which reflects the diversity of hidden features.

In detail, we denote the input of model as $X \in \mathbb{R}^{C_{freq} \times T}$, where C_{freq}, T denote the dimension of frequency and time axes of heart sound MFCCs, respectively. The output of i-th TCFE-TDNN layer are defined as $H^i \in \mathbb{R}^{C_i \times T_i}$ for $i = 1, 2, 3, 4$ and $H^i = \{h_1^i, h_2^i, \ldots, h_{T_i}^i\}$ where $h_t^i \in \mathbb{R}^{1 \times C_i}$ is the output frame-level feature at time captured by i-th TCFE-TDNN layer. The output of each TCFE-TDNN layer can be expressed as:

$$H^i = \begin{cases} BN(\mathrm{ReLU}(\hat{F}_i(X))), i = 1 \\ BN(\mathrm{ReLU}(\hat{F}_i(H^{i-1}))), i = 2, 3, 4 \end{cases}, \quad (1)$$

where BN and $ReLU$ are batch normalization and ReLU function. $\hat{F}_i$ denote the convolution of i-th TCFE-TDNN layer. The first TCFE-TDNN layer (TCFE-TDNN-1) contains a valid-convolution with kernel size of 5. TCFE-TDNN-2, TCFE-TDNN-3 and TCFE-TDNN-4 contains the valid convolutions with kernel size 3, and their dilation rates are 2, 3 and 4, respectively. The C_i satisfies $C_4 = 2C_3 = 4C_2 = 8C_1 = 1024$. It should be noted that each TCFE-TDNN layer also contains a convolution with kernel size of 1, which precedes the valid convolution and does not change the feature dimension. Its structure can be found in Fig. 2.

With different dilation rates and valid convolution, we construct the TCFE-TDNN. This module extracts frame-level hidden features of the heart sound MFCCs with different time-frequency resolutions layer by layer, so that the network can observe the potential abnormal performance of each frame in the low-dimensional and high-dimensional frequency hidden feature space. This conversion mechanism from temporal features to frequency features of TCFE-TDNN enriches the representation of frame-level features. These features are used to generate HS-vector, which improves the frequency resolution even more and focuses the features in the time domain.

B. Dynamic Masked-Attention Statistics Pooling

Statistics pooling is a global pooling structure that implements aggregation by calculating the mean and standard deviation of the features on each channel dimension. In time series related tasks, using statistics pooling instead of global average pooling can better capture the long-term temporal characteristics. Besides, given the hierarchical structure of TCFE-TDNN, the features at each level exhibit different details, which are closely related to heart sound classification. Thus, we constructed a Dynamic Masked-Attention statistics pooling (DMA Stats Pooling), which dynamically masks out irrelevant frame-level features by applying masked-attention statistics pooling (MA stats pooling) to different TCFE-TDNN layers so that HS-vector emphasizes the representation of important regions. This module enables the model to aggregate different resolution features to obtain a HS-vector with good representation ability. The Masked-Attention Statistics Pooling is shown in Fig. 6.

We denote the output of a TCFE-TDNN layer as $H \in \mathbb{R}^{C \times \mathrm{T}}$ and $H = \{h_1, h_2, \ldots, h_T\}$ where $h_t \in \mathbb{R}^{C \times 1}$ is the output feature frame at time t captured by hidden layer. T, C denote the dimension of time and channel axes of the frame-level features, respectively. We first obtain an attention matrix. The channel-dependent attention [33] is defined as follows:

$$B = \upsilon^T f(WH + bo) + ko, \quad (2)$$

where the parameters $W \in \mathbb{R}^{(C/r) \times C}$ and $b \in \mathbb{R}^{(C/r) \times 1}$ compress the attention information by r times. This combines fuse the features of each channel and reduces the amount of computation. All elements of the matrix $o \in \mathbb{N}^{1 \times T}$ are 1. $f(\cdot)$ is an activation function and *Tanh* function is used here. The weights $\upsilon \in \mathbb{R}^{(C/r) \times C}$ and the bias $k \in \mathbb{R}^{C \times 1}$ are the parameters of the 1D-Conv layer. $B \in \mathbb{R}^{C \times T}$ is the output of the bottleneck layer and the scalar $e_{t,c}$ in B is expressed as:

$$e_{t,c} = \upsilon_c^T f(Wh_t + b) + k_c, \quad (3)$$

where $\upsilon_c \in \mathbb{R}^{(C/r) \times 1}$ denotes the c-th weight vector in υ and k_c denotes the c-th scalar in k. The $e_{t,c}$ represents the scalar score of the t-th feature frame on the c-th channel. Then, the *softmax* function is applied along the time axis of the B:

$$\alpha_{t,c} = \frac{\exp(e_{t,c})}{\sum_{\tau}^{T} \exp(e_{\tau,c})}, \quad (4)$$

where $\alpha_{t,c}$ represents the importance of the t-th feature frame on the c-th channel. The attention matrix $A \in \mathbb{R}^{T \times C}$ is defined

