Parameter-Efficient Densely Connected Dual Attention Network for Phonocardiogram Classification

Keying Ma, Jianbo Lu, and Benzhuo Lu

***Abstract*—Cardiac auscultation, exhibited by phonocardiogram (PCG), is a non-invasive and low-cost diagnostic method for cardiovascular diseases (CVDs). However, deploying it in practice is quite challenging, due to the inherent murmurs and a limited number of supervised samples in heart sound data. To solve these problems, not only heart sound analysis based on handcrafted features, but also computer-aided heart sound analysis based on deep learning have been extensively studied in recent years. Though with elaborate design, most of these methods still use additional pre-processing to improve classification performance, which heavily relies on time-consuming experienced engineering. In this article, we propose a parameter-efficient densely connected dual attention network (DDA) for heart sound classification. It combines two advantages simultaneously of the purely end-to-end architecture and enriched contextual representations of the self-attention mechanism. Specifically, the densely connected structure can automatically extract the information flow of heart sound features hierarchically. Alongside, improving contextual modeling capabilities, the dual attention mechanism adaptively aggregates local features with global dependencies via a self-attention mechanism, which captures the semantic interdependencies across position and channel axes respectively. Extensive experiments across stratified 10-fold cross-validation strongly evidence that our proposed DDA model surpasses current 1D deep models on the challenging Cinc2016 benchmark with significant computational efficiency.**

***Index Terms*—Cinc2016, dense block, dual attention, phonocardiogram.**

Manuscript received 14 April 2022; revised 26 December 2022 and 6 May 2023; accepted 8 June 2023. Date of publication 15 June 2023; date of current version 6 September 2023. This work was supported by the National Natural Science Foundation of China under Grants 11771435 and 22073110. *(Corresponding authors: Jianbo Lu; Benzhuo Lu.)*

Keying Ma and Benzhuo Lu are with the State Key Laboratory of Scientific and Engineering Computing, National Center for Mathematics and Interdisciplinary Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: makeying@lsec.cc.ac.cn; bzlu@lsec.cc.ac.cn).

Jianbo Lu is with the National Human Genetics Resource Center, National Research Institute for Family Planning, Beijing 100081, China (e-mail: jblu@lsec.cc.ac.cn).

Digital Object Identifier 10.1109/JBHI.2023.3286585

I. Introduction

CARDIOVASCULAR diseases (CVDs) are consistently heavy threats to human health [1]. Early detection of CVDs, including pathological conditions related to blood vessels and heart valves, is quite important for follow-up counseling and medical therapy to reduce mortality. Heart sound auscultation, as an integral part of the physical examination, enables the detection of many pathological heart diseases and can be used as a basis for further diagnostic workup [2]. Therefore, studying a system that can assist in the diagnosis of heart sound auscultation is necessary for the early treatment of heart disease, especially in some areas that lack cardiologists.

Compared to other cardiac-related signals [3], [4], phonocardiogram (PCG) is regarded as a low-cost cardiac acoustic signal for diagnosing CVDs [5]. The PCG signal is a time series of audio recordings of sounds transformed on the surface of the chest [2]. When recording the heart sounds of a cardiac cycle, the PCG signal may also have the additional third heart sound (S3), fourth heart sound (S4), clicks, and heart murmurs resulting from the galloping blood flow through the valve. Some heart murmurs may be innocent owing to physiologic high-flow conditions, while those caused by valvular stenosis (such as aortic stenosis) or regurgitation (such as mitral regurgitation) are pathological murmurs [6], [7], [8]. These unavoidable murmurs, which overlap with normal and abnormal heart sounds in the frequency domain [9], bring great difficulties to the classification of heart sounds, and thus PCG classification becomes notoriously challenging.

Early approaches use machine learning models to tackle the heart sound classification task [10], [11], [12], [13], [14], [15], [16], [17]. For example, Sun et al. [15] extract four diagnostic time and frequency domain features and use a support vector machine (SVM)-based classification boundary method. Uğuz et al. [16] apply wavelet transform, short-time Fourier transform, and wavelet entropy during the feature extraction stage and feed features into the hidden Markov model (HMM). Avendaño et al. [17] employ eigenplane-based principal component analysis (PCA) and partial least squares (PLS) techniques to perform feature extraction, and then the features extracted are used to feed a simple k-nearest neighbor (k-NN) classifier. Despite the decent performance, these machine learning-based methods usually heavily rely on the manually designed pre-processing and feature extraction approaches based on expert experience, which

generally suffer from time-consuming experienced engineering and poor generalization.

Recently, deep learning technologies have made remarkable achievements in the processing of regular data, e.g., image [18], [19], video [20], and speech [21]. Therefore, there has been a growing focus on leveraging the power of deep learning for PCG classification [6], [22], [23], [24], [25], [26], [27], [28]. The pioneering studies of [25], [29], [30], [31] employs mel-frequency cepstral coefficients (MFCC) [32] to transfer 1D raw signals to 2D representations and then perform further feature extraction using deep neural networks, which have acoustical patterns with finer granularity. However, this line of work has two main drawbacks: 1) involving an extra set of hyperparameters to optimize the transformation process; 2) requiring heavier computation to train 2D convolution kernels. To circumvent these issues, some researchers propose to take the 1D raw waveform signals as the input without any additional transformation process [26], [28], [33]. Albeit showing the potential for PCG classification, they usually realize the end-to-end architecture by simply applying 1D convolution on local segmentations, which could be powerless in accurately capturing diverse physiological patterns due to the inherent murmurs in heart sound data.

In this article, we develop a novel framework, termed densely connected dual attention network (DDA), for PCG classification. DDA can automatically capture the discriminative features, while not requiring elaborately designed pre-processing to manually choose features. Technically, the proposed DDA introduces a 1D densely connected convolutional neural network (CNN) along with two parallel self-attention modules, a position attention module (PAM), and a channel attention module (CAM). Specifically, the densely connected structure allows each layer to have direct access to update gradients from the loss function and the original input signal, which can improve the information flow and parameter efficiency of the entire network. To further capture informative and discriminative heart sound signals, we introduce a dual self-attention mechanism (DA) to learn feature dependencies in the position and channel dimensions respectively, which endows our DDA with the capability of adaptively emphasizing important features and suppressing unimportant features. The outputs of position and channel attention modules are fused by a simple summation operation to further boost the feature representation. In this way, our DDA combines the best of two worlds, i.e., automated training of end-to-end architecture and enhanced feature representations of the dual attention mechanism. Experimental results show that our simple-yet-effective method contributes towards a significant improvement across the PCG classification task. A vivid example is that, compared with Xiao's work [33], the proposed DDA obtains performance gains of up to 4.73% absolute overall score, which strikes the best trade-off in classification performance and model size.

The key contributions can be summarized as follow:

- We propose a novel densely connected dual attention network (DDA) for heart sound classification. It can automatically learn informative and discriminative features from heart sound signals, without requiring manual pre-processing for selecting specified features.
- A dual attention mechanism is introduced to learn the position and channel inter-dependencies of features in parallel. It markedly improves the representation ability of our model with negligible extra computational cost.
- Extensive experiments demonstrate that our DDA model achieves new state-of-the-art results on widely used metrics (such as accuracy, score, sensitivity, and specificity) on the challenging PhysioNet/CinC 2016 benchmark.

