The Visual Computer (2023) 39:3811–3822
https://doi.org/10.1007/s00371-023-03031-5

ORIGINAL ARTICLE

PCTMF-Net: heart sound classification with parallel CNNs-transformer and second-order spectral analysis

Rongsheng Wang[1] · Yaofei Duan[1] · Yukun Li[1] · Dashun Zheng[1] · Xiaohong Liu[2] · Chan Tong Lam[1] · Tao Tan[1]

Accepted: 6 July 2023 / Published online: 20 July 2023

Abstract

Heart disease is a common condition worldwide and has become one of the leading causes of death worldwide. The electrocardiogram (PCG) is a safe, painless, and non-invasive test that captures bioacoustic information reflecting the function of the heart by capturing the acoustic signal of the patient's heart. Nowadays, based on biosignal processing and artificial intelligence technologies, automated heart sound classification is playing an increasingly important role in clinical applications. In this paper, we propose a new parallel CNNs-transformer network with multi-scale feature context aggregation (PCTMF-Net). It combines the advantages of CNNs and transformer to achieve efficient heart sound classification. In PCTMF-Net, firstly, the heart tone signal features are extracted using the second-order spectral analysis, and a transformer-based MHTE-4 (multi-head transformer encoder with four attention heads) is designed to encode and aggregate the contextual information, and then, two CNNs feature extractors are designed in parallel with MHTE-4 to capture the hierarchical features. Finally, the feature vectors obtained from CNNs and MHTE-4 through feature fusion in PCTMF-Net will be fed into the fully connected layer for predicting the classification results of heart sounds. In addition, we perform validation based on two publicly available mutually exclusive heart sound datasets and conduct extensive experiments and comparisons of existing algorithms under different metrics. The experimental results show that our proposed method achieves 99.36% accuracy on the Yaseen dataset and 93% accuracy on the PhysioNet dataset. It surpasses current algorithms in terms of accuracy, recall and $F1$-score metrics. The aim of this study is to apply these new techniques and methods to improve the diagnostic accuracy and validity of heart disease for clinical use.

Keywords Classification of heart sound · Heart sound signal · Higher-order spectrum · Parallel convolution and transformer

1 Introduction

The heart is a vital organ of the human body. Cardiovascular disease (CVD) is characterized by burstiness and recidivism, making it one of the leading causes of morbidity and mortality in the world population [1]. With an estimated 17.5 million deaths from CVD-related disease in 2012, this accounts for 31% of all deaths worldwide [2]. However, the burden is particularly problematic in developed countries (LMICs), where high-quality diagnostics are often difficult to access in resource-limited areas. Engineering a mobile health tool to assess and manage cardiovascular disease risk is a promising endeavor. Although ultrasound and magnetic resonance imaging have replaced auscultation in wealthier economies [3], heart sound auscultation remains an important diagnostic method for outpatient physicians. However, with the patient-to-physician ratios as high as 50,000:1 in some parts of the world, access to specialist diagnosis is often hampered [4].

Traditionally, medical professionals have used heart sounds to detect heart disease through auscultation. Numerous disease disorders of the cardiovascular system are reflected in several heart-related signals, such as heart electronic signals (i.e., electrocardiographs or ECGs) and heart sound signals (i.e., phonocardiograms or PCGs). Compared to the conventional electrocardiogram (ECG), PCG is cost-effective, reproducible, and informative. The heart sound signal carries early pathological information about cardiovascular disease and is beneficial for the early identification of underlying cardiovascular disease [5]. Heart sounds generated by the

✉ Tao Tan
taotanjs@gmail.com

[1] Faculty of Applied Sciences, Macao Polytechnic University, Rua de Luís Gonzaga Gomes, Macao 999078, China

[2] Faculty of John Hopcroft Center, Shanghai Jiao Tong University, Shanghai 200240, China

mechanical activity of the myocardium can be heard in PCG recordings, and in addition to the correct heart sounds such as S1 and S2 [6], pathological murmurs can also be heard. Depending on their localization, different pathologies and structural defects can be diagnosed. PCG is a non-invasive diagnosis, cost-effective, and requires minimal equipment [7]. Therefore, PCG is well suited for cardiac screening, especially in small primary care clinics.

However, traditional heart sound auscultation has some drawbacks, and the results may be misdiagnosed or missed due to subjective factors such as the doctor's hearing and experience. Statistically, the accuracy rate of auscultation by cardiologists is about 80%, while that of primary care physicians is about 20–40% [8]. Moreover, PCG recordings are limited by the audible frequency range, and ambient noise and variations in the recording area are the biggest challenges in auscultation. To address these shortcomings, a cost-effective automated diagnostic system in ambulatory monitoring is a practical and effective way to do so. Such an approach could effectively reduce financial costs and more efficiently utilize the vast potential of expert resources. This screening method would also provide testing opportunities for a large population of potential CVDs, further providing additional diagnostic tests for medical evaluation. Consequently, there is an urgent need for an objective and automated computer-aided tool for PCG signal analysis aimed at an automatic classification of PCG signals. Today, based on biosignal processing and artificial intelligence techniques, automated heart sound classification has the potential to screen for pathology in a variety of clinical applications, thereby reducing costly and time-consuming manual examinations, and automated analysis of heart sound signals using computer technology is emerging as a promising area of research.

Early heart disease detection is important for patients to take preventive measures to reduce the harm caused by potential diseases. Improving the accuracy of automated heart sound auscultation is an important matter. In this paper, we propose classifying heart sound using parallel CNNs-transformer networks with second-order spectral analysis.

Our main contributions are highlighted as follows:

(1) A second-order spectral analysis is used in heart sound feature extraction. The second-order spectral analysis can be well applied to non-smooth medical signals, such as EEG, ECG, and PCG, which can effectively retain the useful features in the signal and reduce the noise.

(2) A PCTMF-Net parallel architecture with CNNs-transformer is proposed for heart sound classification, where MHTE-4 is designed to encode and aggregate contextual information, and a two-way CNNs structure in parallel with MHTE-4 is used to capture hierarchical features.

(3) We conducted experimental evaluations on two publicly available datasets. Our proposed method achieved the best performance in comparison with four state-of-the-art heart sound classification models on both datasets. Specifically, our method achieved the highest classification accuracy compared to other methods and demonstrated excellent advantages in evaluation metrics such as precision and recall. These results fully demonstrate that our proposed method has high accuracy and stability in heart sound classification tasks.