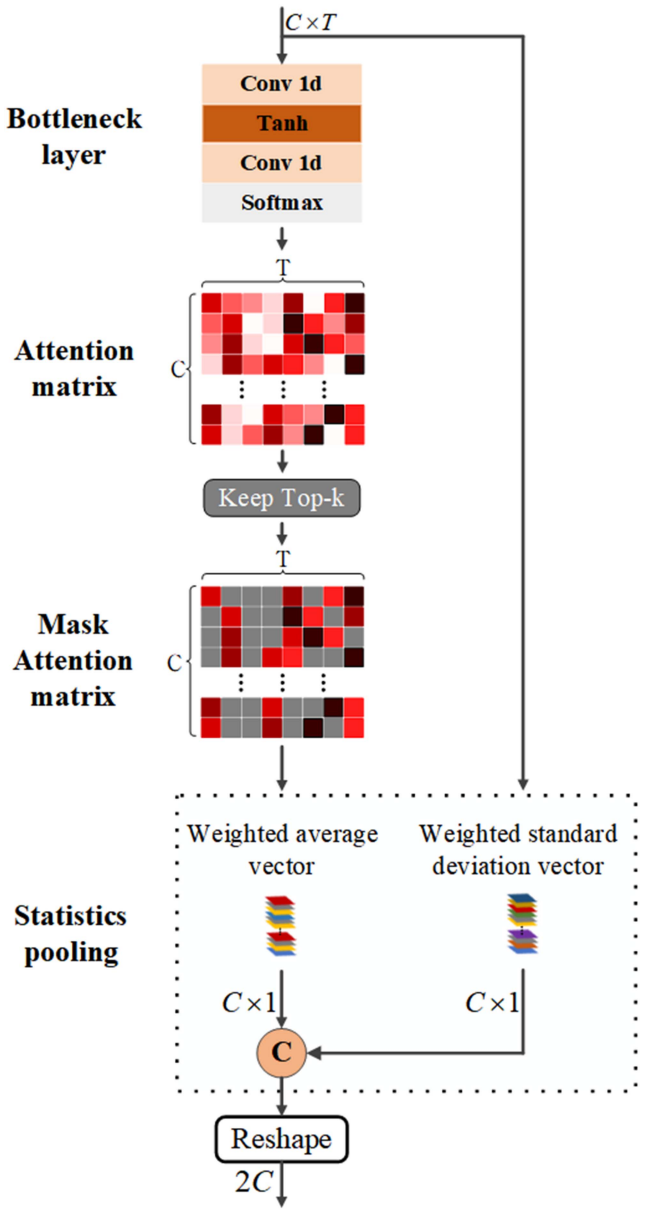


Fig. 6. Masked-Attention Statistics Pooling.

as follows:

$$A = \begin{bmatrix} \alpha_{1,1} & \alpha_{1,2} & \alpha_{1,3} & \cdots & \alpha_{1,C} \\ \alpha_{2,1} & \ddots & \vdots & \vdots & \alpha_{2,C} \\ \alpha_{3,1} & \cdots & \ddots & \vdots & \alpha_{3,C} \\ \vdots & \cdots & \cdots & \ddots & \vdots \\ \alpha_{T,1} & \alpha_{T,2} & \alpha_{T,3} & \cdots & \alpha_{T,C} \end{bmatrix}. \tag{5}$$

The potential pathological features are an essential diagnostic basis for the heart sound signal. In our proposed DMA module, the learnable attention matrix A represents the level of attention paid to the features by the model. The features with high attention weights positively affect the diagnostic outcome, and the features with low attention weights negatively affect it, which is verified in the experimental section. Therefore, our masked-attention masks out irrelevant features. First, the mask matrix M is defined as follows:

$$M = \begin{bmatrix} m_{1,1} & m_{1,2} & m_{1,3} & \cdots & m_{1,C} \\ m_{2,1} & \ddots & \vdots & \vdots & m_{2,C} \\ m_{3,1} & \cdots & \ddots & \vdots & m_{3,C} \\ \vdots & \cdots & \cdots & \ddots & \vdots \\ m_{T,1} & m_{T,2} & m_{T,3} & \cdots & m_{T,C} \end{bmatrix}, \tag{6}$$

where

$$m_{t,c} = \begin{cases} 1 & a_{t,c} \in top_k\{a_{\tau,c} | \tau = 1, 2, \ldots, T\} \\ 0 & else \end{cases}. \tag{7}$$

here, $top_k\{\cdot\}$ means the top k numbers in the descending list, which is used to obtain features with high attention weights. The mask matrix $M \in \{0,1\}^{T\times C}$ is a binarized output. Then, we create an attention matrix using mask $A' \in \mathbb{R}^{T\times C}$ is expressed as:

$$A' = M \otimes A, \tag{8}$$

where $\otimes$ denotes the element-wise multiplication. A' represents the attention matrix that only retains the top k weight scores on each channel dimension, which emphasizes the attention to important regions. It is worth noting that $A' = A$ when $k = T$. For convenience, we define *mask rate* v as the ratio of the removed feature number divide the original feature number, that is $v = \frac{T-k}{T}$.

