IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

Efficient anomaly detection algorithm for heart sound signal

ZHIHAI LIU[1], WEN LIU[2,3], ZHENG GU[2], and Feng Yang[4]

[1]School of Computer Science and Technology, Xinjiang Normal University, 830000 Urumqi, China
[2]Artificial Intelligence and Smart Mine Engineering Technology Center, Xinjiang Institute of Engineering, Urumqi, China
[3]Xinjiang Changsen Data Technology Co., Ltd, Urumqi 830011, China
[4]Computer Information Center, Xinjiang Institute of Engineering, Urumqi, China

Corresponding author: Wen Liu (e-mail: 627952@qq.com).

This work is supported by Tianshan Talent of Xinjiang Uygur Autonomous Region - Young Top Talents in Science and Technology (2022TSYCCY0008) and NSFC under grant 61962058.

ABSTRACT According to the latest report by the WHO, cardiovascular disease claims approximately 17.9 million lives annually, making it one of the leading causes of mortality [1]. Hence, early screening and detection of cardiovascular diseases are important for their prevention. Heart sound signals contain a wealth of information on cardiac function and health status. Researchers have recently utilized deep learning methods to detect abnormal features in heart sound signals, thereby facilitating disease diagnosis. Currently, existing heart sound datasets suffer from imbalanced data proportions, complex feature types, and low discriminative power between systolic and diastolic murmurs, resulting in the suboptimal performance of deep learning algorithms in detection. Therefore, we propose a heart sound abnormality detection algorithm based on the Swin Transformer architecture. Firstly, we enhance the ability to extract local texture features of heart sound signals by introducing a convolutional embedding module into the positional encoding layer of the backbone network. Second, we augmented the model's capability to extract the frequency-domain features of heart-sound signals by incorporating a discrete convolutional mapping structure. This structure utilizes discrete cosine transformation in conjunction with convolutional projection to acquire feature matrices, thereby improving classification accuracy. Finally, we employed a Focal Loss function to prioritize abnormal heart-sound samples, enhancing the generalization ability of the model and evaluating the proposed algorithm using the PhysionNet/CinC 2016 public dataset. The results demonstrated an Accuracy of 93.4%, a Specificity of 90.4% and a Sensitivity of 95.7%.

INDEX TERMS abnormality detection; cardiovascular disease; convolution embedding; discrete convolution projection; heart sound signal

I. INTRODUCTION

CARDIOVASCULAR diseases (CVDs) remain the leading cause of morbidity and mortality worldwide, posing a significant burden on healthcare systems [2]. Early detection and diagnosis are crucial for improving patient outcomes, especially in resource-limited environments with restricted access to medical equipment. Auscultation, listening to heart sounds using a stethoscope, is a simple yet effective tool for assessing cardiac health. However, the accuracy of auscultation largely depends on the experience and skill of the healthcare professional. While cardiologists can achieve high diagnostic accuracy, it is often challenging for general practitioners and non-specialists to make precise diagnoses, which may lead to misdiagnosis or missed diagnoses of diseases [3] [4]. This limitation highlights the urgent need to develop an automated system capable of accurately and efficiently analyzing heart sounds, which would play a crucial role in the early detection of cardiovascular diseases. For example, literature [5] introduces an expert diagnosis system for early-stage heart disease based on fuzzy inference technology. This system demonstrates the potential of fuzzy logic in handling uncertainty and complexity and provides a new perspective for researchers in the field of heart sound classification. By applying fuzzy inference techniques to the processing and classification of heart sound signals, the inherent uncertainty and complexity of heart sound data can be more effectively addressed. This research direction helps improve automated heart sound analysis systems' performance and

offers a promising pathway for developing future cardiovascular disease diagnostic technologies.Artificial intelligence technology has recently been widely applied in medical research. For instance, literature [6] proposes a convolutional neural network (CNN)-based algorithm that utilizes CT scan images to identify lung nodules. The study demonstrates that this algorithm significantly improves the accuracy of lung nodule detection. This suggests that researchers could explore converting time-domain and frequency-domain features of heart sounds into images and apply artificial intelligence techniques to the automated classification of heart sound signals. Such an approach could assist doctors in more accurately assessing patients' cardiac health.By doing so, not only can diagnostic accuracy be improved, but the workload of physicians can also be reduced, which is particularly important in resource-limited environments where such systems can play a crucial role.

Heart sounds are generated by the mechanical activity of the heart during the cardiac cycle, and they contain vital information about cardiac health. A typical cardiac cycle consists of four phases (S1, S2, S3, S4), with S1 and S2 being the most diagnostically significant [7]. Abnormal heart sounds, such as murmurs, may indicate underlying pathologies, including valvular dysfunction and heart failure. Traditionally, signal processing techniques such as time-domain analysis and Mel Frequency Cepstral Coefficients (MFCC) have been employed to extract features from heart sound signals. However, these methods exhibit limitations in capturing the complex nonlinear characteristics of heart sounds, especially in noisy environments or when abnormalities are subtle.

In recent years, the rapid development of deep learning technologies has revolutionized several fields, including speech recognition, image classification, and medical signal analysis. Deep learning methods, particularly Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks, have demonstrated exceptional performance in handling complex, high-dimensional data tasks [8] [9] [10]. A vital advantage of these methods is their ability to automatically extract features from raw data, eliminating manual feature engineering commonly required in traditional machine learning approaches. Deep learning models have also shown vital classification accuracy and robustness in capturing the spatiotemporal patterns present in heart sound signals [11].

This study is motivated by two key factors. First, the increasing global burden of cardiovascular diseases (CVDs) demands more efficient diagnostic tools, especially in regions with scarce specialized cardiology services. Automated heart sound classification systems have the potential to make early CVD detection more accessible, reducing reliance on expert interpretation and enabling non-specialists to perform diagnoses. Second, the limitations of traditional signal processing techniques in analyzing complex heart sounds highlight the need for more advanced machine-learning models capable of learning directly from raw data. By leveraging the strengths of deep learning, this study aims to overcome the challenges of conventional methods and provide a more reliable and scalable solution for detecting abnormal heart sounds.

The primary objective of this study is to develop a deep learning-based framework for the automatic classification of abnormal heart sounds aimed at facilitating the early detection of cardiovascular diseases (CVDs). Specifically, this research proposes an innovative network architecture that integrates convolutional concepts with the Swin-Transformer, designed to capture both local patterns and long-range dependencies in heart sound signals. Unlike traditional methods that rely on manual segmentation and feature extraction, this approach processes raw heart sound signals directly, using an end-to-end framework to streamline the workflow and reduce the risk of information loss during intermediate steps. Ultimately, we propose the Dcv-Swin Transformer method for efficient heart sound anomaly detection. The critical contributions of our work are summarized as follows:

- Introduction of a Swin-Transformer-based Architecture for Abnormal Heart Sound Detection: This model leverages the hierarchical local attention mechanism of the Swin-Transformer to effectively capture both local and global features of heart sound signals.
- Design of a Fifth-Order Butterworth Filter: This filter significantly reduces the interference of heart sound noise on classification tasks by filtering out high-frequency noise.
- Development of a Convolutional Embedding Module: This module enhances the model's feature extraction capability by better capturing local features of heart sound signals while preserving the positional information of Mel-spectrogram features.
- Creation of a Convolutional Mapping Module with Discrete Cosine Transform: This module uses convolutions with varying strides to obtain the q, k, and v matrices for attention calculation. It reduces data dimensions and retains frequency domain information, thus improving the model's generalization ability.
- Incorporation of Focal Loss and Linear Interpolation: This approach effectively addresses the class imbalance between normal and abnormal heart sounds, enhancing the model's classification performance on imbalanced datasets.