II. Related Work

A. Heart Sound Classification

In this section, we briefly review existing methods for PCG classification.

Traditional heart sound classification methods usually adopt machine learning models to separately perform three steps in sequence: signal segmentation, feature extraction, and final classification. The first step, signal segmentation, is to perform precise segmentation for the fundamental heart sound shards, with the goal of facilitating extracting the features of each heartbeat.

Heart sounds include fundamental heart sounds (FHSs), typically with the first (S1) and second (S2) heart sounds occurring successively [9]. The initial S1 arises at the early stage of systole due to the continuous closure of the atrioventricular mitral and tricuspid valves. The subsequent heart sound S2 arises at the early stage of the diastole owing to the closure of the aortic and pulmonary valves. To precisely segment these fundamental shards, the machine learning models include envelope-based methods [34], feature-based methods [35], machine learning [36], and HMM-based methods [37]. Furthermore, feature extraction is usually considered the most critical step for traditional heart sound classification methods based on machine learning [38]. These methods mainly rely on the segmented S1 and S2 heart sounds to manually extract features. Typically, wavelet transform and Fourier transform are used to extract time [39], frequency [40], or time-frequency features [12], [41], [42], [43], revealing some physiological and pathological information for subsequent classification. Finally, a softmax classifier based on machine learning is trained with the extracted discriminative feature as input, where the canonical ones contain SVM [10], [11], [12], HMM [14], k-NN [13], etc. However, due to heavy reliance on expert experience and manually precise processing, designing proper machine learning models to handle heart sound classification remains quite an intractable challenge.

Plenty of follow-up research is devoted to developing deep learning methods for heart sound classification [6], [22], [23], [24], [25], [26], [27], [28]. They either introduce fixed-length segmentation using overlapping windows [23], [26] or abandon the segmentation to use the filling method, such as [27], [44]. Among them, the original high-dimensional heart sound signals are first projected into low-dimensional features through a variety of mathematical transformations. For this purpose, they typically convert the heart sound classification task into the image classification task by extracting 2D feature maps, yet a large parameter overhead is required for the transformation of features into 2D feature maps, which is not conducive to

the application on mobile devices. The canonical approaches mainly include MFCCs [25], [29], [30], [31], log mel-frequency spectral coefficients (MFSCs) [29], [45], [46], power spectral density (PSD) [23], and neuromorphic auditory sensors [6]. Some studies also directly employ 1D features for computational efficiency [26], [47], [48], [49], [50]. For example, Ryu et al. [26] adopt the Windowed-sinc Hamming filter as the pre-processing step to assist a simple 1D CNN. Different from these methods, our proposed DDA can directly perform heart sound classification with simple fixed-size segmentations and remove the need for manual pre-processing steps.

Recently, some researchers have explored performing PCG classification in a purely end-to-end manner. These methods take raw waveform data as input, combine feature extractor and classifier, and automatically extract the most discriminative features. Xu et al. [28] develop a 1D CNN model, which adopts a block-stacked style architecture for parameter efficiency. Xiao et al. [33] introduce the 1D CNN-based model that focuses on feature reusing, which can obtain an overall score of 90.51%. Compared with these methods, benefitting from the dual attention mechanism, our proposed DDA can automatically extract informative features and employ parallel position and channel attention modules to improve bidimensional information flow in a computationally efficient pattern.

B. Attention Mechanism

In convolution neural networks, convolutional filter weights are quite intractable to adapt dynamically to input variation since they are typically fixed after training. Thus, a significant number of attention-based methods are developed to resolve this issue [51], [52], [53], [54], [55], [56]. SE-Net [52] is a milestone of these efforts by modeling channel attention with encouraging performance gains. Subsequently, the non-local network [55] attempts to capture long-term dependencies in both space and time, which shows the potential of the attention mechanism for accurate video classification yet requires high memory and computation overheads. The Criss-cross network [54] harvests the contextual information of all the pixels on the criss-cross path to generate sparse attention maps for reducing complexity. Liu et al. [57] apply a self-attention generative adversarial network to boost the image completion task. Bello et al. [58] adopt the self-attention mechanism to promote image classification performance. In this article, we are the first to propose a dual self-attention mechanism to capture rich contextual information for heart sound classification studies. Here we wish to push forward the auscultation of heart sounds by focusing on combining the merits of the self-attention and end-to-end convolution neural network for strengthening discriminant feature representations.

III. Methodology

In this section, we develop a 1D CNN-based combined dual attention mechanism system for CVD diagnosis. To start with, we introduce the heart sound segmentation strategy. And then the details of our model for heart sound shards classification are provided. Finally, the decision rule transforming the shard-level results into a recording-level diagnosis is proposed.

A. Segmentation

Following the assumptions in [23], the abnormality of heart sounds can be observed with each heartbeat if the heart pathology is present. The abnormal heart sound recording can be diagnosed within a few seconds, which has been concurred by general practitioners [23]. Therefore, we employ the fixed-length sliding window for shard-level prediction by segmenting each heart sound recording into a series of flattened shards. The window length and slide interval are empirically selected as three seconds and one second, respectively. Each shard is labeled according to the class corresponding to the intact heart sound before segmentation. This segmentation strategy requires no additional hyperparameters and increases the number of data samples, alleviating the problem of a quite small sample size.

B. DDA Architecture

The proposed method contains a densely connected convolutional network (DenseNet) equipped with a dual self-attention mechanism. Intuitively, the framework of our DDA net is illustrated schematically in Fig. 1. Specifically, we first adopt the original 1D signals as input to capture the time-frequency features using DenseNet. Then, the learned features are fed into dual attention modules, which are comprised of two parallel branches – the position attention module and the channel attention module. Technically, for the position attention module, we introduce the feature similarities between different positions of the extracted feature map to model long-term dependencies of PCG signals via a self-attention mechanism. That is, each feature is encouraged to perform an interaction with ones with different positions to produce rich feature representations, where their distance may be far apart in the position dimension. Likewise, the channel attention module models the correlations between any two-channel maps using the weighted sum of channel maps via a similar self-attention mechanism, which renders the ability to increase the inter-channel dependencies. In this way, our model can learn to emphasize important features and suppress the unnecessary ones, from the perspective of the channel and position axes respectively. In the following, we introduce the detailed design of our proposed architecture.