Statistics pooling is used to create a uniform feature embedding representation for all frames, which consists of various statistics computed across frames for each channel. To enhance the feature embedding representation of heart sounds, the proposed MA stats pooling suggests applying the mask attention matrix A' to the statistics calculation. Specifically, statistics pooling is utilized to obtain the weighted mean vector $\tilde{\mu} \in \mathbb{R}^{C\times 1}$ and weighted standard deviation vector $\tilde{\sigma} \in \mathbb{R}^{C\times 1}$ along the time axis. The weighted mean $\tilde{\mu}_c$ and standard deviation $\tilde{\sigma}_c$ of their corresponding each channel are expressed as:

$$\tilde{\mu}_c = \sum_t^T \alpha'_{t,c} h_{t,c}, \tag{9}$$

$$\tilde{\sigma}_c = \sqrt{\sum_t^T \alpha'_{t,c} h_{t,c}^2 - \tilde{\mu}_c^2}. \tag{10}$$

P is the result of connecting $\tilde{\mu}$ and $\tilde{\sigma}$ along the channel axis. We perform a flatten operation to reshape P into a vector E:

$$P \in \mathbb{R}^{2C\times 1} \rightarrow E \in \mathbb{R}^d, \tag{11}$$

where $d = 2 \cdot C$. E represents the segment-level features of heart sounds, which is the output of MA stats pooling aggregating frame-level features H along the time axis.

To take full advantage of the multi-scale frame-level features extracted by TCFE-TDNN, a dynamic masked-attention statistics pooling(DMA stats pooling) is proposed, which constructs MA stats pooling for each frame-level feature and concatenates all the segment-level features as output. As shown in Fig. 7.

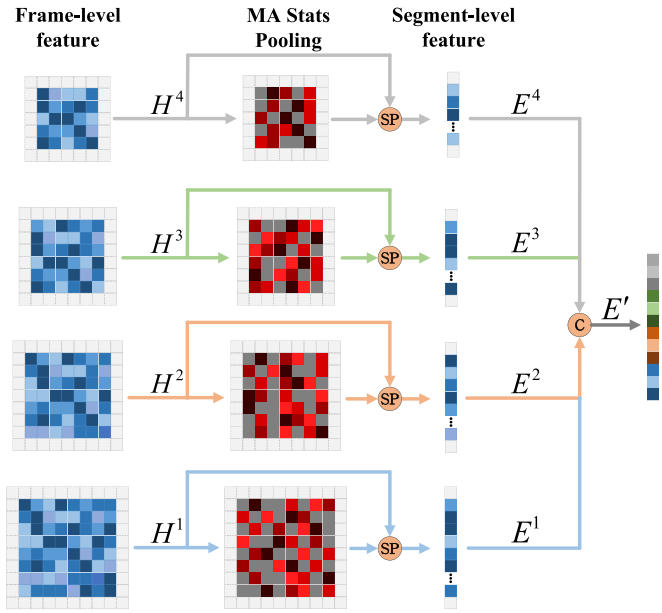


Fig. 7. Dynamic Masked-Attention Statistics Pooling.

Since the four frame-level features of a heart sound sample have different representations, DMA stats pooling can dynamically identify the different features of each layer and mask out the irrelevant features. It learns dynamic mask functions to match the heart sound embedding in the same period at a different level, then combines the information from all functions to obtain the final representation. The final output E' of DMA stats pooling is the concatenation of all output vectors:

$$E' = [E^1, E^2, E^3, E^4], \tag{12}$$

where E^i represents the output vector of the i-th TCFE-TDNN layer.

C. Loss Function

In the proposed HS-Vectors heart sound classification neural network, we mainly use focal loss [36] and center loss [37] to constrain the learnable parameters. The former is used to improve the classification accuracy of the model for hard-to-classify samples. The focal loss function is defined as follows:

$$L_F = -\frac{1}{m}\sum_{i=0}^{m}(1-\hat{y}_i)^{\gamma}\log(\hat{y}_i), \tag{13}$$

where $\hat{y}_i$ denote the probability that the i-th sample is predicted to be its actual class, and its value is calculated by *Softmax*. The γ is an adjustable hyperparameter that can be adjusted to control the classification of hard-to-classify and easy-to-classify samples. The size of mini-batch is m.

In addition, to improve the convergence speed of the model training and enhance the generalization ability of the model, the center loss is applied to aggregate HS-vector. With the center loss, the HS-vector of each heart sound sample is aggregated to the corresponding class center vector in high-dimensional space. The center loss function can be expressed as:

$$L_C = \frac{1}{2}\sum_{i=1}^{m}\|x_i - c_{y_i}\|_2^2, \tag{14}$$

where $x_i \in \mathbb{R}^{d'}$ denote the HS-vector of the i-th sample. $c_{y_i} \in \mathbb{R}^{d'}$ represents the class center vector corresponding to label y_i of the i-th sample.