In summary, this paper presents an effective method for detecting abnormal heart sounds to enhance cardiac diagnostics' accuracy and efficiency. The remainder of this paper is organized as follows: Section II reviews relevant work on heart sound classification and signal processing. Section III introduces the proposed method, including the network architecture and innovative aspects. Section IV describes the model training process, presents experimental results, and discusses performance metrics. Section V Provides an objective analysis and discussion of the strengths and limitations of our model, and proposes feasible solutions to address the current limitations and enhance model performance in future

work. Finally, Section VI concludes the paper and outlines potential directions for future research.

II. RELATED WORK

A comprehensive abnormality detection system for heart-sound signals typically consists of four main components:heart-sound denoising, heart-sound segmentation, heartsound feature extraction, and automatic classification of heart-sound signals. Over the past few decades, the automatic classification of heart-sound signals based on deeplearning models has seen significant advancements. Noman et al. [8] extracted 2D time-frequency features from heartsound signals to construct a 2D convolutional neural network(CNN) model for heart-sound classification. Walker et al. [12] utilized a Double Bayesian Resnet(DBRes) along with the time-frequency features of heart-sound signals to build a classification model. Moreover, representative recurrent neural networks (RNNs) and long short-term memory networks (LSTMs) are more effective in extracting temporal features from sequential heart-sound signals. Ahmad et al. [9] combined LSTM networks with discrete wavelet transformation and extracted the Mel Frequency Cepstral Coefficients (MFCCs) from heart-sound signals for classification. Deng et al. [10] fused CNNs with RNNs to form a single CNN. Liu et al. [13] introduced a Temporal Convolutional Network (TCN) that leverages dilated and causal convolutions, which are suitable for time-series data classification tasks. Li et al. [14] employed BiLSTM networks to learn the feature representations of time-series data while incorporating attention mechanisms for classification. Deperlioglu et al. [15] attempted to extract instantaneous energy from heart-sound signals as features and combined them with stacked autodecoder networks for heart-sound classification tasks. Banerjee et al. [16] used a 2D CNN combined with neural network model of a global average pooling layer to achieve five classifications of heart sound signals. Chen et al. [17] used a combination of CNN and wavelet transform to cut and classify heart sound signals. Duggento et al. [18] used a CNN model and the Mel epstral coefficient of the heart sound signal to identify abnormal heart sounds. Ranipa et al. [19] used a mutil-modal CNN fusion architecture to extract fusion architecture to extract frequency domain features of heart sound signals and classify heart sound signals. Xiao et al. [20] used a CNN that combined spatial and channel attention to improve the performance of CNN in classifying heart sound signals. Oh et al. [21] proposed a deep WaveNet model, which can be used to automatically classify heart sound signals. Gao et al. [22] used a model based on GRU and HMM to perform feature extraction and classification of heart sound signals without denoising. Shuvo et al. [23] proposed a lightweight model CardioXNet based on recurrent neural networks to achieve automatic recognition of heart sound signals. Alaskar et al. [24] applies AlexNet's pretrained model to the heart sound signal classification task to achieve automatic classification of heart sound signals. M [25] et al. utilized a transfer learning approach, employing log-mel spectrogram features and the MobileNet network for heart sound signal classification. Their study focused primarily on the time-domain characteristics of heart sound signals and leveraged transfer learning to enhance the model's generalization capability. Yadav [26] et al. combined convolutional neural networks (CNN) with bidirectional gated recurrent units (Bi-GRU), where CNNs were used to extract local features of the signals, while Bi-GRU captured long-term dependencies, facilitating the detection of abnormal heart sounds. Shuvo [27] et al. proposed a novel lightweight heart sound classification model, NRC-Net, optimized for heart sound signal classification by integrating convolutional recurrent neural networks (CRNN) with continuous wavelet transforms (CWT). Ballas et al. [28] developed a self-supervised learning method for automatic heart sound detection by training a CNN using normalized temperature-scaled cross-entropy loss. Lu et al. [29] introduced a model combining lightweight CNNs with random forests, utilizing Mel-spectrogram features as input. They applied noise augmentation and spectral enhancement techniques to improve the accuracy and generalization of the heart sound classification model. Fakhry et al. [30] compared three optimization algorithms—RMSprop, Adam, and stochastic gradient descent with momentum (SGDM)—in training a BiLSTM network, with results indicating that SGDM yielded the best classification performance. Lastly, Takezaki et al. [31] proposed two data augmentation methods: Window Slicing with Spectrogram (WSS) and Synthetic Spectrograms based on GANs (SSG). These methods were combined with CNNs to address the generalization challenges of sparse data in publicly available heart sound datasets.

In summary, although significant progress has been made in heart-sound signal classification based on deep learning in recent years, several unresolved issues and research gaps still exist. Firstly, in terms of noise robustness, many studies rely on preprocessing steps to mitigate the effects of noise, such as the time-frequency features and discrete wavelet transform (DWT) used in [8] and [9]. However, these preprocessing steps can increase system complexity and may not be feasible in real-time applications. Therefore, designing a deep learning model that can directly process raw signals without requiring complex preprocessing remains a challenge. Secondly, while some studies, such as [15], have attempted to extract instantaneous energy as features, most work focuses on time-domain and frequency-domain features. There is still insufficient exploration of the nonlinear characteristics and complex spatiotemporal patterns inherent in heart-sound signals. This indicates the necessity of further developing new feature representation methods that can capture the intrinsic nonlinear properties of heart-sound signals.Lastly, although some existing models, such as those proposed in [22] and [23], can reduce the impact of noise to some extent, their robustness and generalization capabilities when dealing with large-scale and diverse clinical datasets have yet to be fully validated. Consequently, despite the variety of deep learning models applied to heart-sound signal classification, there are still significant research gaps that need to be addressed in

developing efficient noise reduction techniques and achieving accurate end-to-end automated heart-sound anomaly detection systems. Addressing these gaps can significantly advance the development of reliable real-time heart-sound analysis systems that can function effectively in diverse clinical environments.

Although deep-learning models have achieved high accuracy in heart-sound signal classification tasks, they still face several challenges. First, there is a lack of sufficient and balanced publicly available datasets, particularly in terms of the disproportionate ratio between normal and abnormal heart sound samples, which affects the ability of the model to identify accurately abnormal heart sounds. Second, the diversity and complexity of the features in heart-sound signals result in significant differences in their time-frequency domain representations, thereby increasing the difficulty of feature extraction and classification accuracy. Third, heart murmurs often coexist with other environmental noises, and these cross-frequency interference factors reduce the accuracy of the models in classifying abnormal heart sounds. Therefore, enriching the diversity of data samples, balancing sample distributions, optimizing feature representations, and effectively suppressing noise interference have become core issues for improving model performance. We addresses the issues above by proposing an improved heart sound signal abnormality detection algorithm, the DCT Convolution Swin Transformer (DCv-Swin Transformer), which builds upon the Swin Transformer model [32]. The DCv-Swin Transformer algorithm employs various techniques, such as Butterworth filtering [14], Focal Loss [33] as the loss function, manipulation of the playback speed of heart-sound signals for database augmentation, Hierarchical Audio Transformer architecture [34], Discrete Cosine Transform (DCT), Convolutional Token Embedding (ConvEmbed), and Convolution Projection [35]. These techniques aimed to shorten the model training time while enhancing the accuracy of abnormal heart-sound recognition. An experimental evaluation of the DCv-Swin Transformer was conducted using the PhysioNet/CinC 2016 dataset [36]. The results demonstrate the high accuracy of the model in classifying heart-sound signals.

III. METHODOLOGY

With the flourishing development of research in the artificial intelligence, we observed the widespread application of models based on the transformer algorithm [37] in speech classification. Notably, both heart sound and speech signals belong to the realm of acoustic signals. Given the high degree of similarity in their characteristics, this study employs the Swin Transformer algorithm, derived from the transformer algorithm, to detect abnormalities in heart sound signals.

A. SWIN TRANSFORMER LAYER

The Swin Transformer model primarily comprises patch partitions, Swin Transformer blocks, and patch-merging components.The structure of the Swin Transformer is illustrated in Fig 1.