1) Dense Block: The densely connected convolutional network achieves highly competitive performance benefiting from its efficient information flow. To extract informative and rich features, we migrate this effective dense structure framework to the one-dimensional. Specifically, the feature maps of all previous layers in a dense block are directly connected to subsequent layers. Technically, given a shard-level PCG signal $\mathbf{X}_0$, we define a composite function $H(\cdot)$ including multiple operations, such as convolutional layers and batch normalization layers. The updated process of feature maps can be formulated by

$$\mathbf{X}_l = H(\{\mathbf{X}_0, \mathbf{X}_1, \ldots, \mathbf{X}_{l-1}\}), \tag{1}$$

where $\mathbf{X}_1, \mathbf{X}_2 \ldots, \mathbf{X}_{l-1}$ represent the feature maps produced in layers $1, 2, ..., l-1$ respectively, and $\{\cdot\}$ signifies the concatenation operation along channels. Details of the dense layers are shown in Fig. 2. Different from the traditional dense layers using two convolutional layers with the kernel sizes of 1×1 and 1

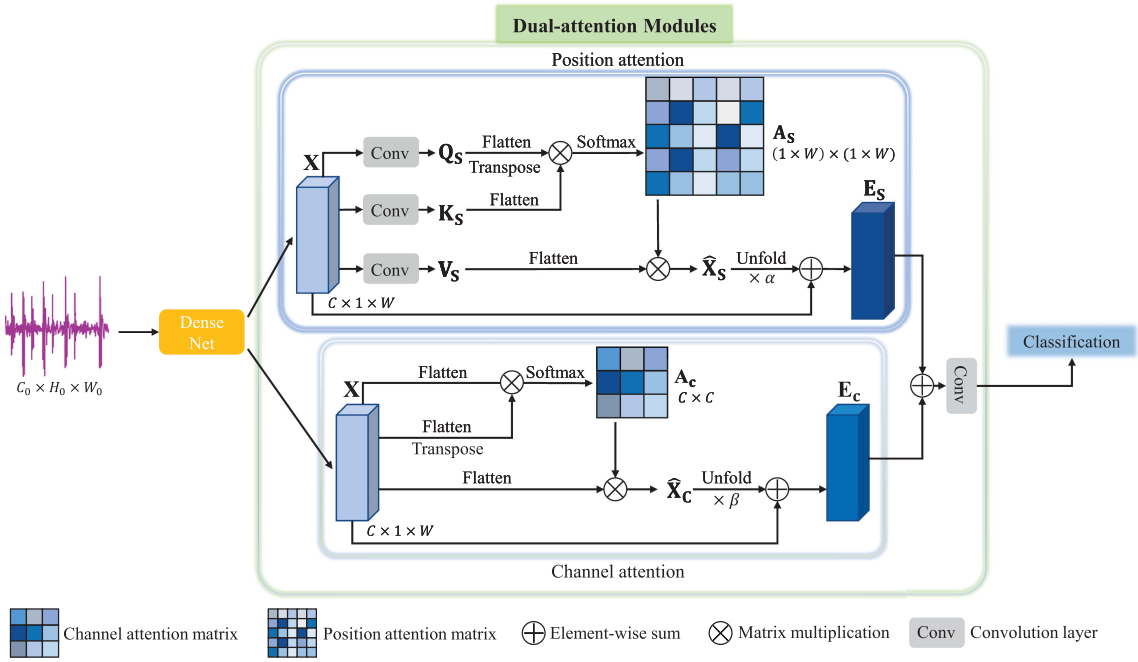


Fig. 1. Overview of our proposed densely connected dual attention model (DDA). Convolutional layers in the grey background are with kernel size 1×1, stride 1, and zero padding.

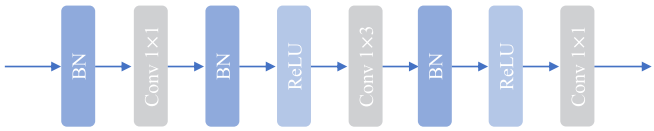


Fig. 2. Structure of dense layer. "Conv1×3" denotes the convolutional layer with the kernel size 3 and the padding size (0,1).

$\times$ 3 respectively, we add an extra convolutional layer with the kernel size of 1×1 at the end of each dense layer to realize the information interactions between channels, which can enhance the information flow in each dense layer as pointed out in [59], [60], [61]. Following the original setting in DenseNet [62], we use 4 stacked dense blocks, each of which has a growth rate (k) of 12.

2) Transition Layer: Although the direct connection between layers enhances the information flow, the number of connections and the number of dense block layers have a quadratic relationship, which will take up a lot of computing resources. In our work, to avoid the inability to densely concatenate features when downsampling features, the transition layers with a 1×1 convolutional layer followed by a 1×2 average pooling layer are adopted to divide the networks into 4 dense blocks, each of which extracts different hierarchical features.

3) Dual Self-Attention Mechanism: To further increase the representation power, we parallelly emphasize the discriminative features along position and channel axes, which are based on the resultant feature map with high-level semantics from DenseNet. Next, we will elaborate on the processes.

Position attention module: The topmost part of Fig. 1 illustrates the position attention module. Formally, given the feature map $\mathbf{X} \in \mathbb{R}^{C \times 1 \times W}$ from DenseNet, we first feed $\mathbf{X}$ into the convolution layers to produce two new feature representations $\mathbf{Q_s}, \mathbf{K_s} \in \mathbb{R}^{\frac{C}{8} \times 1 \times W}$ (denoted as *query* and *key*) respectively, whose channels are reduced by 8 times for efficient computation. We flatten $\mathbf{Q_s}, \mathbf{K_s}$ to dimension $\frac{C}{8} \times W$ and then perform a dot product between $(F(\mathbf{Q_s}))^\top$ and $F(\mathbf{K_s})$ to obtain a similarity matrix $\mathbf{A_s} \in \mathbb{R}^{W \times W}$ with a softmax operation, which represents the pairwise similarity between each element in $\mathbf{Q_s}$ and $\mathbf{K_s}$. Here $F(\cdot)$ flattens the input feature map into the vector representation. The higher similarity scores of the two elements indicate the greater relevance between them. Meanwhile, another new feature representation $\mathbf{V_s} \in \mathbb{R}^{C \times 1 \times W}$ (denoted as *value*) will be generated by feeding $\mathbf{X}$ to a convolution layer, which is further flattened to dimension $C \times W$. The convolution layers for Q, K, and V are with kernel size 1×1, stride 1, and zero padding. Then we carry out a matrix multiplication between $\mathbf{A_s}$ and $F(\mathbf{V_s})$ and then unfold it to obtain the attention score $\hat{\mathbf{X}}_\mathbf{s} \in \mathbb{R}^{C \times W}$. Finally, we adopt a learnable parameter α to perform the weighted residual-style feature blending, i.e., residual connection with the weighted feature map $\alpha \hat{\mathbf{X}}_\mathbf{s}$ and the original feature map $\mathbf{X}$. This process can be written as

$$\begin{aligned} \mathbf{A_s} &= \text{softmax}\left((F(\mathbf{Q_s}))^\top F(\mathbf{K_s})\right), \\ \hat{\mathbf{X}}_\mathbf{s} &= F(\mathbf{V}_s) \cdot \mathbf{A}_s, \\ \mathbf{E_s} &= \mathbf{X} + \alpha U(\hat{\mathbf{X}}_\mathbf{s}), \end{aligned} \tag{2}$$

where $U(\cdot)$ unfolds the input vector from dimension $C \times W$ into $C \times 1 \times W$.