Finally, we add L_2 regularization on this CNN model to prevent overfitting:

$$L_2 = \frac{1}{2}\|W\|_2^2, \tag{15}$$

where W denotes the parameter to be learned in the network. The final loss function is:

$$L = L_F + \lambda_1 L_C + \lambda_2 L_2, \tag{16}$$

where λ_* is a scalar to balance the corresponding loss function.

IV. Experiments

In this section, we present the implementation details and evaluation metrics in the experiments and evaluate our proposed algorithm on two heart sound datasets. Subsection IV-A introduces the datasets used for the experiments. Subsection IV-B describes the preprocessing and feature extraction of heart sounds. Subsection IV-C introduces the implementation details and evaluation metrics in the experiments. In Subsection IV-D, we compared the impact of different modules on model classification. Finally, we analyzed the advantages of our proposed algorithm in comparative experiments in Subsection IV-E.

A. Datasets Acquisition

The heart sound datasets used in our experiment are the 2016 PhysioNet/CinC Challenge Dataset (PCCD) and the Pediatric Heart Sound Dataset (PHSD), respectively.

1) 2016 PhysioNet/CinC Challenge Dataset (PCCD): The 2016 PhysioNet/CinC Challenge Dataset [38] consists of six sub-datasets(A through F) from seven different research groups, containing a total of 3240 heart sound recordings, which were collected in either a clinical or nonclinical environment, lasting from 5 seconds to just over 120 seconds. These heart sound recordings were divided into two types: normal and abnormal heart sound recordings and each recording was resampled to 2000 Hz. In addition, this dataset has the characteristic of an imbalance in the number of normal and abnormal samples, which contains 665 Abnormal and 2575 Normal heart sound recordings. The detailed dataset information is listed in Table I.

2) Pediatric Heart Sound Dataset (PHSD): The Pediatric Heart Sound Dataset is a publicly available heart sound dataset that we constructed in our previous work [4]. This dataset contains 528 pediatric heart sound recordings with durations ranging from 3 to 249 seconds. The child subjects involved ranged in age from one month to 12 years. These heart sound recordings are collected by using the Thinklabs One digital stethoscope with a sampling frequency of 44.1 kHz and 16 bits per sample. There is

TABLE I
DETAILED PROFILES FOR THE AVAILABLE HEART SOUND DATASETS

Subset	Data source	Subject number	Normal recordings	Abnormal recordings	RecordingsLength(s)	Acquisition Device
2016 PhysioNet/Cinc Challenge Database						
a	MITHSDB	121	117	292	9∼37	Welch Allyn Meditron
b	AADHSDB	106	386	104	8	3M Littmann E4000
c	AUTHHSDB	31	7	24	10∼122	Welch Allyn Meditron
d	UHAHSDB	38	27	28	6∼49	Infral Corp. Prototype
e	DLUTHSDB	356	1958	183	3∼312.5	MLT201/Piezo/3M Littmann
f	SUAHSDB	112	80	34	16∼88	JABES
Pediatric Heart Sound Dataset						
-	-	137	193	335	3∼249	Thinklabs

also an imbalance in the number of samples in this dataset, which contains 193 Normal and 335 Abnormal heart sound recordings. All the heart sounds are grouped into seven categories: Normal, Atrial Septal Defect (ASD), Ventricular Septal Defect (VSD), both ASD and VSD, Tetralogy of Fallot (TOF), both ASD and TOF, and other heart-related diseases, such as Mitral Regurgitation, Aortic Stenosis, Pulmonary Stenosis, etc.

B. Data Preprocessing and Feature Extraction

As mentioned in Section I, data preprocessing and feature extraction are important pre-operation. In our method, heart sound signals were preprocessed with a second-order Butterworth filter from 25 Hz to 400 Hz for denoising. Since the real heart sound signal is non-fixed length, we employed a sliding window algorithm to intercept the heart sound signal into 3 s length patches with a stride of 1 s. It should be pointed out that this is not a segmentation (S1, systole, S2, diastole) but a simple slice. Then, the 81-dimensional MFCCs from a 256 ms window with a 128 ms frame shift were fed into our model.

C. Implementation Details and Evaluation Metrics

The proposed HS-Vectors is trained from scratch on a NVIDIA GeForce RTX 3090 GPU. The adaptive moment (Adam) estimation algorithm is used as the optimizer. The hyperparameters γ, λ_1 and λ_2 of the proposed loss function are set to 2, 1 and 0.01, respectively. Our proposed model is trained for 60 epochs and a batch size of 512. The learning rate is initially set to 0.002, and the cosine annealing algorithm is used to decay the learning rate to 0.00002 during the training. Finally, the 10-fold cross-validation is adopted to evaluate the performance of our algorithm.