The complete Swin Transformer consisted of five stages. In the first stage, the input data are fed into the patch-embedding module, where a convolutional layer is utilized to segment the input data into non-overlapping patches. Each segmented patch is defined as a 'token. ' In the second stage, the segmented tokens were input into a Linear Embedding layer, where each token was mapped to a dimension of C (C=96). Subsequently, the tokens enter the two Swin Transformer block modules. This module primarily consists of Layer Normalization, two window-based self-attention mechanisms (W-MSA and SW-MSA), and an ((MLP). Within the Swin Transformer Block, self-attention calculations were performed on the tokens. The third stage involves Patch Merging to reduce the number of tokens while reducing their feature dimensions. This operation, known as downsampling, merges 2 tokens and concatenates them along the C dimension, resulting in a dimension of 4C. Subsequently, a linear mapping layer was employed to transform the dimensions from 4C to 2C. Following down-sampling, the number of tokens was reduced by a factor of four. Similar to the second stage, the tokens underwent self-attention calculations using two Swin Transformer Blocks. In the fourth stage, further Patch Merging is conducted to decrease the token count, whereas down-sampling is performed to reduce the dimensionality of the feature vectors. Self-attention calculations were performed using six Swin Transformer blocks to transform the features into deeper layers. In the fifth and final stages, another round of Patch Merging was performed, followed by self-attention calculations using two Swin Transformer Blocks to obtain the final features. The Swin Transformer, utilizing window-based self-attention mechanisms (including W-MSA and SW-MSA) and a hierarchical structure design, demonstrates significant advantages in capturing multiscale features across both the temporal and frequency domains of heart-sound signals. This is crucial for enhancing the accuracy and efficiency of heart-sound classification. However, its complex architecture has a higher parameter count and computational demands, extends the training cycles, and places higher demands on computing resources. Furthermore, the common issue of sample imbalance in heart-sound datasets often limits the generalization ability of the model. In light of these challenges, this study proposes a specialized algorithm for abnormal heart-sound recognition of DCv-Swin transformers aimed at addressing the limitations above. Through experiments conducted on the standard heart sound classification dataset PhysioNet/CinC 2016, we validated the effectiveness of the DCv-Swin Transformer algorithm in improving the model adaptability and classification performance. This, in turn, provides a more efficient and robust solution for heart sound signal analysis.

B. NETWORK STRUCTURE

1) Backbone feature network design

The framework of the proposed algorithm, the DCv-Swin Transformer, is presented in Fig2. In this study, we introduce a ConvEmbed structure instead of the original patch embedded within the heart-sound mel spectrogram encoder module.

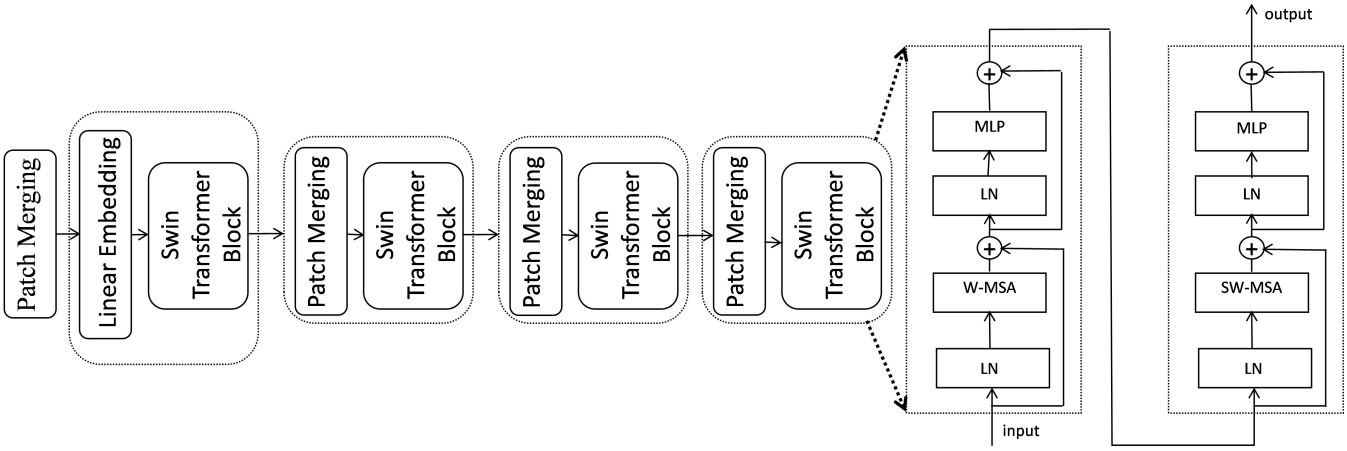

FIGURE 1: **Swin Transformer Layer Network.**

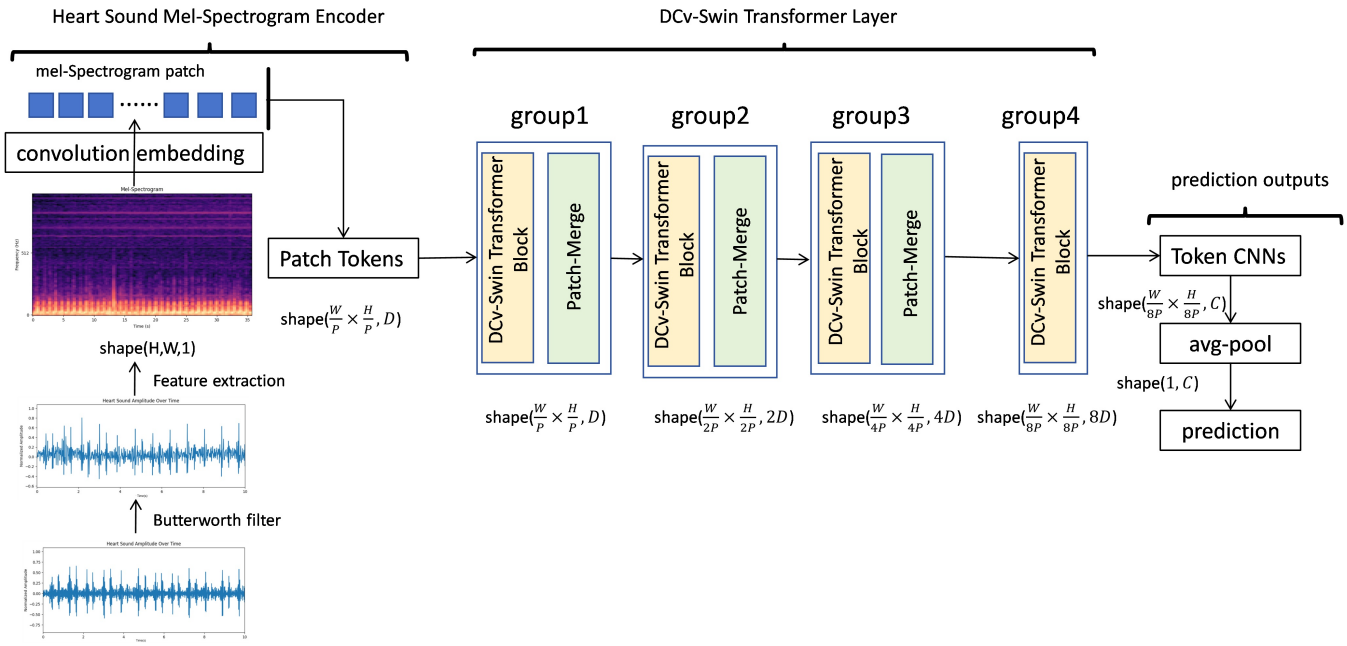

FIGURE 2: **The overall framework of our proposed DCT Comvolution Swin Transformer(DCv-Swin Transformer).**

This modification involves feeding the sequence of mel-spectrogram patches extracted by the model into the DCv-Swin Transformer Layer. The DCv-Swin Transformer Layer adopts a hierarchical design comprising four layers, with the first three layers performing one downsampling operation each. Downsampling reduces the number of training parameters, thereby enhancing the training speed. Within the DCv-Swin Transformer block, a discrete convolution projection (DCP) module was utilized to obtain Q, K, and V (as depicted in Fig 3) for window-based attention calculations. The DCv structure facilitates more efficient extraction of frequency-domain features from heart-sound signals, thereby enhancing the model's resistance to interference. Finally, the prediction outputs were generated to provide the model's prediction results.