2 Related work

The standard process for the computer-based heart sound analysis can be summarized into the following steps: (1) preprocessing; (2) feature extraction; (3) classifier design.

In the past decades, fruitful methods have been reported for each step of the aforementioned heart sound analysis process. Oliveira et al. [9] proposed the need for a heartbeat alignment step, and evaluated different machine learning algorithms. Fatmawati et al. [10] compared the Empirical Mode Decomposition (EMD) and Double-Density Discrete Wavelet Transform (DD-DWT) method as a denoising system to minimize the noise effect in the PCG signal.

Zabihi et al. [11] extracted 40 time-frequency features from unsegmented heart sound signals and performed heart sound abnormality detection. Schmidt et al. [12] extracted different kinds of spectral features, including spectral parameter models, instantaneous frequency and amplitude (IFA), and octave power, to describe time-frequency properties. Kumar et al. [13] used fast wavelet decomposition to extract high-frequency heart sound features. Cristhian et al. [14] proposed to use MFCC (mel-frequency cepstral coefficient) cepstrum coefficients to convert one-dimensional PCG audio signal extraction into a two-dimensional time-frequency representation for heat map visualization and CNNs classification. Nilanon et al. [15] proposed to use spectrograms for feature extraction of heart sounds, and spectrograms can effectively capture the frequency, amplitude, and time information of heart sound signals.

As for heart sound classification, Stasis et al. [16] used a decision tree algorithm for the diagnostic task. Hadrina et al. [17] constructed a hidden Markov model for heart sound classification and achieved good experimental results. Wang et al. [18] used a combination of hidden Markov model (HMM) and MFCC features to classify abnormal heart sound signals. Ali et al. [19] used three integration techniques, namely Bagging, AdboostM1, and random subspace to improve the recognition rate of low performance-based classifiers.

In the course of rapid development of artificial intelligence algorithms, deep learning neural networks (DNN) have been explored for human heart sound classification in recent years. Unlike conventional heart sound classification algorithms, the particular advantage of deep learning algorithms lies in their feature extraction capabilities from complex heart sound signals. Bozkurt et al [20] extracted three different types of heart sound features, including Mel spectral map, MFCC, and subband envelope and compared with different feature fusion and segmentation strategies based on feed-forward CNN. A robust heart sound classification method combining a deep CNN feature extractor and a support vector machine (SVM) was presented and evaluated in Tschannen et al. [21], but no comparison with other models. Thomae et al. [22] proposed to develop deep end-to-end neural networks in which an RNN was constructed as a convolutional front end, but the model's structure is relatively simple and may not fully explore the important features in the heart sound signal. Latif et al. [23] proposed an RNN for abnormal heartbeat detection. They investigated the performance and computational complexity of four RNN models, namely Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU), Bidirectional Long Short-Term Memory (BLSTM), and Bidirectional Gated Recurrent Units (BGRU). Qaisar et al. [24] proposed a cardiovascular disease classification network using spectral feature maps obtained from continuous wavelet transforms and a self-aware transformer, but the dataset of this work includes only 250 samples, so the results have limited generalizability.

Although extensive work has been presented on cardiac tone classification, most of them suffered from degraded performance [25] because of the complicated and variable acoustic environment of the heart. The main reason for this is that many early works in the field of heart sound classification were based on low-order feature extraction methods, such as power spectral feature extraction based on Fourier transform, wavelet transform, etc. However, these methods are difficult to handle complex time-frequency information and are easily affected by interference factors such as signal noise, limiting their application in the field of heart sound classification. In addition, the single CNN network is not conducive to global modeling, and it is easy to segment the input heart sound signal into blocks for processing, which cannot retain the time series information well. Although CNNs-LSTM networks can better capture long sequence dependencies, the problems of larger model complexity and slower training speed when dealing with data involving high dimensionality also limit their application scope. Therefore, in the field of heart sound signal classification, more advanced data pre-processing and feature extraction methods combining multi-scale feature information for modeling are needed to further explore to better handle the complex information in heart sound signals and improve the classification accuracy and robustness.

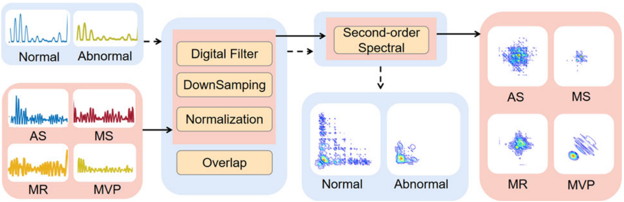


Fig. 1 Pre-processing and second-order spectral analysis feature extraction, where aortic stenosis (AS), mitral stenosis (MS), mitral regurgitation (MR), mitral valve prolapse (MVP) are four abnormal categories

3 Methodology

3.1 Solution design overview

In this paper, we propose an improved method for heart sound classification. Figure 1 shows pre-processing using digital filtering, downsampling, normalization, overlap, and feature extraction using second-order spectral analysis. This process contains all the pre-processing performed on the one-dimensional heart sound signal.

A classification model based on CNNs and transformer is proposed to perform heart sound classification, as shown in Fig. 2. The specific steps are as follows:

(1) Heart sound second-order spectral analysis feature map using the MHTE-4 module built by transformer to encode and aggregate contextual information.
(2) Heart sound second-order spectral feature map increases the diversity of local feature extraction and accelerates feature extraction by parallel two-way CNNs modules.
(3) The rich heart sound local feature map information extracted by parallel two-way CNNs and the global contextual information aggregated by MHTE-4 are fused by a fully connected network.
(4) Finally, the classification vector of fused features is used for the outcome prediction of the heart sound category. It is worth noting that the prediction head here is a single-branch prediction head. The reason is that the large differences in the sample data of the two tasks make it difficult to optimize into a model with an average view. And the multi-branch prediction head generates a large amount of computation and memory overhead, which leads to a slower model runtime.