For the evaluation of the proposed algorithm, four evaluation metrics are introduced for experiment: *Accuracy*, *Sensitivity*, *Specificity*, and *F1-score* which are given as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{17}$$

$$Sensitivity = Recall = \frac{TP}{TP + FN}, \tag{18}$$

$$Specificity = \frac{TN}{TN + FP}, \tag{19}$$

$$Precision = \frac{TP}{TP + FP}, \tag{20}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \tag{21}$$

where TP is the number of true positive results, TN is the number of true negative results, FP is the number of false positive results and FN is the number of false negative results. We considered the correctly diagnosed abnormal heart sound recordings as true positive samples in this article.

D. Evaluation of the Proposed HS-Vectors

To analyze the relative contributions of different components of our architecture, we evaluate some variants of the proposed method with different settings on the PCCD dataset.

1) Model Performance With Different Mask Rates: As mentioned subsection III-B, *mask rate* is an important parameter in the process of model generating HS-vector. In this part, we find the best mask rate for the task of abnormal heart sound detection. In order to ensure that the total number of features is within an appropriate range, we take the mask rate v from the interval [0,0.9]. Fig. 8 shows the classification results for different mask rates on PCCD. It shows that our model can achieve better classification performance when focusing on a small number of important features, even if we mask out 90% of the features. It proves that the mask-attention can balance the relationship between the discriminative ability and the feature dimension and remove redundant information. Besides, the *F1-score* changes with the *mask rate* and has the highest value when the *mask rate* is 0.5. Therefore, we selected 0.5 as the *mask rate* of the proposed HS-Vectors according to the priority order of *F1-score*, *accuracy*, *specificity* and *sensitivity*.

2) Model Performance With Different Architectures: To evaluate the proposed TCFE-TDNN and DMA stats pooling, we construct three architectures including: TCFE-TDNN&DMA Stats Pooling (named: *M1*), TDNN&DMA Stats Pooling (named: *M2*), and TCFE-TDNN&MA Stats Pooling (named: *M3*) to analyze the contribution of each module to classification. *M1* is our proposed model. *M2* replaces TCFE-TDNN

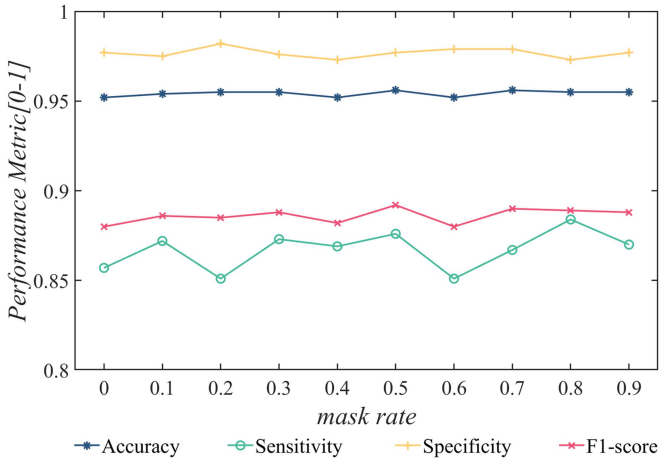


Fig. 8. Classification performance on the PCCD with respect to different mask rates. When v is equal to 0, it means we are not using the mask mechanism.

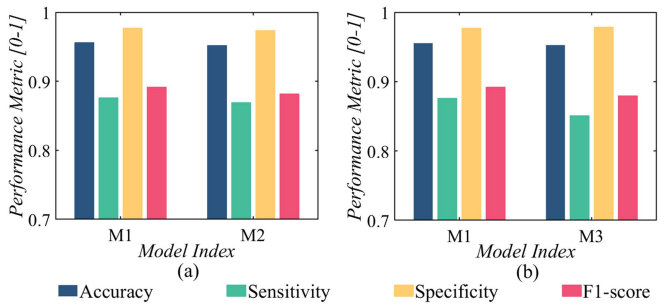


Fig. 9. Classification performance on the PhysioNet/CinC 2016 between models with different network architectures.

in *M2* with TDNN. *M3* uses single-scale DMA stats pooling for feature aggregation. The *mask rate* for all models is 0.5.

M1 vs. M2: Since TCFE-TDNN is an improved version of TDNN, we compared the abnormal heart sound detection performance of TCFE-TDNN module and TDNN module. For TDNN in *M2*, the same-convolution is used for frame-level feature extraction, and each layer outputs features of the same scale. Specifically, the output $H^i \in \mathbb{R}^{C_i \times T_i}$ for i-th layer of TDNN satisfies $T_1 = T_2 = T_3 = T_4$ and $C_1 = C_2 = C_3 = C_4 = 512$. Fig. 9(a) shows the experimental results, in which we can observe that the proposed model *M1* has a higher index score. It shows the superiority of our TCFE-TDNN module.

M1 vs. M3: To see the effectiveness of the DMA stats pooling, we compared multi-scale *M1* and single-scale *M3*. The *M3* only uses MA stats pooling to perform feature aggregation on H^4. Fig. 9(b) shows the comparison results, in which the multi-scale model *M1* has better performance in *sensitivity* and *F1-score*. Since the DMA stats pooling aggregates frame-level features from different TCFE-TDNN layers, the impact of potential lack of detailed features on the model generation HS-vector is reduced.