2) Token Embedding

The complexity and diversity of features in heart-sound signals often lead to instances of omission and misjudgment in abnormal heart-sound recognition. To address this challenge, this study attempted to incorporate a ConvEmbed structure to enhance the classification accuracy of the model. Initially, heart-sound signals are transformed into mel spectrograms. Then ConvEmbed is utilized to map them into a sequence of feature vectors (tokens), with dimension transformation and standardization preprocessing implemented in between. This structure strengthens the model's ability to capture intrinsic temporal and frequency domain features of heart sounds and converts the signals into a sequential form (patches) suitable for model input, thereby significantly enhancing the accuracy of abnormal heart sound recognition. The specific opera-

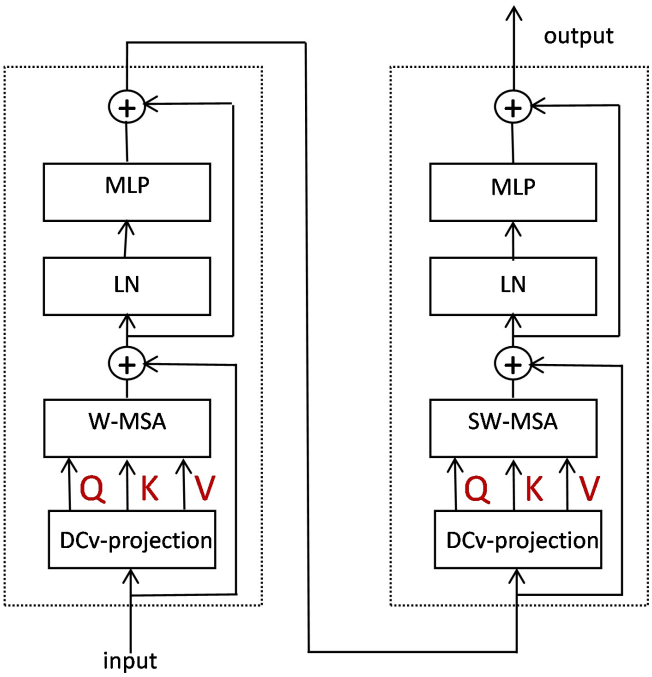

FIGURE 3: **Discrete convolutional window self attention mechanism.**

tions are as follows: First, we utilized the Spectrogram and LogmelFilterBank methods from the PyTorch framework to extract Mel spectrogram features from heart-sound signals. The shape of the Mel spectrogram is denoted as (C, F, T), where F represents the frequency dimension of the heart-sound signal, and T represents the time dimension. Additionally, the default number of channels C in the Mel spectrogram is set to 1, resulting in . Secondly, the input Mel spectrogram undergoes a convolution operation $f_{(x)}$ to be transformed into feature vectors Y , denoted as $Y = f_{(x)}$. The size of the convolutional kernel (kernel size) was set to $k \times k$, the stride (stride) was set to s, and the padding was set to p to maintain the consistency of the input and output dimensions, thereby avoiding loss of heart-sound feature information. Y can be expressed as $Y \in \mathbb{R}^{D\times H\times W}$, where D represents the depth of the feature vectors, H represents the height dimension, and W represents the width dimension. In the third step, a reshaping operation is applied to the feature vectors Y, denoted as $Y^{'} = rearrange(Y^{'}, bchw \rightarrow b(hw)c)$, where $Y^{'} \in \mathbb{R}^{B\times(H\times W)\times D}$. The purpose of this operation is to merge the dimensions $H \times W$ of the feature vector matrix into a single dimension, thereby transforming the Mel spectrogram data into sequential data suitable for model input. Finally, the feature vectors were normalized, denoted by $Y^{'} = Norm(Y^{'})$. The mathematical expression for ConvEmbed is given in Equation (1):

$$Y_{output} = rearrange(Norm(rearrange(Conv2d(X)))) \quad (1)$$
$$Y_{output} \in \mathbb{R}^{B\times C\times H\times W}$$

In this study, we introduce a Convolutional Token Embedding (ConvEmbed) structure that dynamically adjusts its parameters to adapt to the dimensions and quantity of feature vectors at each stage. Specifically, this structure gradually reduces the sequence length while simultaneously increasing the dimensions of the individual feature vectors. As processing levels progress, this enables the representation of more complex heart-sound features, mirroring the deepening of the feature hierarchy in convolutional neural networks. This design enhances the efficient capture of local features and facilitates deeper learning in higher-dimensional feature spaces, thereby improving the accuracy of heart-sound signal classification models.

3) Discrete Depthwise Convolution

In abnormal heart-sound recognition, a critical step involves analyzing heart murmurs occurring between the S1 (first heart sound, corresponding to diastole) and S2 (second heart sound, corresponding to systole) phases of the cardiac cycle, serving as the basis for classification. However, heart murmur signals contain a mixture of genuine cardiovascular state reflections and non-target interferences, such as environmental noise and respiratory sounds. This interference spans the frequency distributions, thus hindering the accuracy of the models in recognizing abnormal heart sounds. Therefore, this study introduces a Discrete Depthwise Convolution (DDC) module based on the Discrete Cosine Transform (DCT) to optimize the computation process of queries (q), keys (k), and values (v) within the attention mechanism as illustrated in Fig 4.

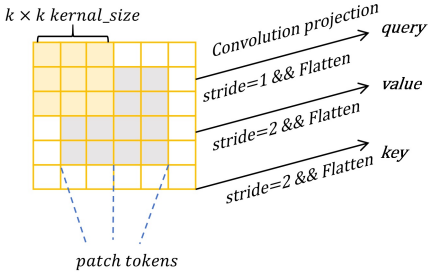

FIGURE 4: **Multi stripe depthwise separable convolution.**

The application of DCT strengthens the extraction capabilities of frequency-domain features from heart-sound signals while effectively filtering out irrelevant high-frequency and low-frequency interference, thereby enhancing the classification performance. Additionally, the DDC module optimizes the classification accuracy and training efficiency during the attention mechanism. The specific operations are as follows:

First, we apply a Discrete Cosine Transform (DCT) to process the mel spectrogram of heart-sound signals. The DCT effectively mitigates the interference of heart-sound noise on the classification performance of the model, thereby enhancing the robustness of the model to heart-sound noise. Finally, we utilize a depthwise separable convolution method with a convolution kernel size of $k \times k$ to implement a convolutional

projection.Using the Discrete Depthwise Convolution (DDC) module, we reduce the computational cost of the attention mechanism. Operating a convolution kernel of size $k \times k$ can decrease the number of feature vectors using a stride greater than one. In our algorithm design, we set the stride size of K and V to 2, and the stride size of Q to 1. This reduces the number of vectors for K and V by fourfold, thereby decreasing the computational cost of the attention calculation by fourfold. The structure of the DDC module is shown in the Fig 4. The mathematical formula for the DDC module is given by Equation (2):

$$x_i^{q/k/v} = Flatten(conv2d(reshape2D(DCT(x_i)), k)) \quad (2)$$

In this context, $x_i^{q/k/v}$ represents the input vector matrix for the attention mechanism in the Swin Transformer Block, $conv2D$ denotes the depth-wise separable convolution, which alters the dimensions of feature vectors, $Reshape2D$ refers to the operation of applying Discrete Cosine Transform to input feature vectors, and DCT represents the size of the convolutional kernel. The experimental results demonstrate that the improved DCv-Swin Transformer algorithm, which implements the convolutional projection operation, can enhance the classification accuracy of the model while significantly reducing the training time under the same training parameter conditions.