3.2 Second-order spectral analysis

Feature extraction can extract useful information from the sound signal to better realize the processing of the sound signal. Short-Time Fourier Transform (STFT) [26] is a time-domain signal to frequency-domain signal transformation,

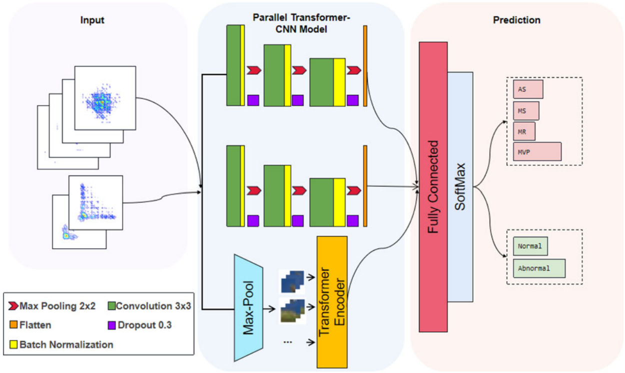


Fig. 2 Model training and inference pipeline

which splits the time-domain signal into multiple short time periods, and then performs Fourier transform on each short time period to obtain a series of frequency-domain signals. The STFT can be used to analyze the frequency characteristics of the time domain signal and the trend of different frequency components in the time domain signal. Wavelet Transform [27] decomposes a signal into components of different scales to better characterize the signal. It can be used to detect mutations in a signal, extract features in a signal, and detect noise in a signal. All of the above are low-order feature extraction methods.

Hiam et al. [28] have demonstrated that modern higher-order spectral analysis methods in the field of digital signal processing extract significantly better features than the results of lower-order feature extraction methods such as short-time Fourier transform and wavelet transform. Bispectrum is one of the higher-order spectral analyses of signals. It quantifies the degree of quadratic phase coupling (QPC) and nonlinearity interactions in non-stationary signals. Several types of medical signals are non-stationary signals, such as ECG, EEG, and PCG.

A two-dimensional feature matrix (256 × 256 × 1) can be generated using second-order spectral analysis, and we can also visualize the extracted feature matrix as a contour map (256 × 256 × 3) and a heat map (256 × 256 × 3).

$$S_2^x(\omega_1,\omega_2)=\sum_{\tau_1=-\infty}^{+\infty}\sum_{\tau_2=-\infty}^{+\infty}c_3^x(\tau_1,\tau_2)\exp(A), \tag{1}$$

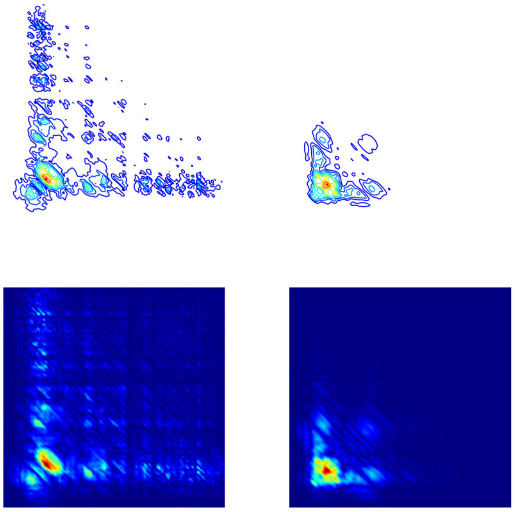
where

$$A=-j(\omega_1\tau_1+\omega_2\tau_2). \tag{2}$$

$$c_3^x(\tau_1,\tau_2)=E\{x(n)x(n+\tau_1)x(n+\tau_2)\} \tag{3}$$

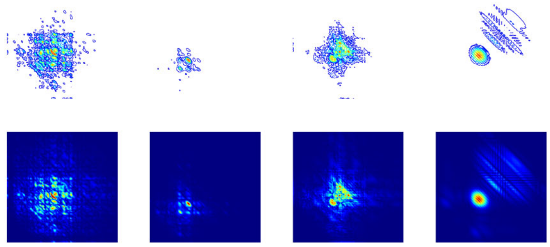
Eq. 1 represents the second-order Fourier change, and Eq. 3 is the third-order accumulation.

Figure 3 shows the contour map and heat map of normal and abnormal heart sounds extracted by the second-order spectral analysis, while Fig. 4 shows the contour map and heat map of four types of abnormal heart sounds extracted by the second-order spectral analysis. The feature maps obtained by second-order spectral analysis can be well distinguished.



(A) Normal (B) Abnormal

Fig. 3 The first line is a contour map and the second line is a heat map. **A** Bispectrum of normal class. **B** Bispectrum of abnormal class



(A) AS (B) MS (C) MR (D) MVP

Fig. 4 The first line is a contour map and the second line is a heat map. **A** Bispectrum of AS class. **B** Bispectrum of MS class. **C** Bispectrum of MR class. **D** Bispectrum of MVP class

3.3 PCTMF-Net

There are two effective approaches to audio classification. One approach is to convert audio data into fixed-length time series features, such as MFCC, and then pass them as input to a deep learning model. However, this approach may miss some critical temporal information when dealing with long audio data. Another approach is to extract the Mel Spectrogram of audio data using a model combining convolutional neural network (CNN) and recurrent neural network (RNN), which can extract the frequency and spatial features of audio data using convolutional layers, and then use recurrent layers

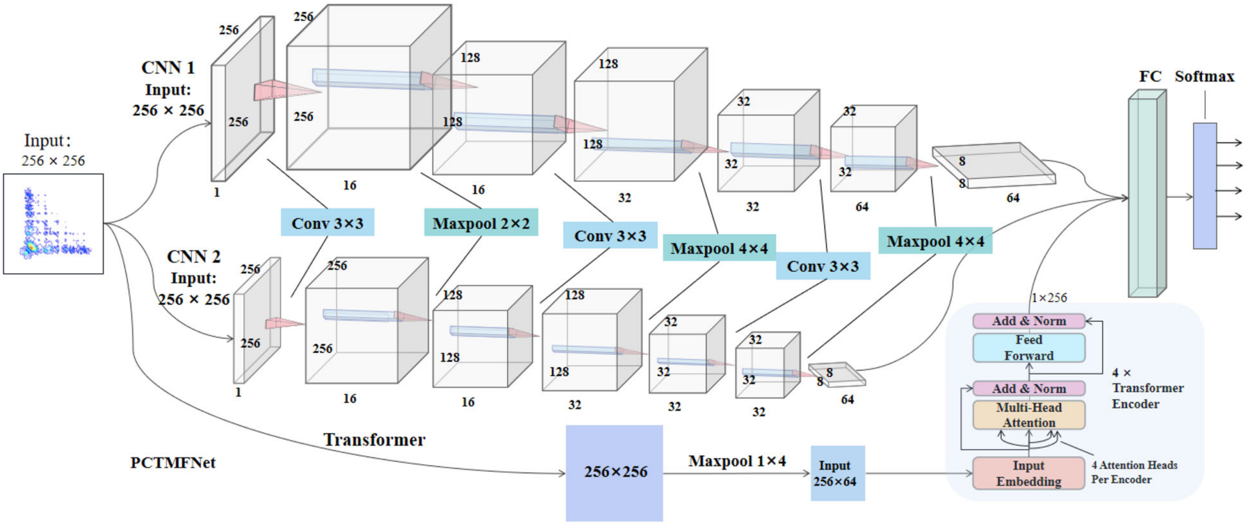


Fig. 5 PCTMF-Net, consisting of two-way parallel CNNs module and multi-head transformer encoder with four attention heads (MHTE-4) module