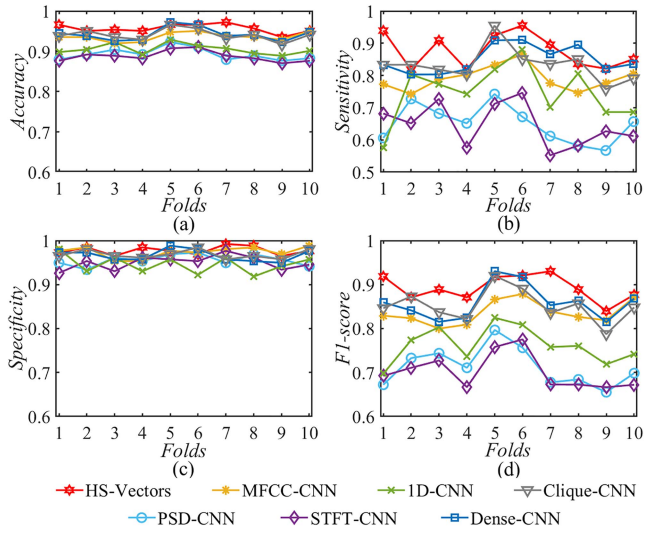


Fig. 10. The accuracy, sensitivity, specificity and F1-score in different cross-validation folds for the different implemented methods.

E. Comparative Experiments and Discussion

1) 2016 PhysioNet/CinC Challenge Dataset: To verify the superiority of our proposed HS-Vectors, we compared our method with several state-of-the-art methods for heart sound classification on PCCD. These methods include the ones using 1D heart sound features [4], [10], [26], [28] and others based on 2D heart sound features [27], [39], [40], [41], [42]. The results of the comparative experiments on PCCD are shown in Table II. We have implemented all the methods except those marked with * in Table. The experimental results of these unimplemented methods are derived from their papers. From Table II, it can be observed that our proposed method has a decent performance compared to the state-of-the-art methods. In particular, our method is more than 3% ahead of other methods in the F1-score indicator.

Furthermore, the *Accuracy*, *Sensitivity*, *Specificity* and *F1-score* in different cross-validation folds for the different implemented methods are illustrated in Fig. 10. It can be found that our method has stable and better performance in each fold, which can prove the superiority of our method.

2) Pediatric Heart Sound Dataset: The results of the comparative experiments on PHSD are shown in Table III. It can be seen that our proposed method still achieves the best performance, and its evaluation indicators: *Accuracy*, *Specificity*, *F1-score* are close to 1 and *Sensitivity* is equal to 1. It proves that our method, which extends the HS-vector embedding into a heart sound deep learning model, can not only represent the heart sound characteristics without a large number of datasets but also make the model better adapt to the datasets, achieving stronger generalization.

V. Conclusion

In this study, we propose the TCFE-TDNN module, which adopts a variable hidden feature extraction strategy to diversify

TABLE II
CLASSIFICATION PERFORMANCE COMPARISON ON THE PCCD BETWEEN OUR PROPOSED METHOD AND OTHER HEART SOUND CLASSIFICATION METHODS

Methods	Accuracy	Sensitivity	Specificity	F1-score
*AlexNet-SVM [39]	0.877	0.837	0.899	-
*AdaBoost-CNN [28]	-	0.778	0.942	-
*CNN+RNN [26]	-	0.830	0.960	-
*CWT-CNN [40]	0.860	0.874	0.867	-
STFT-CNN [41]	0.887	0.646	0.949	0.701
PSD-CNN [42]	0.892	0.649	0.955	0.712
MFCC-CNN [27]	0.936	0.790	0.973	0.836
1-D CNN [10]	0.905	0.747	0.946	0.762
Dense-CNN [4]	0.942	0.849	0.966	0.858
Clique-CNN [4]	0.940	0.832	0.968	0.851
HS-Vectors(**Our method**)	**0.956**	**0.876**	**0.977**	**0.892**

TABLE III
PERFORMANCE ON THE INDEXES OF OUR METHOD ARE COMPARED ON THE PHSD WITH THE AVERAGE PERFORMANCE OF OTHER HEART SOUND CLASSIFICATION METHODS

Methods	Accuracy	Sensitivity	Specificity	F1-score
STFT-CNN [41]	0.764	0.954	0.435	0.837
PSD-CNN [42]	0.847	0.832	0.877	0.872
MFCC-CNN [27]	0.965	0.978	0.942	0.973
1-D CNN [10]	0.929	0.942	0.906	0.944
Dense-CNN [4]	0.979	0.996	0.948	0.983
Clique-CNN [4]	0.917	0.934	0.892	0.923
HS-Vectors(**Our method**)	**0.998**	**1**	**0.995**	**0.998**

the hidden features. The network can observe the potential abnormal performance of each frame in both low and high-dimensional frequency hidden feature space. Then, we apply the Dynamic Masked-Attention Statistics Pooling to each TCFE-TDNN layer, which dynamically masks out irrelevant features so that the HS-vector generated by the model focuses on the representation of essential regions and improves the classification performance of heart sounds. Compared with previous studies, our method adds a global feature representation process after the local feature extraction stage, which enhances the generalization of the model to long-time features. Experiments have proved that our method has strong robustness in terms of *Accuracy*, *Sensitivity*, *Specificity* and *F1-score*.