C. FOCAL LOSS FUNCTION

To address the phenomenon of a significantly larger number of normal heart-sound samples than abnormal samples in the heart-sound classification dataset, this study introduced a Focal Loss function to mitigate the issue of data imbalance. The primary idea behind Focal Loss is to adjust the weights of the loss function, focusing the model on abnormal heart sounds to address the problems of low classification accuracy and generalization caused by data imbalance. The Focal Loss function introduces a focal parameter that adjusts its value to reduce the loss of normal heart-sound samples and amplify the loss of abnormal heart-sound samples. This design allowed the model to prioritize abnormal heart-sound samples, thereby improving its ability to recognize these samples. The mathematical formula for the Focal Loss function is given by Equation (3):

$$FL_{(p_t)} = -\alpha_t(1-p_t)^{\gamma}\log p_t \quad (3)$$

In the formula, p_t represents the probability value predicted by the model, indicating the probability that a sample belongs to the positive class; α_t denotes the class weights used to adjust the priority of positive and negative samples; γ is the focal parameter, controlling the weight parameter distribution of normal and abnormal heart sound samples. The experimental results demonstrate that the Focal Loss function, through conditional weighting parameters, can effectively address the issue of imbalance between normal and abnormal heart-sound samples in heart-sound classification tasks, thereby enhancing the accuracy and robustness of the model in identifying abnormal heart-sounds.

IV. EXPERIMENTS

A. DATA PREPROCESSING

This study utilized a heart-sound dataset publicly available from the PhysioNet/CinC 2016 Challenge to train the DCv-Swin Transformer algorithm. The dataset comprised 3240 heart-sound samples, with 2757 samples categorized as normal heart-sounds and 665 samples as abnormal heart-sounds. The sampling rate across the dataset was uniformly set at 2000 Hz. It was observed that there exists an imbalance in the dataset, with a ratio of approximately 4:1 between normal and abnormal heart sounds, posing a challenge as the DCv-Swin Transformer algorithm tends to bias towards learning to discern normal heart sounds, thereby exhibiting relatively weaker performance in identifying abnormal heart sounds. Moreover, environmental noise interference while recording heart-sound data further exacerbates the discriminatory capabilities of the model. To address these issues, this study employs data augmentation techniques by accelerating the heart-sound sample rate to balance the dataset and designing Butterworth filters to mitigate the heart-sound noise.

1) Balanced Dataset

This study employs linear interpolation techniques to address the imbalance in the dataset, where there are fewer abnormal heart-sound samples leading to weaker model generalization. Specifically, non-normal heart-sound audio samples were subjected to a $1.5\times$ acceleration and $0.5\times$ deceleration. This augmentation strategy aims to increase the dataset volume and diversity by expanding the number of abnormal heart sound samples to 1995, thus balancing the ratio between normal and abnormal heart sound data samples. This approach enhanced the generalization capabilities of the model for heart-sound classification.

2) Noise Filtering

Heart-sound signals typically encompass noise from sources such as respiratory sounds and environmental noise present during recording, all of which qualify as high-frequency noise. To counteract the potential interference of such noise on the classification performance of the model, this study introduced a fifth-order Butterworth bandpass filter for heart-sound noise filtration. Given that heart sound signals are characterized by low-frequency components, with the frequency distribution of the first and second heart sounds primarily falling between 50-500 Hz [22], the Butterworth filter employed in this study filters out noise components above 500 Hz and below 50 Hz from the heart-sound signal. The Butterworth bandpass filter effectively denoised the heart-sound signal, thereby providing a more comprehensive representation of the complete cardiac cycle. The waveform plots in Fig 5 depict the original heart-sound data (left) and the heart-sound data after filtration with a Butterworth bandpass filter (right).

B. MODEL TRAINING

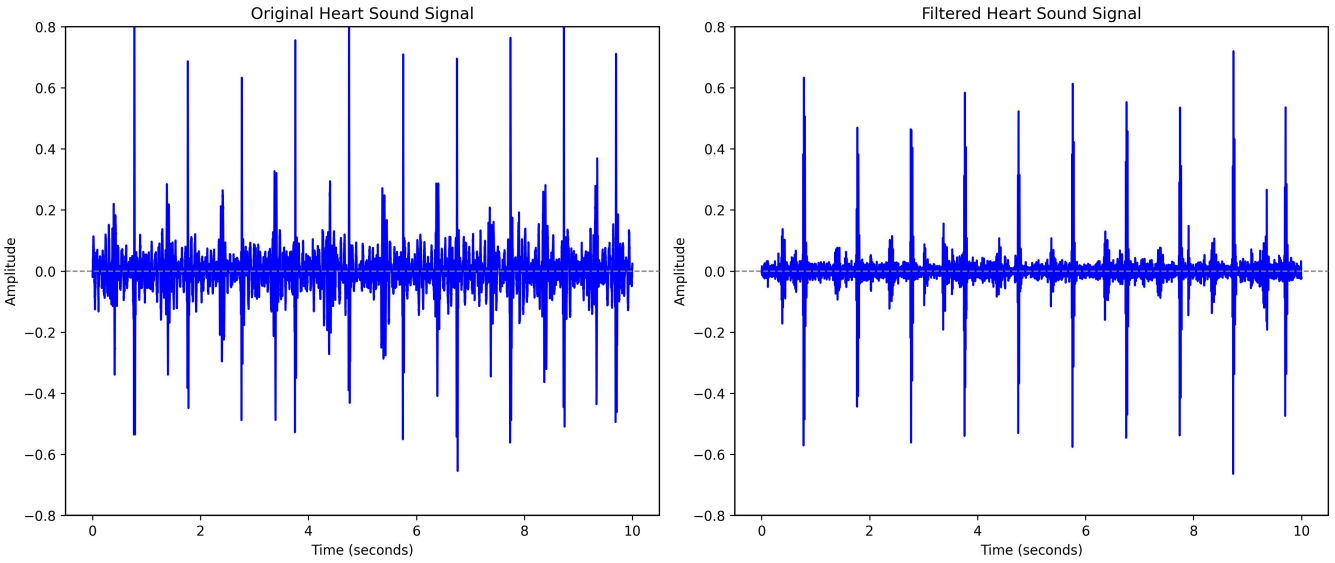

FIGURE 5: **Comparison of time-domain waveforms of heart sound samples after Butterworth denoising.**

1) Lab Enviroment

The server configuration parameters for training the DCv-Swin transformer algorithm were as follows: Windows operating system with hardware comprising an Intel(R) Gold 6140 CPU, 32 GB of RAM, an NVIDIA GeForce RTX4090 GPU, and a 1 TB SSD hard drive. PyTorch is a deep learning framework. Python version 3.8 was employed, along with GPU acceleration software CUDA 11.7, and CUDNN 8.4.1.

2) Training Details

The heart-sound dataset used for training maintained a sampling rate of 2000 Hz for all samples. To ensure that the algorithm adequately learned the underlying features of heart-sound samples, we partitioned the dataset into a ratio of 8:1:1 for training, validation, and testing sets to enhance the model's classification efficiency. During the training process, the model was trained on the training set, validated using the validation set to optimize the hyperparameters, and finally evaluated on the testing set to assess its classification performance, with a batch size of 24 to minimize its impact on the classification model. For samples within the same batch, the maximum sequence length was chosen as the uniform length, with zero padding applied to ensure uniform sequence length across all heart sound samples. In feature extraction, a window size 1024, a hop size 320, and 64 Mel bins were utilized for the Short-Time Fourier Transform to obtain the Mel spectrogram. To address the issue of early-stage underfitting in transformer-based models, a warm-up training strategy was employed for the DCv-Swin transformer algorithm with 100 epochs. For the initial three epochs, learning rates of 0.02, 0.05, and 0.1 were applied, followed by a learning rate of 0.001 from the 4th epoch onwards. Additionally, a CosineAnnealingLR strategy was used to control the learning rate decay during training. To effectively address the issue of dataset imbalance, we introduced a Focal Loss Function, while the Adam optimizer was used to accelerate the model's fitting speed.