Channel attention module: We produce a channel attention map by mining the dependencies among different channels of the resultant feature map from DenseNet. This channel attention

TABLE I
ARCHITECTURE DETAILS OF OUR PROPOSED DENSELY CONNECTED DUAL-ATTENTION MODEL (DDA)

Layer name	Output size	Configurations
Inputs	1×1×6000	
Conv. layer	24×1×2998	Conv 1×7 with stride (1,2), padding (0,1)
Pooling layer	24×1×1499	Maxpool 1×3 with stride (1,2), padding (0,1)
Dense block	96×1×1499	Dense layer × 6, k=12
Transition layer	48×1×749	Conv 1×1 with stride 1, padding 0 Avepool 1×2 with stride (1,2)
Dense block	120×1×749	Dense layer × 6, k=12
Transition layer	60×1×374	Conv 1×1 with stride 1, padding 0 Avepool 1×2 with stride (1,2)
Dense block	132×1×374	Dense layer × 6, k=12
Transition layer	66×1×187	Conv 1×1 with stride 1, padding 0 Avepool 1×2 with stride (1,2)
Dense block	138×1×187	Dense Layer × 6, k=12
Dual attention	138×1×187	Dual Attention modules
Pooling layer	138×1	Avepool 1×187 with stride (1,187)
Classification	2	Fully connected layer, softmax

module is illustrated in the middle part of Fig. 1. Notably, the channel attention models the inter-channel relationship of features, which is complementary to the position attention. To this end, we first flatten $\mathbf{X} \in \mathbb{R}^{C\times 1\times W}$ to obtain three identical feature representations $F(\mathbf{X}) \in \mathbb{R}^{C\times W}$ as *query*, *key*, and *value* matrices. Similar to the position attention, the similarity matrix $\mathbf{A_c} \in \mathbb{R}^{C\times C}$ between the inter-channel features is modeled by the similarity between the query-key pairs $(F(\mathbf{X}))^\top$ and $F(\mathbf{X})$. Then the attention scores $\hat{\mathbf{X}}_\mathbf{c} \in \mathbb{R}^{C\times W}$ are computed as the weighted sum with the *value* matrix $F(\mathbf{X})$. Finally, a learnable parameter β is used to control the weight of $\hat{\mathbf{X}}_\mathbf{c}$ and $\mathbf{X}$ in the residual connection. Thus, the channel attention computation can be written as

$$\begin{aligned}\mathbf{A_c} &= \text{softmax}\left(F(\mathbf{X})(F(\mathbf{X}))^\top\right),\\ \hat{\mathbf{X}}_\mathbf{c} &= \mathbf{A}_c \cdot F(\mathbf{X}),\\ \mathbf{E_c} &= \mathbf{X} + \beta U(\hat{\mathbf{X}}_\mathbf{c}).\end{aligned} \tag{3}$$

C. Final Decision Rule

Our densely connected dual-attention model classifies based on three-second heart sound shards, and then we employ a majority voting strategy to generate a final diagnosis of the entire heart sound recording. In this voting mechanism, we count the number of shards marked in each record and set an anomaly threshold to determine the class of the entire heart sound recording. More concretely, if the number of shards in the abnormal category is not less than 40% of the total number of shards among one heart sound record, the record can be diagnosed as abnormal, otherwise, it is normal.

D. DDA Model Overview

Following the basic configuration of DenseNet [62], we build our proposed DDA architecture. As depicted in Table I, DDA is mainly stacked by 4 dense blocks and transition blocks between them, along with a dual attention block. The input of our model is 1×1×6000, where the first '1' represents the channel dimension and the latter '1 × 6000' represents the position dimension of heart sound signals. A convolution layer with a relatively larger kernel size (i.e., 1 × 7) is first applied to capture abundant low-level features, instead of directly feeding the input signals into the first block. Then, a max pooling layer is adopted to downsample the feature maps. Sequentially, dense blocks with 6 dense layers and a growth rate of 12 are utilized to extract hierarchical features. In the following, to avoid large-scale parameters, the channel and position dimensions of feature maps are reduced to half through the transition layer. After 4 dense blocks, the feature maps are fed into a dual self-attention module to extract rich contextual features. Finally, a fully connected layer with a sigmoid activation function is used for classification. The weighted cross-entropy (WCE) loss function we adopt is defined as follows:

$$\begin{aligned}\mathcal{L}(\mathbf{x}_o, y) = &-w_+ \cdot y \log p(Y=1|X=\mathbf{x}_o)\\ &-w_- \cdot (1-y)\log p(Y=0|X=\mathbf{x}_o),\end{aligned} \tag{4}$$

where $\mathbf{x}_o$ denotes the network output, y denotes the class label, and $p(Y=i|X=\mathbf{x}_o)$ is the probability that the network assigns to the label $i, i \in \{0,1\}$. In our experiments, the weights of positive (abnormal) and negative (normal) classes are set as $\omega_+ = 0.8$ and $\omega_- = 0.2$ respectively, since the ratio of abnormal samples to normal samples is approximately $\frac{1}{4}$. Note that, the model size of our DDA is only 0.23 M.

IV. EXPERIMENTS AND RESULTS

In this section, we conduct extensive experiments to verify the heart sound classification ability of the proposed DDA model. We first introduce the public datasets used in our experiments. Then, we detail the experimental settings and evaluation metrics. Moreover, we experiment on the challenging benchmark with the 10-fold cross-validation to evaluate the effectiveness of DDA. Finally, we provide comprehensive ablation studies to analyze DDA thoroughly.

A. Dataset

We conduct classification experiments on the challenging PhysioNet/CinC 2016 heart sound dataset. PhysioNet/CinC 2016 is a large-scale heart sound dataset containing 3240 annotated heart sound records taken from 764 subjects, including 2575 normal heart sound records and 665 abnormal heart sound records. Each PCG record lasts from 5 seconds to 120 seconds. All recordings are resampled to 2000 Hz.

B. Experiment Settings

In our experiment, the model architecture has 4 operations: 4 dense blocks, 3 transition layers, dual attention modules, and a fully connected layer. Each dense block has 6 dense layers with a growth rate of 12. The training setting follows the common practice as in [33], [62], [63], [64]. Specifically, the model is trained for 250 epochs with batch size 64. Stochastic gradient

TABLE II
COMPARISON OF OUR DDA AND OTHER ADVANCED METHODS ACROSS STRATIFIED 10-FOLD CROSS-VALIDATION ON PHYSIONET/CINC 2016

Method	Params (M)	Acc. (%)	Score (%)	Precise Segmentation	Input Feature	End-to-end
1D CNN [26]	0.19	89.33	84.45	No	1D filtered signals	Yes
MFCC-CNN [25]	12.41	93.31	88.91	Yes	MFCC features	No
PSD-CNN [23]	0.24	89.05	86.26	No	Spectrograms	No
AdaBoost-CNN [24]	-	-	85.00	Yes	Time and frequency features, decomposed four frequency bands	No
DRGE [27]	-	-	90.00	No	1D raw signals	Yes
1D Dense [33]	0.11	93.56	90.51	No	1D raw signals	Yes
1D Clique [28]	0.19	93.28	90.69	No	1D raw signals	Yes
LSTM [50]	-	-	92.35	No	Spectrograms, temporal quasi-periodic features	No
DDA (ours)	0.23	**95.15**	**95.24**	No	1D raw signals	Yes

The bold values highlight the highest value(s) of each metric.

descent (SGD) is employed with an initial learning rate of 0.1, momentum of 0.9, and weight decay of 10^{-5}. The dropout of 0.1 is set for regularization. A cosine annealing is used to schedule the learning rate with a minimum learning rate of 10^{-4}. Considering the imbalanced data set, a cross-entropy loss with a weight of 0.8-0.2 (abnormal to normal) is adopted as the loss function as is common practice [33], [49]. Following [65], [66], [67], we employ Synthetic Minority Over-sampling Technique (SMOTE) [68] as the sampling strategy. SMOTE is a combination of over-sampling the minority (abnormal) class and under-sampling the majority (normal) class. Particularly, SMOTE is only used for the training process and not for the validation process. Empirically, the voting mechanism judges the input heart sound signal to be abnormal when the fraction of abnormal shards exceeds 40% of the total number of heart sound shards.