Although our method achieves superior performance, these metrics suggest that the method still has some drawbacks for clinical application. For example, low-quality heart sound signals in the dataset may reduce the model's ability to generalize to valid data. Furthermore, our model cannot be applied to the diagnosis scenario of multiple anomaly categories. However, as a heart sound-aided diagnosis system, the superior performance shown by our method makes it sufficient to assist doctors in diagnosis.

This thesis has provided a deeper insight into heart sound detection. Our method can capture pathological information and perform abnormal heart sound detection within the effective interval of clinical PCG signals, opening a series of perspectives for future research and clinical applications. The task of identifying different types of abnormalities will be our future research work.

REFERENCES

[1] WHO, "Cardiocascular diseases(cvds)," Jun. 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] U. Alam, O. Asghar, S. Q. Khan, S. Hayat, and R. A. Malik, "Cardiac auscultation: An essential clinical skill in decline," *Brit. J. Cardiol.*, vol. 17, no. 1, pp. 8–10, 2010.

[3] S. Li, F. Li, S. Tang, and F. Luo, "Heart sounds classification based on feature fusion using lightweight neural networks," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–9, 2021.

[4] B. Xiao et al., "Follow the sound of children's heart: A deep-learning-based computer-aided pediatric chds diagnosis system," *IEEE Internet Things J.*, vol. 7, no. 3, pp. 1994–2004, Mar. 2020.

[5] E. Kay and A. Agarwal, "Dropconnected neural networks trained on time-frequency and inter-beat features for classifying heart sounds," *Physiol. Meas.*, vol. 38, no. 8, 2017, Art. no. 1645.

[6] D. Chakraborty, S. Bhattacharya, A. Thakur, A. R. Gosthipaty, and C. Datta, "Feature extraction and classification of phonocardiograms using convolutional neural networks," in *Proc. IEEE 1st Int. Conf. Convergence Eng.*, 2020, pp. 275–279.

[7] M. U. Khan, Z. Mushtaq, M. Shakeel, S. Aziz, and S. Z. H. Naqvi, "Classification of myocardial infarction using MFCC and ensemble subspace KNN," in *Proc. Int. Conf. Elect., Commun., Comput. Eng.*, 2020, pp. 1–5.

[8] L. Zhiming and M. Sheng, "Multi-label classification of heart sound signals," in *Proc. IEEE Int. Conf. Comput. Eng. Artif. Intell.*, 2021, pp. 355–360.

[9] M. Markaki, I. Germanakis, and Y. Stylianou, "Automatic classification of systolic heart murmurs," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2013, pp. 1301–1305.

[10] H. Ryu, J. Park, and H. Shin, "Classification of heart sound recordings using convolution neural network," in *Proc. IEEE Comput. Cardiol. Conf.*, 2016, pp. 1153–1156.

[11] A. I. Humayun, S. Ghaffarzadegan, M. I. Ansari, Z. Feng, and T. Hasan, "Towards domain invariant heart sound abnormality detection using learnable filterbanks," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 8, pp. 2189–2198, Aug. 2020.

[12] F. Renna, J. Oliveira, and M. T. Coimbra, "Deep convolutional neural networks for heart sound segmentation," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 6, pp. 2435–2445, Nov. 2019.

[13] O. Deperlioglu, "Heart sound classification with signal instant energy and stacked autoencoder network," *Biomed. Signal Process. Control*, vol. 64, 2021, Art. no. 102211.