C. MODEL EVALUATION INDEX

We evaluated the performance of the Dcv-Swin Transformer model using a confusion matrix to calculate Accuracy(Acc), Precision(P), Sensitivity(Se), Specificity(Sp), and F1-score. Equation (4) defines Acc as the proportion of correctly predicted positive and negative samples to the total number of samples. Equation (5) defines P as the proportion of true positives among all samples predicted as positive. Equation (6) defines Se as the proportion of actual positive samples correctly identified by the model. Equation (7) defines Sp as the proportion of negative samples correctly predicted. Finally, Equation (8) describes the F1-score, which measures the model's overall performance by balancing Precision and Sensitivity.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$P = \frac{TP}{TP + FP} \tag{5}$$

$$Se = \frac{TP}{TP + FN} \tag{6}$$

$$Sp = \frac{TN}{TN + FP} \tag{7}$$

$$F1\text{-}score = 2 \times \frac{P \times R}{P + R} \tag{8}$$

In Equations (4) to (8), TP represents the number of true positives, FP denotes the number of false positives where negative samples are incorrectly identified as positive, FN represents the number of false negatives where positive samples are incorrectly identified as harmful, and TN denotes the

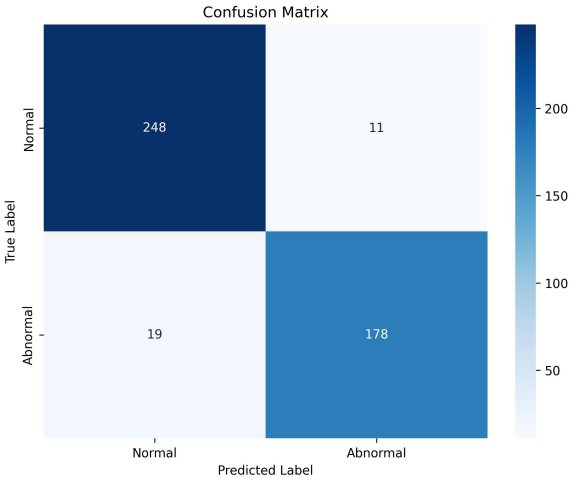

FIGURE 6: **DCv-Swin Transformer confusion matrix.**

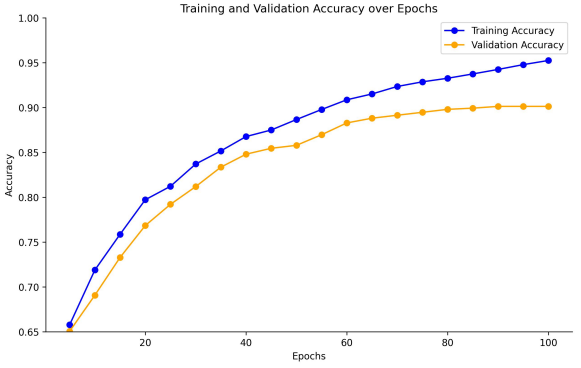

FIGURE 7: **DCv-Swin Transformer training set and validation set Accuracy conversion curve.**

number of true negatives where negative samples are correctly identified.

V. ANALYSIS OF RESULTS

A. COMPARATIVE EXPERIMENT

We trained the DCv-Swin Transformer model using the previous section's training methods. To visually present the best training results of the model, we display the confusion matrix in Fig 6, which clearly shows the values of true positives (*TP*), true negatives (*TN*), false positives (*FP*), and false negatives (*FN*). Additionally, to analyze the model's performance during the training process further, we present the changes in accuracy for both the training set and the validation set over time in Fig 7. These charts provide quantitative metrics of the model's performance and reveal its stability and convergence throughout the training process.

To better evaluate the performance of the DCv-Swin Transformer algorithm on the PhysionNet/CinC 2016 dataset, this study compared the experimental results with different methods using the same dataset (PhysionNet/CinC Challenge 2016). Table 1 presents the methods used in these studies and their respective results and compares them with the experimental results of this study.

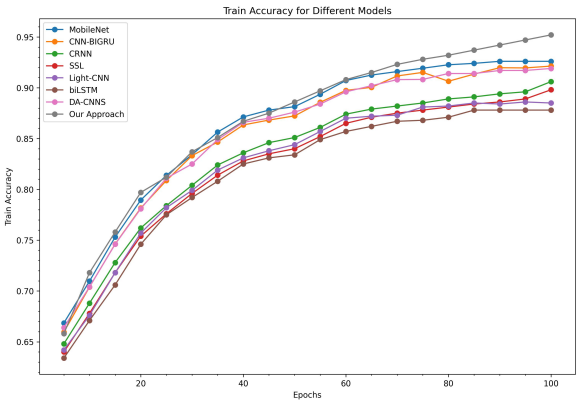

FIGURE 8: **Comparison of Accuracy changes in training sets of different models.**

To comprehensively evaluate the performance of different models, we present the accuracy change curves for various models in Fig 8. This comparison helps readers understand the advantages of the DCv-Swin Transformer relative to other models and demonstrates its performance at different stages. The comparison results from the Tabel 1 indicate that the proposed DCv-Swin Transformer algorithm performs better than other heart-sound classification algorithms on the publicly available PhysionNet/CinC 2016 dataset. Overall, among the eight algorithms considered, the DCv-Swin Transformer algorithm achieved the highest accuracy of 93.4% and a specificity of 90.4%. Additionally, Our approach Sensitivity reached 95.7%, and its F1-score value was 94.4%, demonstrating its superior performance in heart sound classification tasks.

These results show that the DCv-Swin Transformer has high accuracy in identifying abnormal heart sounds and reflects its advantage in capturing abnormal signals, as evidenced by its high Sensitivity and F1 score. This indicates that the algorithm possesses strong generalization abilities and represents an improvement over previous algorithms, especially in handling complex heart sound signals.

Finally, we statistically compared the average training time per epoch for the models mentioned in Table 1. As shown in Fig 9, the DCv-Swin Transformer employs a hierarchical design with multiple downsampling layers, significantly reducing the training time. The experimental results indicate that the average training time per epoch for the DCv-Swin Transformer is 46 seconds, demonstrating its clear advantage in computational efficiency. This efficient training time accelerates the model iteration process and reduces the demand for computational resources, making it more suitable for handling large datasets and real-time applications.

B. ABLATION EXPERIMENT

The primary innovations of the proposed DCv-Swin Transformer algorithm are the introduction of two ConvEmbed modules, the DCP, and the utilization of the Focal Loss

TABLE 1
Comparison of Model Evaluation Indices.

Method	Acc(%)	Sp(%)	Se(%)	P(%)	F1-score(%)
MobileNet [25]	92.6	92.4	93.3	87.0	90.1
CNN-BIGRU [26]	91.7	90.0	90.7	91.4	91.1
CRNN [27]	90.6	89.8	90.0	88.4	89.7
SSL [28]	89.8	91.9	88.2	84.2	86.1
Light-CNN [29]	88.7	86.6	90.1	84.3	87.1
biLSTM [30]	87.7	85.2	90.2	87.3	88.6
DA-CNNS [31]	91.7	92.4	88.7	92.6	90.5
Our Approach	93.4	90.4	95.7	92.8	94.4

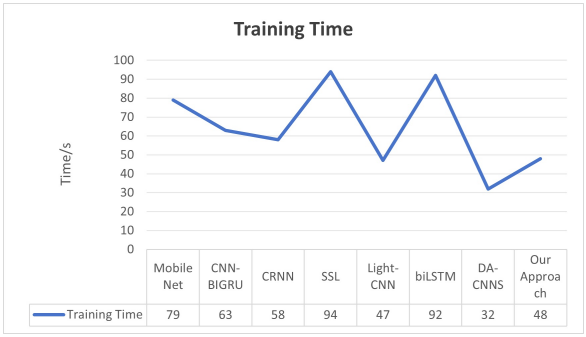

FIGURE 9: **Comparison of training time for different heart sound models.**

function. To validate the effectiveness of the proposed algorithm, experiments were conducted on the publicly available PhysionNet/CinC 2016 dataset for heart-sound signal classification, along with ablation studies on the algorithm. In this study, the swine transformer served as the standard model and was used as a reference model for the ablation experiments. Within the Swin Transformer, the ConvEmbed structure replaced the original Patch Embedding to enhance the model's ability to extract heart-sound features and improve classification accuracy. In addition, a Focal Loss function was introduced to address the issue of dataset imbalance. The resulting model is named CeF-SwinT. To further enhance the ability of the model to extract frequency-domain features from heart-sound signals, the DDC structure replaced the linear mapping used for attention computation in the swine transformer. Similarly, a Focal Loss function is incorporated. The resulting model is named DcF-SwinT. The experimental results are presented in Table 2.