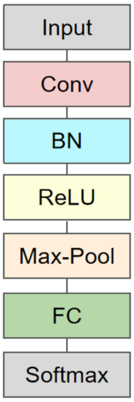
to process the time series. This approach is more effective in processing long audio data. However, LSTM and RNN sequence models cannot be computed in parallel and the computation time consumes long. Taking a separate CNN model cannot have a good global view to extract global features effectively.

For this purpose, combining CNNs and transformer is an effective way to explore. CNNs will extract the most expressive local feature representation at a low computational cost, while the transformer is used to encode and fuse information from the context, ultimately focusing on the global feature structure. Figure 5 shows the specific network design of PCTMF-Net. The fused features here come from a two-way parallel CNNs module and multi-head transformer encoder with four attention heads (MHTE-4) module, and this parallel CNNs and transformer structure can extract richer and more comprehensive features and improve the classification accuracy.

3.3.1 Two-way parallel CNNs module

In deep learning, CNNs are a class of artificial neural networks (ANN) commonly applied to image processing. CNNs are considered shift-invariant and spatially invariant and are based on a shared weight architecture of convolutional kernels or filters that slide along the input features and provide a translational isovariant response called the feature map. CNNs utilize multiple different levels of convolutional kernels to collect local features of an image for representation and have a unique advantage in extracting local features of an image. As the depth of the network increases, CNNs can enrich the extraction of hierarchical features and enhance their representation. Figure 6 depicts a one-layer CNN network, but adding more convolutional layers, pooling layers, and batch normalization layers can create a more sophisticated network.

Fig. 6 A simple one-layer CNN network composition



The VGGNet [29] has shown that using a large kernel across heavily stacked CNN layers is not cost-effective. Two benefits of using small stacked filters are computational efficiency and the expressivity of feature representation. Therefore, this paper utilizes small stacked filters. Many tasks have shown that a single CNN structure requires stacking a large number of convolutional layers, pooling layers, etc. to achieve good feature extraction, to reduce the difficulty of updating a large number of parameters during model learning and the computational speed during processing. We propose a parallel two-way CNNs structure. The parallel two-way CNNs can better learn different features from each convolutional layer to increase the diversity of information flow. Table 1 shows the structural design of the one-way CNN.

3.3.2 MHTE-4 module

Transformer [30] is a model proposed by Google in 2017, which is applied in the field of natural language process-

Table 1 Architecture of the one-way CNN Network

Layer	Output size	KSize	Stride	*P*
Input	256 × 256			
Conv1	256 × 256	3 × 3	1	
BN1		16		
MaxPool1	128 × 128	2 × 2	2	
Dropout1				0.3
Conv2	128 × 128	3 × 3	1	
BN2		32		
MaxPool2	32 × 32	4 × 4	4	
Dropout2				0.3
Conv3	32 × 32	3 × 3	1	
BN3		64		
MaxPool3	8 × 8	4 × 4	4	
Dropout3				0.3

ing, due to its powerful performance, it has been gradually introduced into computer vision. The transformer network structure is mainly composed of attention mechanisms, and its significant feature is the global receptive field. From another perspective, transformer is actually a special CNN, with a global feeling field. Figure 7 shows a complete transformer structure, which consists of two big structures, Encoder and Decoder.

In convolutional neural networks, convolution and pooling continuously refine the edges of the object for feature extraction. The encoder of the transformer corresponds to the convolution in a convolutional neural network, while the decoder structure is similar to the deconvolution. Both are used to extract features and perform feature map interactions. Transformer encodes the input data and then performs self-attention to generate a new feature vector, which is then mapped back to its original location by the decoder. In heart sound classification, convolutional neural networks can extract local features of heart sounds on a second-order spectral feature map, while the global features of the second-order spectral feature map are also important. The entire beating of the heart affects the entire frequency sequence, not just a time step.

In order to improve the classification of heart sound diseases, we propose the use of a transformer-encoder for global structure modeling of heart sound feature maps. By switching from the transformer's decoder to its encoder structure, neural networks can be made lighter and require fewer computer resources. Furthermore, the transformer has a stronger parameter-sharing capability, making it easier to share features, and thus better equipped to capture contextual information from heart sound interchanges. Our proposed structure, the MHTE-4, uses four transformer-encoder modules in the network, containing four independent self-attention heads each. The MHTE-4 aims to enhance the model's capability for multi-scale information modeling with stronger perceptual contextual semantics. Each self-attention head can learn separate attention weights that efficiently capture feature representations from different locations and achieve feature reconstruction prior to a specific downsampling layer, through a multi-head parallel structure.