[14] S. Sun, Z. Jiang, H. Wang, and Y. Fang, "Automatic moment segmentation and peak detection analysis of heart sound pattern via short-time modified Hilbert transform," *Comput. Methods Programs Biomed.*, vol. 114, no. 3, pp. 219–230, 2014.
[15] D. B. Springer, L. Tarassenko, and G. D. Clifford, "Logistic regression-HSMM-based heart sound segmentation," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 4, pp. 822–832, Apr. 2016.
[16] F. Noman, S.-H. Salleh, C.-M. Ting, S. B. Samdin, H. Ombao, and H. Hussain, "A Markov-switching model approach to heart sound segmentation and classification," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 3, pp. 705–716, Mar. 2020.
[17] T. Fernando, H. Ghaemmaghami, S. Denman, S. Sridharan, N. Hussain, and C. Fookes, "Heart sound segmentation using bidirectional LSTMs with attention," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 6, pp. 1601–1609, Jun. 2020.
[18] E. Messner, M. Zöhrer, and F. Pernkopf, "Heart sound segmentation-an event detection approach using deep recurrent neural networks," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 9, pp. 1964–1974, Sep. 2018.
[19] S. Ari, K. Hembram, and G. Saha, "Detection of cardiac abnormality from PCG signal using LMS based least square SVM classifier," *Expert Syst. with Appl.*, vol. 37, no. 12, pp. 8019–8026, 2010.
[20] F. Safara, S. Doraisamy, A. Azman, A. Jantan, and A. R. A. Ramaiah, "Multi-level basis selection of wavelet packet decomposition tree for heart sound classification," *Comput. Biol. Med.*, vol. 43, no. 10, pp. 1407–1414, 2013.
[21] H. Uğuz, "Adaptive neuro-fuzzy inference system for diagnosis of the heart valve diseases using wavelet transform with entropy," *Neural Comput. Appl.*, vol. 21, no. 7, pp. 1617–1628, 2012.
[22] S. Patidar, R. B. Pachori, and N. Garg, "Automatic diagnosis of septal defects based on tunable-q wavelet transform of cardiac sound signals," *Expert Syst. Appl.*, vol. 42, no. 7, pp. 3315–3326, 2015.
[23] T. Dissanayake, T. Fernando, S. Denman, S. Sridharan, H. Ghaemmaghami, and C. Fookes, "A robust interpretable deep learning classifier for heart anomaly detection without segmentation," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 6, pp. 2162–2171, Jun. 2021.
[24] A. Quiceno-Manrique, J. Godino-Llorente, M. Blanco-Velasco, and G. Castellanos-Dominguez, "Selection of dynamic features based on time–frequency representations for heart murmur detection from phonocardiographic signals," *Ann. Biomed. Eng.*, vol. 38, no. 1, pp. 118–137, 2010.
[25] H. Uğuz, "A biomedical system based on artificial neural network and principal component analysis for diagnosis of the heart valve diseases," *J. Med. Syst.*, vol. 36, no. 1, pp. 61–72, 2012.
[26] C. Thomae and A. Dominik, "Using deep gated RNN with a convolutional front end for end-to-end classification of heart sound," in *Proc. IEEE Comput. Cardiol. Conf.*, 2016, pp. 625–628.
[27] J. Rubin, R. Abreu, A. Ganguli, S. Nelaturi, I. Matei, and K. Sricharan, "Classifying heart sound recordings using deep convolutional neural networks and mel-frequency cepstral coefficients," in *Proc. IEEE Comput. Cardiol. Conf.*, 2016, pp. 813–816.
[28] C. Potes, S. Parvaneh, A. Rahman, and B. Conroy, "Ensemble of feature-based and deep learning-based classifiers for detection of abnormal heart sounds," in *Proc. IEEE Comput. Cardiol. Conf.*, 2016, 2016 pp. 621–624.
[29] M. Deng, T. Meng, J. Cao, S. Wang, J. Zhang, and H. Fan, "Heart sound classification based on improved MFCC features and convolutional recurrent neural networks," *Neural Netw.*, vol. 130, pp. 22–32, 2020.
[30] J. P. Dominguez-Morales, A. F. Jimenez-Fernandez, M. J. Dominguez-Morales, and G. Jimenez-Moreno, "Deep neural networks for the recognition and classification of heart murmurs using neuromorphic auditory sensors," *IEEE Trans. Biomed. Circuits Syst.*, vol. 12, no. 1, pp. 24–34, Feb. 2018.
[31] D. R. Megalmani, B. Shailesh, A. Rao, S. S. Jeevannavar, and P. K. Ghosh, "Unsegmented heart sound classification using hybrid CNN-LSTM neural networks," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2021, pp. 713–717.
[32] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process*., 2018, pp. 5329–5333.
[33] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proc. Interspeech* 2020, 2020, pp. 3830–3834.
[34] Y. Zhu, T. Ko, D. Snyder, B. Mak, and D. Povey, "Self-attentive speaker embeddings for text-independent speaker verification," in *Proc. Interspeech*, 2018, vol. 2018, pp. 3573–3577.
[35] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.
[36] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.
[37] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 499–515.
[38] G. D. Clifford et al., "Classification of normal/abnormal heart sound recordings: The physionet/computing in cardiology challenge 2016," in *Proc. IEEE Comput. Cardiol. Conf.*, 2016, pp. 609–612.
[39] H. Alaskar, N. Alzhrani, A. Hussain, and F. Almarshed, "The implementation of pretrained alexnet on PCG classification," in *Proc. Int. Conf. Intell. Comput.*, 2019, pp. 784–794.
[40] A. Meintjes, A. Lowe, and M. Legget, "Fundamental heart sound classification using the continuous wavelet transform and convolutional neural networks," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2018, pp. 409–412.
[41] Q. Chen, W. Zhang, X. Tian, X. Zhang, S. Chen, and W. Lei, "Automatic heart and lung sounds classification using convolutional neural networks," in *Proc. IEEE Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2016, pp. 1–4.
[42] T. Nilanon, J. Yao, J. Hao, S. Purushotham, and Y. Liu, "Normal/abnormal heart sound recordings classification using convolutional neural network," in *Proc. IEEE Comput. Cardiol. Conf.*, 2016, pp. 585–588.