From the Table 2, it can be observed that the DCv- Swin Transformer network proposed in this study achieved improvements of approximately 2.4%, 1.0%, and 2.7% in accuracy, specificity, and Sensitivity, respectively, compared with the standard model on the PhysionNet/CinC 2016 dataset.

Finally, regarding the utilization of the SwinTransformer with a hierarchical structure to enhance training efficiency and reduce training time, thereby reducing the consumption of hardware resources, such as GPUs, during algorithm training, this study compared the training times of the Swin and DCv-swin transformers under the same number of training epochs. The training process took approximately 4.6 hours,whereas the DCv-Swin transformer only required 1.28 hours. It is evident that the hierarchical DCv-Swin transformer significantly reduces the algorithm training time and improves training efficiency.

VI. DISCUSSION

This paper proposes a novel method for heart sound classification by incorporating convolutional concepts into the Swin-Transformer model. From the results discussed above, the strengths and limitations of our model can be summarized as follows:

First, the model effectively reduces noise interference in signal classification by applying a Butterworth filter to denoise heart sound signals. Additionally, we introduced modifications in the patch-embedding layer of the Swin-Transformer, employing Patch Embedding to enhance the model's ability to capture temporal features of heart sound signals. Moreover, the original linear projection layer was replaced with a depth-wise separable convolution module combined with Discrete Cosine Transform (DCT), which significantly improves the extraction of fine-grained features, such as the first heart sound (S1), second heart sound (S2), murmurs, and pathological signals. Lastly, to address the issue of imbalanced heart sound datasets, where abnormal samples far outnumber normal ones, we employed a linear interpolation method for data augmentation. At the same time, the Focal Loss function was utilized to enhance the model's performance on imbalanced data.

Compared to traditional convolutional neural networks (CNNs) and models based on RNNs or LSTMs, our proposed DCv-SwinTransformer model demonstrates superior classification performance, as shown in Table 1, when benchmarked against several prominent heart sound classification models from the past two years. The unique hierarchical feature extraction capability of the DCv-SwinTransformer enables it to more effectively capture both local and global features of heart sound signals, resulting in improved adaptability and performance in heart sound classification tasks.

However, as evidenced in Table 1, the model exhibits relatively low specificity, indicating a tendency to miss normal heart sounds. This issue may be attributed to two primary factors: (1) The Swin Transformer-based model relies on large datasets for optimal performance, yet the available pub-

TABLE 2
Comparison of Ablation Experimental Data.

Method	Acc(%)	Sp(%)	Se(%)	P(%)	F1-score(%)
SwinT	91.0	88.3	93.0	92.0	92.1
Cef-SwinT	92.3	90.3	93.4	92.7	93.0
Dcf-SwinT	92.1	91.9	92.2	93.4	93.2
Our Approach	93.4	90.4	95.7	92.8	94.4

lic datasets for heart sound classification are limited, which restricts the amount of training data available for the DCv-Swin Transformer model. (2) The complex patterns inherent in heart sound signals may not be fully captured by Mel spectrogram features, leading to the misclassification of subtle variations in normal heart sounds as pathological conditions.

In future work, we plan to address the aforementioned issues by employing transfer learning techniques using pre-trained models to alleviate the problem of data scarcity. Additionally, to mitigate misclassification issues arising from the complexity of heart sound signal features, we intend to utilize higher-order spectral analysis methods from the field of digital signal processing. Specifically, we will apply bispectral analysis for feature extraction of heart sound signals and incorporate these features into subsequent anomaly detection tasks.

VII. CONCLUSION

In this paper, we propose an effective method for detecting abnormal heart sounds. This method includes data preprocessing steps and the DCv-Swin Transformer algorithm. Specifically, we designed a fifth-order Butterworth filter to reduce noise interference and adopted the Focal Loss function along with linear interpolation to address the class imbalance between normal and abnormal heart sound samples. Additionally, we utilized ConvEmbed instead of traditional Patch Embedding to enhance the model's ability to capture local features of heart sound signals. By incorporating a Discrete Cosine Transform (DDC) structure, we improved the model's capability to capture the correlation between time-domain and frequency-domain features of heart sounds. Experimental results demonstrate that the DCv-Swin Transformer algorithm exhibits superior performance in heart sound classification tasks. Although significant progress has been made in heart sound classification, there remain several challenges. Future work will focus on improving the model's generalization capabilities and exploring better heart sound signal features to further enhance its performance in practical applications.

REFERENCES

[1] Mahboobeh Jafari, Afshin Shoeibi, Marjane Khodatars, Navid Ghassemi, Parisa Moridian, Roohallah Alizadehsani, Abbas Khosravi, Sai Ho Ling, Niloufar Delfan, Yu-Dong Zhang, et al. Automated diagnosis of cardiovascular diseases from cardiac magnetic resonance imaging using deep learning models: A review. *Computers in Biology and Medicine*, 160:106998, 2023.

[2] Writing Group Members, Donald Lloyd-Jones, Robert J Adams, Todd M Brown, Mercedes Carnethon, Shifan Dai, Giovanni De Simone, T Bruce Ferguson, Earl Ford, Karen Furie, et al. Heart disease and stroke statistics—2010 update: a report from the american heart association. *Circulation*, 121(7):e46–e215, 2010.

[3] Spencer L Strunic, Fernando Rios-Gutiérrez, Rocío Alba-Flores, Glenn Nordehn, and Stanley Burns. Detection and classification of cardiac murmurs using segmentation techniques and artificial neural networks. In *2007 IEEE symposium on computational intelligence and data mining*, pages 397–404. IEEE, 2007.

[4] MZC Lam, TJ Lee, PY Boey, WF Ng, HW Hey, KY Ho, and PY Cheong. Factors influencing cardiac auscultation proficiency in physician trainees. *Singapore medical journal*, 46(1):11, 2005.

[5] Tehseen Mazhar, Qandeel Nasir, Inayatul Haq, Mian Muhammad Kamal, Inam Ullah, Taejoon Kim, Heba G Mohamed, and Norah Alwadai. A novel expert system for the diagnosis and treatment of heart disease. *Electronics*, 11(23):3989, 2022.

[6] Inayatul Haq, Tehseen Mazhar, Muhammad Amir Malik, Mian Muhammad Kamal, Inam Ullah, Taejoon Kim, Monia Hamdi, and Habib Hamam. Lung nodules localization and report analysis from computerized tomography (ct) scan using a novel machine learning approach. *Applied Sciences*, 12(24):12614, 2022.

[7] Ali Harimi, Yahya Majd, Abdorreza Alavi Gharahbagh, Vahid Hajihashemi, Zeynab Esmaileyan, José JM Machado, and João Manuel RS Tavares. Classification of heart sounds using chaogram transform and deep convolutional neural network transfer learning. *Sensors*, 22(24):9569, 2022.

[8] Fuad Noman, Chee-Ming Ting, Sh-Hussain Salleh, and Hernando Ombao. Short-segment heart sound classification using an ensemble of deep convolutional neural networks. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1318–1322. IEEE, 2019.