To show the generalization of our proposed DDA, we adopt the stratified 10-fold cross-validation strategy. We divide the whole dataset with 3240 heart sounds into ten equal parts, each time one is taken as the test set, and the rest is used as the training set. We empirically segment the 1D waveform heart sound recordings into three-second length shards with a stride of one second as the input of our proposed heart sound diagnosis system as is common practice.

In our experiments, we adopt six widely used metrics to evaluate the proposed model: 1) accuracy (Acc.), which measures the classification performance; 2) sensitivity (Sen.) and 3) specificity (Spe.), which indicate the proportion of correctly identified positives and negatives, respectively; 4) the overall score (Score), also known as Macc, which describes the comprehensive diagnosis performance by calculating the average of sensitivity and specificity; 5) Area under the ROC curve (AUC), which evaluates the performance of classification tasks by calculating the area under the receiver operating characteristic curve (ROC); 6) the trainable parameters (Params) of models, which are also calculated to evaluate the model size. Particularly, based on the Cinc2016 competition, the correctly diagnosed abnormal heart sound recordings are taken as true positive samples in this work.

C. Comparison With Current Advanced Methods

Table II presents the quantitative comparisons of our proposed DDA and previous state-of-the-art methods across stratified 10-fold cross-validation on PhysioNet/CinC 2016, where our DDA outperforms other advanced methods. Specifically, DDA brings 6.10% accuracy gains compared with PSD-CNN [23] using few parameters. Our DDA also significantly boosts the accuracy by 1.59% and score by 4.73% compared with 1D Dense [33]. Without any pre-processing, our DDA can gain up to 2.89% absolute improvement on score compared with long short-term memory networks (LSTM) [50]. Additionally, our DDA brings in a gain of 6.3% score with significantly few parameters (0.23 M *vs*. 12.41 M) compared with MFCC-CNN [25], which transforms 1D heart sound signals into 2D time-frequency features. By contrast, only a simple yet effective sliding window-based segmentation strategy with a fixed size is required in our DDA. The detail of stratified 10-fold cross-validation results is listed in Table III. Concretely, the proposed DDA achieves the highest accuracy of 97.22% and the overall score of 96.92% on the 2nd fold cross-validation. These results convincingly verify the effectiveness of our model. Moreover, we further analyze the advantages and disadvantages of our proposed DDA in Section V.

TABLE III
RESULTS OF OUR DDA ACROSS STRATIFIED 10-FOLD CROSS-VALIDATION ON PHYSIONET/CINC 2016

#fold	Acc. (%)	Score (%)	Sen. (%)	Spe. (%)
0	96.91	96.69	**100.00**	93.39
1	**97.22**	**96.92**	98.50	**95.33**
2	95.06	95.04	95.52	94.55
3	94.75	94.78	98.51	91.05
4	93.52	93.68	95.52	91.83
5	94.75	95.74	**100.00**	91.47
6	94.75	94.80	96.97	92.64
7	93.52	94.40	98.49	90.31
8	95.37	95.38	96.97	93.80
9	95.68	95.01	95.46	94.57
Average	95.15	95.24	97.59	92.89

The bold values highlight the highest value(s) of each metric.

D. Ablation Studies

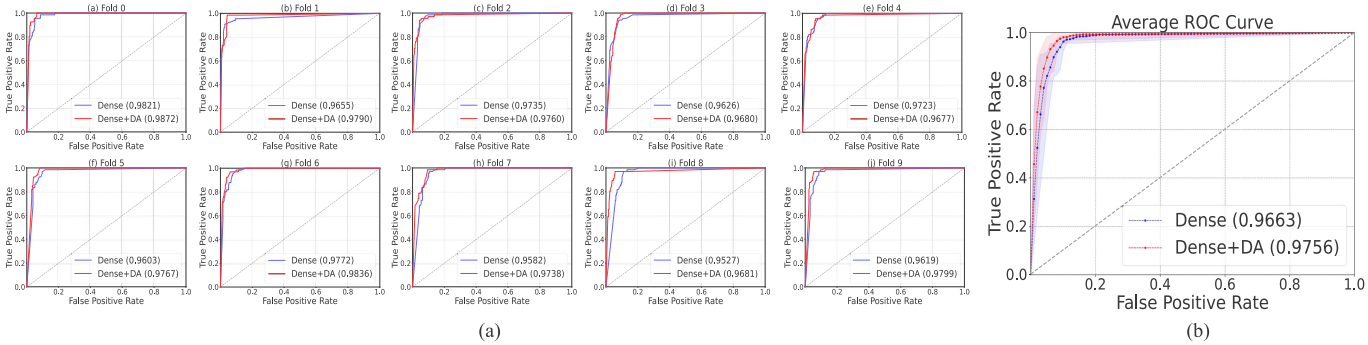
In this section, we perform comprehensive ablation studies to analyze our proposed DDA thoroughly on the PhysioNet/CinC 2016 dataset.

1) Effect of Dual Attention Modules: To verify the effectiveness of the dual attention (DA) modules for heart sound classification, we arrange an experiment on different densely connected backbones with varying network depths. Table IV lists the results on varying dense blocks, dense layers (layers

TABLE IV
ABLATION STUDY ON THE EFFECT OF DUAL ATTENTION (DA) MODULES UNDER DIFFERENT NETWORK DEPTHS

Architecture	Model	Params (M)	Accuracy (%)	Score (%)	Sensitivity (%)	Specificity (%)
[6, 6, 6, 6]*	Dense-53	0.14	92.90	92.33	93.94	90.72
	Dense-53 + DA	0.22	93.98	93.96	96.69	91.22
[6, 6, 6]	Dense-58	0.12	93.24	92.45	93.23	91.65
	Dense-58 + DA	0.17	94.35	94.21	96.39	92.03
[6, 6, 6, 6]	Dense-77	0.15	93.89	93.55	95.64	91.46
	Dense-77 + DA (ours)	0.23	**95.15**	**95.24**	**97.59**	92.89
[6, 6, 6, 6]	Dense-77 ($k = 24$)	0.57	93.07	92.61	94.39	90.84
	Dense-77 + DA ($k = 24$)	0.73	94.20	93.66	95.19	92.12
[6, 6, 6, 6, 6]	Dense-96	0.22	93.91	93.64	94.80	92.88
	Dense-96 + DA	0.29	95.06	94.53	95.19	**93.87**
[6, 12, 24, 16]	Dense-179	0.78	93.89	93.93	96.08	91.77
	Dense-179 + DA	1.10	94.63	94.36	95.79	92.93
[6, 12, 36, 24]	Dense-239	1.40	93.64	93.74	95.93	91.54
	Dense-239 + DA	2.07	94.57	94.20	96.09	92.31

We report the average performance across stratified 10-fold cross-validation on PhysioNet/CinC 2016. Specifically, we vary the overall number of dense blocks, dense layers, and growth rates k. For example, [6, 6, 6, 6] denotes the DenseNet architecture comprised of 4 stacked dense blocks, each of which has 6 dense layers with a growth rate (k) of 12. Particularly, [•, . . . , •]* denotes the original DenseNet architecture using two sequential convolutional layers in each dense layer with the kernel sizes of 1×1 and 1×3. Likewise, [•, . . . , •] denotes our used DenseNet architecture using three sequential convolutional layers in each dense layer with the kernel sizes of 1×1,1×3, 1×1. The model with DA is highlighted with a grey background.
The bold values highlight the highest value(s) of each metric.