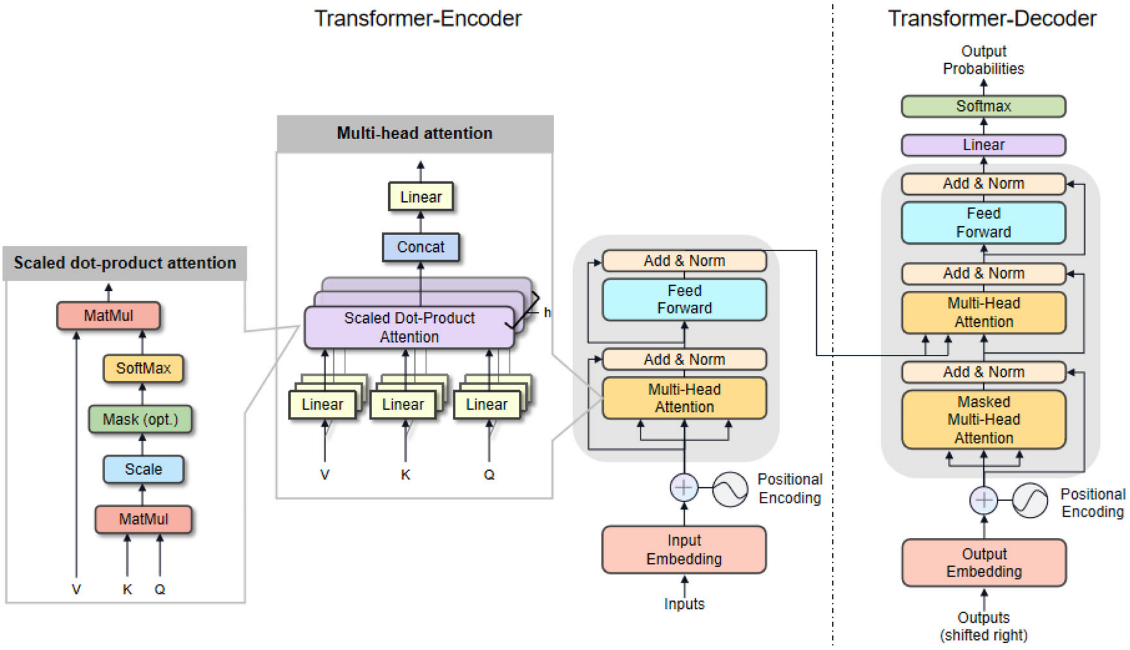


Fig. 7 Architecture of the transformer model

To reduce the computational effort, we use global pooling for the heart sound feature maps and then sample the features before sending them to the MHTE-4 in parallel. Input embeddings incorporate the features and provide location information for context. Furthermore, by leveraging multi-headed attention, the correlations between various heart sound features are learned to create numerous attention vectors. These vectors are then averaged and passed through a normalization layer to simplify the optimization process. Finally, the resulting vectors are given to a feedforward network that converts the data into dimensions readable by the fully connected layer.

4 Experiments and analysis

4.1 Implementation details

In this paper, the proposed algorithm was trained on an Ubuntu−20.04 64-bit operating system, using a 7 vCPU Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz and an NVIDIA RTX 3060 high-performance GPU with 12GB RAM on a single card. The model was built and trained using Pytorch 1.11.0, CUDA 11.3, and CUDNN 8.5. The initial learning rate was set to 0.0001, and the optimization method used was Adam with a batch size of 32. To improve training effectiveness, a cosine annealing restart learning rate mechanism was employed. This allowed the model to restart learning at a high learning rate after the optimized extremum was reached with a small learning rate, eventually converging at the 100th epoch.

4.2 Datasets

This paper presents two publicly available datasets of heart sound signals that have been labeled with appropriate categories.

Yaseen et al. [31] created a dataset of 5 categories of heart sound signals (PCG signals) from various sources in Table 2, containing one normal category (N) and four abnormal categories, the four abnormal categories being: aortic stenosis (AS), mitral stenosis (MS), mitral regurgitation (MR), mitral valve prolapse (MVP), with a total of 1000 audio files for the normal and abnormal categories (200 audio files/category) in wav file format. The length of each audio file is fixed at 2 s.

Liu et al. [32] created a dataset in Table 3 for the 2016 PhysioNet/CinC Computing Challenge. The archive comprises nine different heart sound databases sourced from multiple research groups around the world. It includes 2435 heart sound recordings in total collected from 1297 healthy subjects and patients with a variety of conditions, including heart valve disease and coronary artery disease. The recordings were collected from a variety of clinical or nonclinical (such as in-home visits) environments and equipment. The length of the recording varied from several seconds to several minutes.

Table 2 Yaseen dataset

Class	Heart status	Total
AS	Aortic stenosis	200
MS	Mitral stenosis	200
MR	Mitral regurgitation	200
MVP	Mitral valve prolapse	200

Table 3 PhysioNet dataset

Class	Original	Total(+Overlap)
Normal	2575	23, 839
Abnormal	665	7422

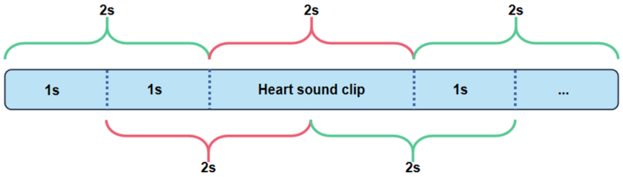


Fig. 8 Next cut with 2 s as a segment with 50% overlap

The criteria for producing audio data samples vary greatly from dataset to dataset. Such differences include audio sample rate, number of channels, length, noise reduction method, etc. These standards need to be unified to the maximum extent possible before fusing the datasets. Therefore, we pre-processed the audio files of the two publicly available datasets and collated them to obtain the processed datasets. The pre-processing process consists of three parts. To filter out high-frequency noise as well as DC noise, a second-order 25–400 Hz Butterworth median is used for digital filtering. To reduce the computational effort of the model, downsampling is performed to 1000 Hz. To reduce the large-scale differences between the audio files, the audio signal is normalized.

In addition to applying the same pre-processing steps to both public datasets, an overlapping method was employed as a data augmentation technique to generate sufficient anomalous and normal classification data. Specifically, each audio file was segmented into 2-second increments with a 50% overlap. Also, to obtain as many data samples as possible, a cut with 50% overlap was chosen. Figure 8 shows the overlap workflow, which intercepts 2 s as one record while overwriting the old 1 s and the new 1 s to form a new record.

Finally, we divide the pre-processed dataset in the ratio of 6: 2: 2 and use them as training set, validation set, and test

Table 4 Comparison of performance metrics of different models on MFCC feature extraction

Methods	Yaseen dataset				PhysioNet dataset			
	Accuracy (Top-1)	Precision	Recall	$F1$-score	Accuracy (Top-1)	Precision	Recall	$F1$-score
Tschannen [33]	0.667	0.761	0.666	0.638	0.676	0.740	0.675	0.696
Li [34]	0.762	0.804	0.762	0.747	0.705	0.705	0.706	0.705
Zheng [35]	0.763	0.812	0.763	0.748	0.750	0.774	0.751	0.761
Maknickas [36]	0.871	0.880	0.872	0.871	0.765	0.795	0.765	0.778
PCTMF-Net (ours)	0.925	0.941	0.929	0.930	0.827	0.788	0.817	0.790

set respectively. This ensures objectivity when evaluating the model performance.