[9] Bilal Ahmad, Faiq Ahmad Khan, Kaleem Nawaz Khan, and Muhammad Salman Khan. Automatic classification of heart sounds using long short-term memory. In *2021 15th International Conference on Open Source Systems and Technologies (ICOSST)*, pages 1–6. IEEE, 2021.

[10] Muqing Deng, Tingting Meng, Jiuwen Cao, Shimin Wang, Jing Zhang, and Huijie Fan. Heart sound classification based on improved mfcc features and convolutional recurrent neural networks. *Neural Networks*, 130:22–32, 2020.

[11] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[12] Ben Walker, Felix Krones, Ivan Kiskin, Guy Parsons, Terence Lyons, and Adam Mahdi. Dual bayesian resnet: A deep learning approach to heart murmur detection. In *2022 Computing in Cardiology (CinC)*, volume 498, pages 1–4. IEEE, 2022.

[13] Keqi Liu, Lei Yuan, Chengji Huang, Wenyuan Wu, Qiangwei Wang, and Gang Wu. Abnormal heart sound detection by using temporal convolutional network. In *2022 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC)*, pages 1026–1029. IEEE, 2022.

[14] Suyi Li, Feng Li, Shijie Tang, and Fan Luo. Heart sounds classification based on feature fusion using lightweight neural networks. *IEEE Transactions on instrumentation and measurement*, 70:1–9, 2021.

[15] Omer Deperlioglu. Heart sound classification with signal instant energy and stacked autoencoder network. *Biomedical Signal Processing and Control*, 64:102211, 2021.

[16] Megha Banerjee and Sudhan Majhi. Multi-class heart sounds classification using 2d-convolutional neural network. In *2020 5th International conference on computing, communication and security (ICCCS)*, pages 1–6. IEEE, 2020.

[17] Yongchao Chen, Shoushui Wei, and Yatao Zhang. Classification of heart sounds based on the combination of the modified frequency wavelet transform and convolutional neural network. *Medical & Biological Engineering & Computing*, 58:2039–2047, 2020.

[18] Andrea Duggento, Allegra Conti, Maria Guerrisi, and Nicola Toschi. Classification of real-world pathological phonocardiograms through multi-instance learning. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 771–774. IEEE, 2021.

[19] Kalpeshkumar Ranipa, Wei-Ping Zhu, and MNS Swamy. Multimodal cnn fusion architecture with multi-features for heart sound classification. In *2021 IEEE International symposium on circuits and systems (ISCAS)*, pages 1–5. IEEE, 2021.

[20] Bin Xiao, Yunqiu Xu, Xiuli Bi, Junhui Zhang, and Xu Ma. Heart sounds classification using a novel 1-d convolutional neural network with extremely low parameter consumption. *Neurocomputing*, 392:153–159, 2020.

[21] Shu Lih Oh, V Jahmunah, Chui Ping Ooi, Ru-San Tan, Edward J Ciaccio, Toshitaka Yamakawa, Masayuki Tanabe, Makiko Kobayashi, and U Rajendra Acharya. Classification of heart sound signals using a novel deep wavenet model. *Computer methods and programs in biomedicine*, 196:105604, 2020.

[22] Shan Gao, Yineng Zheng, and Xingming Guo. Gated recurrent unit-based heart sound analysis for heart failure screening. *Biomedical engineering online*, 19:1–17, 2020.

[23] Samiul Based Shuvo, Shams Nafisa Ali, Soham Irtiza Swapnil, Mabrook S Al-Rakhami, and Abdu Gumaei. Cardioxnet: A novel lightweight deep learning framework for cardiovascular disease classification using heart sound recordings. *ieee access*, 9:36955–36967, 2021.

[24] Haya Alaskar, Nada Alzhrani, Abir Hussain, and Fatma Almarshed. The implementation of pretrained alexnet on pcg classification. In *Intelligent Computing Methodologies: 15th International Conference, ICIC 2019, Nanchang, China, August 3–6, 2019, Proceedings, Part III 15*, pages 784–794. Springer, 2019.

[25] Menghui Xiang, Junbin Zang, Juliang Wang, Haoxin Wang, Chenzheng Zhou, Ruiyu Bi, Zhidong Zhang, and Chenyang Xue. Research of heart sound classification using two-dimensional features. *Biomedical Signal Processing and Control*, 79:104190, 2023.

[26] Harshwardhan Yadav, Param Shah, Neel Gandhi, Tarjni Vyas, Anuja Nair, Shivani Desai, Lata Gohil, Sudeep Tanwar, Ravi Sharma, Verdes Marina, et al. Cnn and bidirectional gru-based heartbeat sound classification architecture for elderly people. *Mathematics*, 11(6):1365, 2023.

[27] Samiul Based Shuvo, Syed Samiul Alam, Syeda Umme Ayman, Arbil Chakma, Prabal Datta Barua, and U Rajendra Acharya. Nrc-net: Automated noise robust cardio net for detecting valvular cardiac diseases using optimum transformation method with heart sound signals. *Biomedical Signal Processing and Control*, 86:105272, 2023.

[28] Aristotelis Ballas, Vasileios Papapanagiotou, Anastasios Delopoulos, and Christos Diou. Listen2yourheart: A self-supervised approach for detecting murmur in heart-beat sounds. In *2022 Computing in Cardiology (CinC)*, volume 498, pages 1–4. IEEE, 2022.

[29] Hui Lu, Julia Beatriz Yip, Tobias Steigleder, Stefan Grießhammer, Maria Heckel, Naga Venkata Sai Jitin Jami, Bjoern Eskofier, Christoph Ostgathe, and Alexander Koelpin. A lightweight robust approach for automatic heart murmurs and clinical outcomes classification from phonocardiogram recordings. In *2022 Computing in Cardiology (CinC)*, volume 498, pages 1–4. IEEE, 2022.

[30] Mahmoud Fakhry and Abeer FathAllah Brery. A comparison study on training optimization algorithms in the bilstm neural network for classification of pcg signals. In *2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, pages 1–6. IEEE, 2022.

[31] Shumpei Takezaki and Kazuya Kishida. Construction of cnns for abnormal heart sound detection using data augmentation. In *Lecture Notes in Engineering and Computer Science: Proceedings of The International MultiConference of Engineers and Computer Scientists*, pages 20–22, 2021.

[32] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[33] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[34] Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 646–650. IEEE, 2022.

[35] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22–31, 2021.

[36] Bruno M Rocha, Dimitris Filos, Luís Mendes, Gorkem Serbes, Sezer Ulukaya, Yasemin P Kahya, Nikša Jakovljevic, Tatjana L Turukalo, Ioannis M Vogiatzis, Eleni Perantoni, et al. An open access database for the evaluation of respiratory sound classification algorithms. *Physiological measurement*, 40(3):035001, 2019.

[37] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

ZHIHAI LIU received a Bachelor's degree in Computer Science and Technology from Luoyang University of Technology in 2022.He is currently a master's student majoring in computer technology at Xinjiang Normal University.His research interests include biomedical signal processing, artificial intelligence, and deep learning.

WEN LIU received the BS degree in computer science from Xinjiang Normal University, Xinjiang, in 2004, and the PhD degree in computer science from Dalian University of Technology, Dalian, China, in 2009. He is currently a professor in College of Control Engineering, Xinjiang Institute of Engineering. His research interests include database, stream data processing, and cloud computing.

ZHENG GU received the BS degree in Measurement and Control Technology and Instrumentation from Tianjin University of Technology and Education, Tianjin, in 2012, and the MS degree in Instrumentation Science and Technology from Southwest Petroleum University, Chengdu, in 2015. She is currently an associate professor in College of Control Engineering, Xinjiang Institute of Engineering. Her research interests include artificial intelligence and image processing.

FENG YANG currently working at the Computer Information Center, Xinjiang Institute of Engineering. His research interests include acoustic signal processing, artificial intelligence, and deep learning.