(a) (b)

Fig. 3. Receiver operating characteristics (ROC) curves of heart sound classification performance across stratified 10-fold cross-validation on PhysioNet/CinC 2016 (a). Particularly, we show the average result of 10-fold cross-validation on the rightmost part (b). The areas Under the ROC curves are shown in parentheses in the legend. The ROC curves fluctuation range of our proposed DDA (Dense+DA) model is marked in the transparent red background. Likewise, the ROC curves fluctuation range of the DenseNet model without dual attention (Dense) is marked in the transparent blue background.

in each dense block), and growth rates. For example, $[6, 6, 6, 6]$ denotes the DenseNet architecture comprised of 4 stacked dense blocks, each of which has 6 dense layers with a growth rate (k) of 12. As shown in Table IV, under the dense network framework with various parameters, the dual attention modules consistently play a major role in heart sound classification. Compared with no dual attention modules, our DDA (Dense-77 + DA) brings in a significant average gain of 1.26% accuracy, 1.69% score, 1.95% sensitivity, and 1.43% specificity across stratified 10-fold cross-validation. Notably, compared with other network configurations (including the ones of the original DenseNet [62] in the 11th and 12th lines of Table IV), "Dense-77 + DA" achieves the best performance, thus we set the dense layers in each dense block as 6 if no special illustration. In addition, the stratified 10-fold cross-validation results of ROC curves are illustrated in Fig. 3. For the metric of the AUC (Area under the ROC curve), we can find that our DDA obtains performance gains by up to 0.0093 AUC on average, demonstrating the effectiveness of the DA module. Regarding each fold, our DDA model works better than the baseline in most folds. For example, our DDA model brings 0.018 AUC benefits at the 9th fold. Our DDA model also achieves the best results with a small number of model parameters as 0.23 M, since small models often have insufficient expression ability, while large models may cause overfitting.

Moreover, we conduct a comparative experiment across stratified 10-fold cross-validation to compare with two single-attention models, adding the CAM module and PAM module to the baseline ("Dense-77") respectively. As listed in Table V, the baseline only achieves an accuracy of 93.89% and a score of 93.55%. The attention mechanisms, CAM and PAM, can model long-range dependencies in terms of position and channel dimensions, which is beneficial for accurate heart sound classification compared with the baseline. This is reasonable since CAM and PAM increase feature reuse to enhance the representation ability. Notably, our DDA ("Dense + DA") considerably surpasses two single-attention models, i.e., over "Dense + PAM" by

TABLE V
ABLATION STUDY ON THE EFFECT OF DUAL ATTENTION (DA) MODULES UNDER DIFFERENT ATTENTION MECHANISMS

Method	Acc. (%)	Score (%)	Sensitivity (%)	Specificity (%)
Dense	93.89	93.55	95.65	91.46
Dense + PAM	94.29	94.19	95.64	92.74
Dense + CAM	94.70	94.44	95.64	**93.24**
Dense + DA	**95.15**	**95.24**	**97.59**	92.89

We report the average performance across stratified 10-fold cross-validation on PhysioNet/CinC 2016. ``Dense" denotes the DenseNet model without the attention mechanism.
The bold values highlight the highest value(s) of each metric.

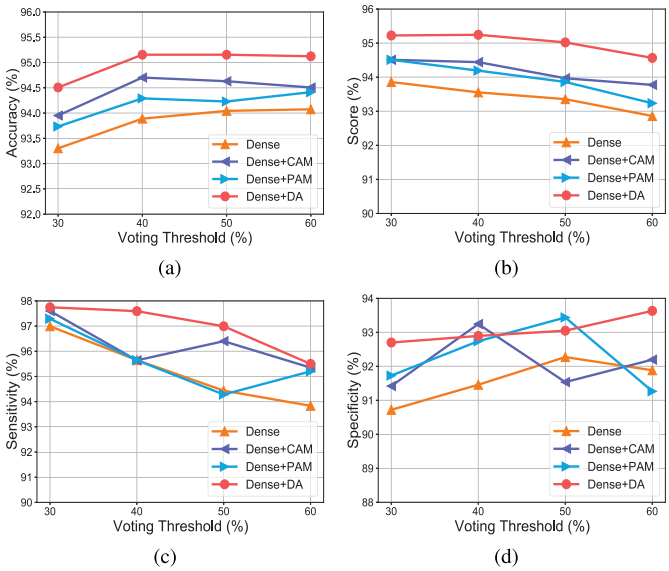


Fig. 4. Ablation study on different thresholds of the voting mechanism on four indicators including (a) accuracy, (b) score, (c) sensitivity, and (d) specificity. Specifically, the thresholds vary from 30% to 60%. We report the average performance across stratified 10-fold cross-validation on PhysioNet/CinC 2016. "Dense" denotes the DenseNet model without the attention mechanism.

0.86% accuracy and over "Dense + CAM" by 0.80% score. DDA also brings in absolute improvements of nearly 2% sensitivity gains compared with "Dense + PAM" and "Dense + CAM". This convincingly verifies the effectiveness of our proposed DDA with both position and channel relation learning.

2) Influence of the Threshold: To investigate the influence of the threshold, we carry out an experiment across stratified 10-fold cross-validation by setting the threshold of the voting mechanism as 30%, 40%, 50%, and 60%, which controls the fraction of abnormal shards in the total heart sound shards for judging whether to be abnormal. The results in Fig. 4 show that, as the threshold changes from 30% to 60%, our DDA can consistently achieve better accuracy and score compared with the baseline model (Dense-77). Furthermore, a higher threshold means that a larger proportion of abnormal segments is required to make the whole heart sound judged to be abnormal. According to Fig. 4, we empirically set the threshold as 40% on our DDA for the best performance.

TABLE VI
ABLATION STUDY ON THE SAMPLING APPROACH SMOTE OF OUR DDA ACROSS STRATIFIED 10-FOLD CROSS-VALIDATION ON PHYSIONET/CINC 2016

#fold	w/o SMOTE		SMOTE	
	Acc. (%)	Score (%)	Acc. (%)	Score (%)
0	95.99	95.43	96.91	96.69
1	95.37	93.77	**97.22**	**96.92**
2	94.75	92.80	95.06	95.04
3	93.21	93.68	94.75	94.78
4	93.52	92.15	93.52	93.68
5	94.14	94.80	94.75	95.74
6	95.06	95.00	94.75	94.80
7	91.98	93.45	93.52	94.40
8	93.83	92.92	95.37	95.38
9	94.14	93.50	95.68	95.01
Average	94.20	93.75	**95.15**	**95.24**

W/o SMOTE: our DDA is trained on the original training data distribution, without SMOTE.
The bold values highlight the highest value(s) of each metric.