4.3 Evaluation metrics

In order to evaluate the merits of the proposed method, some evaluation metrics were used for experimental comparisons.Commonly used concepts in evaluation metrics are expressed as follows:

- True Positive (TP): the number of positive classes predicted to be positive classes.
- True Negative (TN): the number of negative classes predicted as negative classes.
- False Positive (FP): the number of negative classes predicted as positive classes, which is the number of detection errors.
- False Negative (FN): the number of positive classes predicted as negative classes, which is the number of missed detections.

For a single-label task, each sample has only one correct category, and a prediction of that category is a correct classification, and a failure to predict is a misclassification, so the most intuitive indicator of classification is accuracy, which is calculated as follows:

$$\text{Accuracy} = \frac{n_{\text{correct}}}{n_{\text{total}}} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} \tag{4}$$

Precision rate is the ratio of the number of correctly classified positive samples to the number of samples determined to be positive by the classifier. The precision rate is a statistic for some samples, focusing on the data that the classifier determines as positive classes:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

Recall is the ratio of the number of correctly classified positive samples to the number of true positive samples. Recall is also a statistic for partial samples, focusing on the true positive class of samples:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

$F1$-score is the summed average of the precision and recall rates, which is defined as:

$$F1\text{-}score = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

4.4 Experimental results

The first dataset is a binary classification dataset, namely the S1–S2 heart sound classification dataset, used to determine whether the heart is normal or abnormal. The second dataset is the heart sound four-class classification dataset, used to classify heart sound signals into normal, systolic murmurs, diastolic murmurs, and early systolic diagnosis.

Table 4 shows the results of feature extraction and classification using MFCC. Our proposed PCTMF-Net achieves the best results in terms of accuracy, recall, and $F1$-score; however, these results are far from being up to the standard of being applied in reality.

Table 5 shows the results of feature extraction using the second-order spectral analysis and performing classification, and our proposed PCTMF-Net achieves the best results in terms of accuracy, precision, recall, and $F1$-score. Compared with MFCC for feature extraction, the second-order spectral analysis achieves better results, which proves that the second-order spectral analysis can better extract the features of heart sounds. Also, PCTMF-Net has better performance compared to the CNNs network obtained alone.

The confusion matrix is a summary of the predictions for a classification problem. The number of correct and incorrect predictions is summarized using numerical values and broken down by each category, which is the key to the confusion matrix. The confusion matrix shows which part of the classification model is confused when making predictions and provides insight not only into the errors made by the classification model, but more importantly, the types of errors that occur. If a model performs well, then the diagonal of the confusion matrix should have the maximum number

Table 5 Comparison of performance metrics of different models on second-order spectral analysis feature extraction

Methods	Yaseen dataset				PhysioNet dataset			
	Accuracy (Top-1)	Precision	Recall	$F1$-score	Accuracy (Top-1)	Precision	Recall	$F1$-score
Tschannen [33]	0.833	0.852	0.833	0.834	0.756	0.576	0.759	0.655
Li [34]	0.910	0.921	0.910	0.911	0.807	0.792	0.807	0.793
Zheng [35]	0.923	0.932	0.923	0.930	0.833	0.828	0.833	0.830
Maknickas [36]	0.974	0.976	0.974	0.974	0.847	0.841	0.847	0.843
PCTMF-Net (ours)	0.994	0.994	0.993	0.993	0.930	0.929	0.930	0.927

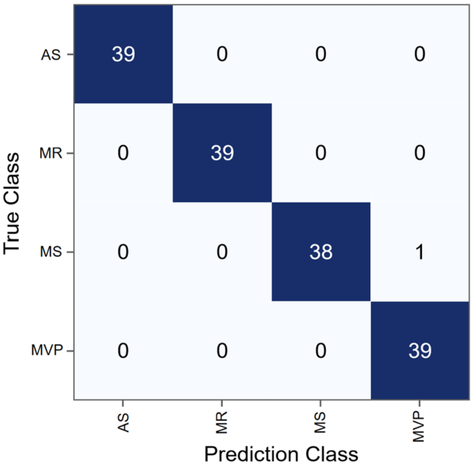


Fig. 9 Confusion matrix of PCTMF-Net on four-category dataset

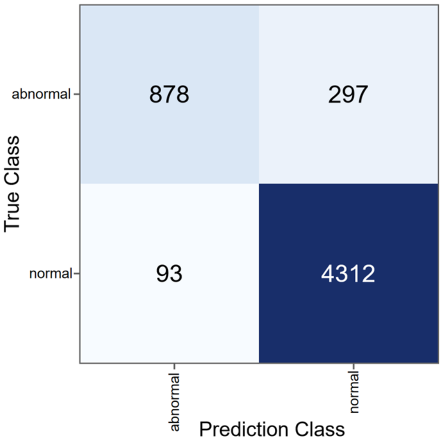
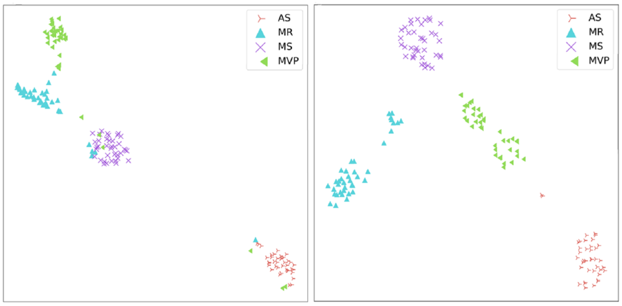


Fig. 10 Confusion matrix of PCTMF-Net on two-category dataset



(A) Maknickas [36] (B) PCTMF-Net(ours)

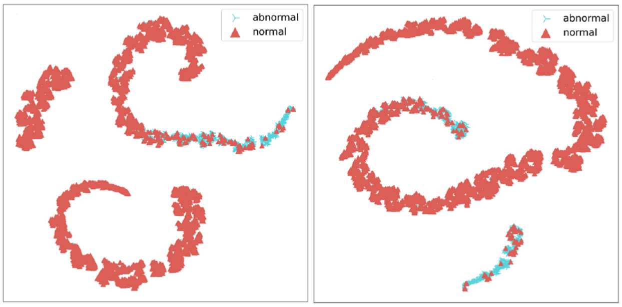
Fig. 11 Reduced dimensional presentation of the features of the four classification test set extracted by two models using t-SNE

of correctly classified samples. Figure 9 shows the results of PCTMF-Net for classifying samples on the four-category dataset and Fig. 10 shows the results of PCTMF-Net for classifying samples on the two-category dataset. It predicts the maximum number of correct samples in two tasks.