3) Impact of Sampling Strategy: Class imbalance is one of the common problems during classifier model training, especially in disease research. The sampling method is only adopted during the training process, which can not destroy the sample distribution of the test set and be in line with the practical scenario. We conduct an experiment across stratified 10-fold cross-validation to analyze the impact of sampling on the heart sound classification. The sampling strategy we use is SMOTE [68]. Note that, as shown in Table VI, except for the 6th fold, the accuracy and score have been significantly improved after applying SMOTE under the other folds. Particularly, our DDA obtains the best accuracy of 97.22% and a score of 96.92% under the first fold. Moreover, the absolute performance gains of accuracy and score brought by SMOTE can be over 1% under the 1st, 3 rd, 4th, 7th, 8th, and 9th fold cross-validation.

V. DISCUSSION

In this section, we summarize the advantages and disadvantages of the proposed DDA. First of all, our DDA only requires a simple yet effective sliding window operation for shard-level segmentation, without the need to elaborately design the precise segmentation algorithms. Second, our DDA takes the 1D raw signals as the input, which can omit the extra transforming procedure from 1D raw signals to 2D representations. Third, our DDA performs end-to-end learning from heart sound signals, instead of a complex ensemble system that manually combines multiple features. Last but not least, our DDA obtains strong performance on the challenging PhysioNet/CinC 2016 benchmark. As shown in Table II, the proposed DDA achieves the best trade-off between performance and parameters compared with current advanced methods. Technically, benefiting from the dual attention mechanism, our DDA can automatically capture discriminative features for heart sound classification, without requiring manual pre-processing for selecting specified features. For convenience, we list the partial advantages of our DDA compared with current advanced methods in Table II.

Understanding the disadvantages of our approach is also critical for improving it. The first disadvantage is that the parameter efficiency of DDA is not the best compared with other state-of-the-art methods, as shown in Table II. In the future, it will be interesting to design a model compression method compatible with our dual attention mechanism to further reduce the model parameters. Besides, our DDA requires more training time overheads than those based on machine learning. Another interesting future work is to speed up the training procedure of the proposed DDA, such as improving the optimization algorithm.

VI. Conclusion

In this work, we present DDA, a lightweight densely connected dual attention mechanism network. DDA is capable of combining the purely end-to-end training pattern and enhanced feature representations based on the self-attention mechanism. For this purpose, a densely connected structure is introduced to improve the information flow with additional direct accesses. Meanwhile, a dual attention mechanism (DA) is developed to capture feature dependencies in the position and channel dimensions, respectively. Technically, the position and channel attention modules are introduced to parallelly integrate local features based on the feature similarities, which are calculated by capturing the dependencies between any two features along position and channel axes, respectively. In this way, DDA can significantly increase contextual modeling capabilities to tackle heart sound classification suffering from inherent murmurs and limited supervised samples while keeping the end-to-end training merit of deep learning. Comprehensive experiments on the challenging Cinc2016 benchmark, as well as thorough ablation studies, have demonstrated the effectiveness of DDA on the heart sound classification task with state-of-the-art trade-offs between performance and parameter efficiency.

Acknowledgment

We would like to thank our group students Sheng Gui, Shiyang Bai, and Yongliang Lv for their valuable discussions.

References

[1] WHO, "Cardiovascular diseases (CVDs)," Accessed: Jun. 11, 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] C. Liu et al., "An open access database for the evaluation of heart sound algorithms," *Physiol. Meas.*, vol. 37, no. 12, 2016, Art. no. 2181.

[3] S. Kiranyaz, T. Ince, and M. Gabbouj, "Real-time patient-specific ECG classification by 1-D convolutional neural networks," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 3, pp. 664–675, Mar. 2016.

[4] T. Choudhary, L. N. Sharma, and M. K. Bhuyan, "Heart sound extraction from sternal seismocardiographic signal," *IEEE Signal Process. Lett.*, vol. 25, no. 4, pp. 482–486, Apr. 2018.

[5] Y. Zhang, L. Sun, H. Song, and X. Cao, "Ubiquitous WSN for healthcare: Recent advances and future prospects," *IEEE Internet Things J.*, vol. 1, no. 4, pp. 311–318, Aug. 2014.

[6] 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.

[7] A. K. Dwivedi, S. A. Imtiaz, and E. Rodriguez-Villegas, "Algorithms for automatic analysis and classification of heart sounds–a systematic review," *IEEE Access*, vol. 7, pp. 8316–8345, 2019.

[8] S. L. Oh et al., "Classification of heart sound signals using a novel deep wavenet model," *Comput. Methods Programs Biomed.*, vol. 196, 2020, Art. no. 105604.

[9] A. Leatham, *Auscultation of the Heart and Phonocardiography*. London, U.K.: Churchill, 1970.

[10] Z. Abduh, E. A. Nehary, M. A. Wahed, and Y. M. Kadah, "Classification of heart sounds using fractional fourier transform based mel-frequency spectral coefficients and traditional classifiers," *Biomed. Signal Process. Control*, vol. 57, 2020, Art. no. 101788.

[11] 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.

[12] D. M. Nogueira, C. A. Ferreira, E. F. Gomes, and A. M. Jorge, "Classifying heart sounds using images of motifs, MFCC and temporal features," *J. Med. Syst.*, vol. 43, no. 6, pp. 1–13, 2019.

[13] 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.

[14] R. Saraçoğlu, "Hidden markov model-based classification of heart valve disease with PCA for dimension reduction," *Eng. Appl. Artif. Intell.*, vol. 25, no. 7, pp. 1523–1528, 2012.

[15] S. Sun, H. Wang, Z. Jiang, Y. Fang, and T. Tao, "Segmentation-based heart sound feature extraction combined with classifier models for a vsd diagnosis system," *Expert Syst. Appl.*, vol. 41, no. 4, pp. 1769–1780, 2014.

[16] H. Uğuz, A. Arslan, and İ. Türkoğlu, "A biomedical system based on hidden Markov model for diagnosis of the heart valve diseases," *Pattern Recognit. Lett.*, vol. 28, no. 4, pp. 395–404, 2007.

[17] L. Avendano-Valencia, J. Godino-Llorente, M. Blanco-Velasco, and G. Castellanos-Dominguez, "Feature extraction from parametric time–frequency representations for heart murmur detection," *Ann. Biomed. Eng.*, vol. 38, no. 8, pp. 2716–2732, 2010.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Sys. 25: 26th Annu. Conf. Neural Inf. Process. Syst.*, Lake Tahoe, Nevada, USA, 2012, pp. 1106–1114. [Online]. Available: https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

[19] R. Faster, "Towards real-time object detection with region proposal networks," in *Proc. Neural Inf. Process. Syst.*, 2015, vol. 9199, no. 10.5555, pp. 2969239–2969250.

[20] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6202–6211.