High-dimensional data visualization frequently employs the potent non-linear dimensionality reduction technique known as t-distributed Stochastic Neighbor Embedding (t-SNE) [33]. Minimizing the divergence between probability distributions over pairs of input and output points relies on mapping similar data points in high-dimensional space to their equivalents in low-dimensional space. Thus, t-SNE enables in-depth analysis of test set data by allowing us to visualize high-dimensional data in two or three dimensions, allowing us to see semantic distances between various classes of data points, identify outliers, categorize fine-grained issues, and view high-dimensional data structures.

The semantic properties of the input images were retrieved using feature maps created from the intermediate layers of the PCTMF-Net heart sound classification model. To better comprehend the connections and similarities between the samples in the test set, we then used t-SNE dimensionality reduction to these characteristics and plotted the outcomes in a two-dimensional coordinate system. This made it possible for us to examine the connections between various heart sound classes and helped us spot any possible misclassifications. Overall, the performance of the classification model was enhanced by the knowledge of the semantic aspects of the heart sound data using t-SNE.

Figures 11 and 12 show the classification performance of PCTMF-Net on a test set of both four and two classifications. In the low-dimensional visualization space, the classification features of the samples of different categories in the test set are obvious, and the samples of the same category are clustered together. This indicates that the second-order spectral analysis feature extraction and PCTMF-Net have good classification ability on heart tone classification.



(A) Maknickas [36] (B) PCTMF-Net(ours)

Fig. 12 Reduced dimensional presentation of two classification test set features extracted by two models using t-SNE

5 Conclusions

In this paper, PCTMF-Net is proposed for heart sound classification. The second-order spectral analysis is first used to extract higher-order features from the heart tone signal, and then, the higher-order extracted feature maps are classified using PCTMF-Net. PCTMF-Net uses a CNNs-transformer architecture, which employs two-way parallel CNNs to extract hierarchical features. Then, a transformer-based MHTE-4 module is designed to encode contextual information in multi-scale features, and contextual aggregation connections are applied to help the fusion and aggregation of features at different levels.

Experiments on Yaseen Dataset and PhysioNet Dataset demonstrate the feasibility of the proposed second-order spectral analysis and PCTMF-Net for heart sound classification. Comparison experiments performed on MFCC feature maps and second-order spectral feature maps show that higher-order extraction methods such as second-order spectral analysis have better characterization than lower-order extraction methods. The effectiveness of the combined two-way parallel CNNs and MHTE-4 is further validated by ablation studies. More specifically, the two-way parallel CNNs help to obtain local semantic properties of the heart sound feature map in terms of edges and morphology and to obtain efficient information flow interaction, while MHTE-4 can efficiently extract global information of the feature map and perform contextual information aggregation. The high accuracy of the two public datasets shows the potential of applying our solution as an aid in heart disease diagnosis.

Funding This work is supported by Macao Polytechnic University Grant (RP/FCA-15/2022).

Data Availability The two public datasets used in this thesis are both open access and the data sources are https://github.com/yaseen21khan/Classification-of-Heart-Sound-Signal-Using-Multiple-Features-/ and https://archive.physionet.org/pn3/challenge/2016/. We have used these datasets to test our findings and have provided relevant citations in the article. We confirm that the data have been used and interpreted correctly and that no data manipulation has taken place. Details of the datasets and how to use them can be found on the datasets' official websites. We hereby declare that the data used in this study are publicly available and that other researchers are free to access and use them to further explore relevant questions.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Deng, M., Meng, T., Cao, J., Wang, S., Zhang, J., Fan, H.: Heart sound classification based on improved MFCC features and convolutional recurrent neural networks. Neural Netw. **130**, 22–32 (2020)
2. Jagannathan, R., Patel, S.A., Ali, M.K., Venkat Narayan, K.M.: Global updates on cardiovascular disease mortality trends and attribution of traditional risk factors. Curr. Diab. Rep. **19**, 1–12 (2019)
3. Hayes, S.N., Kim, E.S.H., Saw, J., Adlam, D., Arslanian-Engoren, C., Economy, K.E., Ganesh, S.K., Gulati, R., Lindsay, M.E., Mieres, J.H., et al.: Spontaneous coronary artery dissection: current state of the science: a scientific statement from the American Heart Association. Circulation **137**(19), e523–e557 (2018)
4. Krishnan, P.T., Balasubramanian, P., Umapathy, S.: Automated heart sound classification system from unsegmented phonocardiogram (PCG) using deep neural network. Phys. Eng. Sci. Med. **43**, 505–515 (2020)
5. Ismail, S., Siddiqi, I., Akram, M.U., Akram, U., Akram, U., Akram, U.: Localization and classification of heart beats in phonocardiography signals—a comprehensive review. EURASIP J. Adv. Signal Process. **1**, 1–27 (2018)
6. Kumar, D., Carvalho, P., Antunes, M., Henriques, J., Eugénio, L., Schmidt, R., Habetha, J.: Detection of s1 and s2 heart sounds by high frequency signatures. In: 2006 International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 1410–1416. IEEE (2006)
7. Zeng, Y., Shudong, X.: New auscultation: can we detect heart failure by auscultation. J. Commun. Med. Pub. Health Rep. **3**, 1 (2022)
8. Deng, M., Meng, T., Cao, J., Wang, S., Zhang, J., Fan, H.: Heart sound classification based on improved MFCC features and convolutional recurrent neural networks. Neural Netw. **130**, 22–32 (2020)
9. Oliveira, J., Nogueira, D., Renna, F., Ferreira, C., Jorge, A.M., Coimbra, M.: Do we really need a segmentation step in heart sound classification algorithms? In: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 286–289. IEEE (2021)
10. Fatmawati, T.Y., Yuliani, A., Afandi, M.A., Zulherman, D.: Comparative analysis of the phonocardiogram denoising system based-on empirical mode decomposition (EMD) and double-density discrete wavelet transform (DDDWT). In: Proceedings of the 1st International Conference on Electronics, Biomedical Engineering, and Health Informatics: ICEBEHI 2020, 8–9 October, Surabaya, Indonesia, pp. 593–604. Springer (2021)
11. Zabihi, M., Rad, A.B., Kiranyaz, S., Gabbouj, M., Katsaggelos, A.K.: Heart sound anomaly and quality detection using ensemble of neural networks without segmentation. In: 2016 Computing in Cardiology Conference (CINC), pp. 613–616 (2016)
12. Schmidt, S.E., Holst-Hansen, C., Hansen, J., Toft, E., Struijk, J.J.: Acoustic features for the identification of coronary artery disease. IEEE Trans. Biomed. Eng. **62**(11), 2611–2619 (2015)