[21] A. v. d. Oord et al., "WaveNet: A generative model for raw audio," in *Proc. 9th ISCA Speech Synth. Workshop*, Sunnyvale, CA, USA, Sep. 13–15, 2016, p. 125. [Online]. Available: http://www.isca-speech.org/archive/SSW%5C_2016/abstracts/ssw9%5C_DS-4%5C_van%5C_den%5C_Oord.html

[22] S. Alam, R. Banerjee, and S. Bandyopadhyay, "Murmur detection using parallel recurrent & convolutional neural networks," 2018, *arXiv:1808.04411*.

[23] 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.

[24] 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, pp. 621–624.

[25] 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.

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

[27] 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.

[28] Y. Xu, B. Xiao, X. Bi, W. Li, J. Zhang, and X. Ma, "Pay more attention with fewer parameters: A novel 1-D convolutional neural network for heart sounds classification," in *Proc. IEEE Comput. Cardiol. Conf.*, 2018, pp. 1–4.

[29] B. Bozkurt, I. Germanakis, and Y. Stylianou, "A study of time-frequency features for CNN-based automatic heart sound classification for pathology detection," *Comput. Biol. Med.*, vol. 100, pp. 132–143, 2018.

[30] T. Alafif, M. Boulares, A. Barnawi, T. Alafif, H. Althobaiti, and A. Alferaidi, "Normal and abnormal heart rates recognition using transfer learning," in *Proc. IEEE 12th Int. Conf. Knowl. Syst. Eng.*, 2020, pp. 275–280.
[31] 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.
[32] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, Aug. 1980.
[33] 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.
[34] 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.
[35] C. D. Papadaniil and L. J. Hadjileontiadis, "Efficient heart sound segmentation and extraction using ensemble empirical mode decomposition and kurtosis features," *IEEE J. Biomed. Health Inform.*, vol. 18, no. 4, pp. 1138–1152, Jul. 2014.
[36] H. Tang, T. Li, T. Qiu, and Y. Park, "Segmentation of heart sounds based on dynamic clustering," *Biomed. Signal Process. Control*, vol. 7, no. 5, pp. 509–516, 2012.
[37] 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.
[38] G. D. Clifford et al., "Recent advances in heart sound analysis," *Physiol. Meas.*, vol. 38, pp. E10–E25, 2017.
[39] S. Ari, K. Hembram, and G. Saha, "Detection of cardiac abnormality from PCG signal using LMS based least square SVM classifier," *Expert Syst. Appl.*, vol. 37, no. 12, pp. 8019–8026, 2010.
[40] 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.
[41] Z. Abduh, E. A. Nehary, M. A. Wahed, and Y. M. Kadah, "Classification of heart sounds using fractional fourier transform based mel-frequency spectral coefficients and stacked autoencoder deep neural network," *J. Med. Imag. Health Inform.*, vol. 9, no. 1, pp. 1–8, 2019.
[42] M. Hamidi, H. Ghassemian, and M. Imani, "Classification of heart sound signal using curve fitting and fractal dimension," *Biomed. Signal Process. Control*, vol. 39, pp. 351–359, 2018.
[43] W. Zhang, J. Han, and S. Deng, "Heart sound classification based on scaled spectrogram and tensor decomposition," *Expert Syst. Appl.*, vol. 84, pp. 220–231, 2017.
[44] M. Zabihi, A. B. Rad, S. Kiranyaz, M. Gabbouj, and A. K. Katsaggelos, "Heart sound anomaly and quality detection using ensemble of neural networks without segmentation," in *Proc. IEEE Comput. Cardiol. Conf.*, 2016, pp. 613–616.
[45] W. Chen et al., "Phonocardiogram classification using deep convolutional neural networks with majority vote strategy," *J. Med. Imag. Health Inform.*, vol. 9, no. 8, pp. 1692–1704, 2019.
[46] V. Maknickas and A. Maknickas, "Recognition of normal–abnormal phonocardiographic signals using deep convolutional neural networks and mel-frequency spectral coefficients," *Physiol. Meas.*, vol. 38, no. 8, 2017, Art. no. 1671.
[47] A. I. Humayun, S. Ghaffarzadegan, Z. Feng, and T. Hasan, "Learning front-end filter-bank parameters using convolutional neural networks for abnormal heart sound detection," in *Proc. IEEE 40th Annu. Int. Conf. Eng. Med. Biol. Soc.*, 2018, pp. 1408–1411.
[48] 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 Inform.*, vol. 24, no. 8, pp. 2189–2198, Aug. 2020.
[49] F. Li, H. Tang, S. Shang, K. Mathiak, and F. Cong, "Classification of heart sounds using convolutional neural network," *Appl. Sci.*, vol. 10, no. 11, 2020, Art. no. 3956.
[50] W. Zhang, J. Han, and S. Deng, "Abnormal heart sound detection using temporal quasi-periodic features and long short-term memory without segmentation," *Biomed. Signal Process. Control*, vol. 53, 2019, Art. no. 101560.
[51] J. Fu et al., "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.
[52] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
[53] J. Hu, L. Shen, S. Albanie, G. Sun, and A. Vedaldi, "Gather-excite: Exploiting feature context in convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. 31: Annu. Conf. Neural Inf. Process. Syst.*, Montréal, Canada, Dec. 3–8, 2018, pp. 9423–9433. [Online]. Available: https://proceedings.neurips.cc/paper/2018/hash/dc363817786ff182b7bc59565d864523-Abstract.html
[54] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 603–612.
[55] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.
[56] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
[57] X. Liu, K. Li, and K. Li, "Attentive semantic and perceptual faces completion using self-attention generative adversarial networks," *Neural Process. Lett.*, vol. 51, no. 1, pp. 211–229, 2020.
[58] I. Bello, B. Zoph, Q. V. Le, A. Vaswani, and J. Shlens, "Attention augmented convolutional networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3286–3295.
[59] M. Lin, Q. Chen, and S. Yan, "Network in network," in *Proc. 2nd Int. Conf. Learn. Representations*, Banff, AB, Canada, Apr. 14–16, 2014.
[60] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.
[61] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. AAlemi, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," in *Proc. 31st AAAI Conf. Artif. Intell.*, San Francisco, CA, USA, Feb. 4–9, 2017, pp. 4278–4284. [Online]. Available: http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14806
[62] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.
[63] Z. Hou et al., "Chex: Channel exploration for cnn model compression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12287–12298.
[64] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.
[65] R. F. Ibarra-Hernández, N. Bertin, M. A. Alonso-Arévalo, and H. A. Guillén-Ramírez, "A benchmark of heart sound classification systems based on sparse decompositions," *Proc. SPIE*, vol. 10975, pp. 26–38, 2018.
[66] M. N. Homsi et al., "Automatic heart sound recording classification using a nested set of ensemble algorithms," in *Proc. IEEE Comput. Cardiol. Conf.*, 2016, pp. 817–820.
[67] B. Unnikrishnan, P. R. Singh, X. Yang, and M. C. H. Chua, "Semi-supervised and unsupervised methods for heart sounds classification in restricted data environments," 2020, *arXiv:2006.02610*.
[68] F. Last, G. Douzas, and F. Bacao, "Oversampling for imbalanced learning based on k-means and smote," 2017, *arXiv:1711.00837*.