13. Kumar, D., Carvalho, P., Antunes, M., Henriques, J., Eugenio, L., Schmidt, R., Habetha, J.: Detection of s1 and s2 heart sounds by high frequency signatures. In: 2006 International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 1410–1416 (2006)
14. Potes, C., Parvaneh, S., Rahman, A., Conroy, B.: Ensemble of feature-based and deep learning-based classifiers for detection of abnormal heart sounds. In: 2016 Computing in Cardiology Conference (CINC), pp. 621–624 (2016)
15. Nilanon, T., Yao, J., Hao, J., Purushotham, S., Liu, Y.: Normal/abnormal heart sound recordings classification using convolutional neural network. In: 2016 Computing in Cardiology Conference (CINC), pp. 585–588 (2016)
16. Stasis, A.C., Loukis, E.N., Pavlopoulos, S.A., Koutsouris, D.: Using decision tree algorithms as a basis for a heart sound diagnosis decision support system. In: 4th International IEEE EMBS Special Topic Conference on Information Technology Applications in Biomedicine, pp. 354–357 (2003)
17. Hadrina Sh-Hussain, M.M., Mohamad, R.Z., Ting, C.-M., Ismail, K., Numanl, F., Hussain, H., Rasul, S.: Classification of heart sound signals using autoregressive model and hidden Markov model. J. Med. Imaging Health Inf. **7**(4), 755–763 (2017)
18. Wang, P., Lim, C.S., Chauhan, S., Foo, J.Y.A., Anantharaman, V.: Phonocardiographic signal analysis method using a modified hidden Markov model. Ann. Biomed. Eng. **35**, 367–374 (2007)
19. Ali, S., Adnan, S.M., Nawaz, T., Obaid Ullah, M., Aziz, S.: Human heart sounds classification using ensemble methods. University of Engineering and Technology Taxila. Tech. J. **22**(1), 113 (2017)
20. Bozkurt, B., Germanakis, I., Stylianou, Y.: A study of time-frequency features for CNN-based automatic heart sound classification for pathology detection. Comput. Biol. Med. **100**, 132–143 (2018)
21. Tschannen, M., Kramer, T., Marti, G., Heinzmann, M., Wiatowski, T.: Heart sound classification using deep structured features. In: 2016 Computing in Cardiology Conference (CINC), pp. 565–568 (2016)
22. Thomae, C., Dominik, A.: Using deep gated RNN with a convolutional front end for end-to-end classification of heart sound. In: 2016 Computing in Cardiology Conference (CINC), pp. 625–628 (2016)
23. Latif, S., Usman, M., Rana, R., Qadir, J.: Phonocardiographic sensing using deep learning for abnormal heartbeat detection. IEEE Sens. J. **18**(22), 9393–9400 (2018)
24. Abbas, Q., Hussain, A., Baig, A.R.: Automatic detection and classification of cardiovascular disorders using phonocardiogram and convolutional vision transformers. Diagnostics **12**(12), 3109 (2022)
25. Clifford, G.D., Liu, C., Moody, B.E., Roig, J.M., Schmidt, S.E., Li, Q., Silva, I., Mark, R.G.: Recent advances in heart sound analysis. Physiol. Meas. **38**, E10–E25 (2017)
26. Durak, L., Arikan, O.: Short-time Fourier transform: two fundamental properties and an optimal implementation. IEEE Trans. Signal Process. **51**(5), 1231–1242 (2003)
27. Zhang, D., Zhang, D.: Wavelet transform. In: Fundamentals of Image Data Mining: Analysis, Features, Classification and Retrieval, pp. 35–44 (2019)
28. Alquran, H., Alqudah, A.M., Abu-Qasmieh, I., Al-Badarneh, A., Almashaqbeh, S.: ECG classification using higher order spectral estimation and deep learning techniques. Neural Netw. World **29**(4), 207–219 (2019)
29. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014). arXiv preprint arXiv:1409.1556
30. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. Adv. Neural Inf. Process. Syst. **30** (2017)
31. Yaseen, G.Y.S., Kwon, S.: Classification of heart sound signal using multiple features. Appl. Sci. **8**(12), 2344 (2018)
32. Liu, C., Springer, D., Li, Q., Moody, B., Juan, R.A., Chorro, F.J., Castells, F., Roig, J.M., Silva, I., Johnson, A.E.W., Syed, Z., Schmidt, S.E., Papadaniil, C.D., Hadjileontiadis, L., Naseri, H., Moukadem, A., Dieterlen, A., Brandt, C., Tang, H., Samieinasab, M., Samieinasab, M.R., Sameni, R., Mark, R.G., Clifford, G.D.: An open access database for the evaluation of heart sound algorithms. Physiol. Meas. **37**(12), 2181 (2016)
33. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. J. Mach. Learn. Res. **9**, 11 (2008)
34. Li, L., Wang, X., Du, X., Liu, Y., Liu, C., Liu, C., Qin, C., Li, Y.: Classification of heart sound signals with bp neural network and logistic regression. In: ACM Cloud and Autonomic Computing Conference (2017)
35. Zheng, Y., Guo, X., Qin, J., Xiao, S.: Computer-assisted diagnosis for chronic heart failure by the analysis of their cardiac reserve and heart sound characteristics. Comput. Methods Programs Biomed. **122**, 372–383 (2015)
36. Maknickas, V., Maknickas, A.: Recognition of normal-abnormal phonocardiographic signals using deep convolutional neural networks and mel-frequency spectral coefficients. Physiol. Meas. **38**, 1671 (2017)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Rongsheng Wang was born in Shandong, China, in 1999. He obtained Bachelor's degree in Engineering from Henan Polytechnic University in 2022. Currently, he is a master's student in the Big Data and IoT program at the Macao Polytechnic University, with a focus on medical image processing and medical AI.

Yaofei Duan received her B.S. degree in Qilu University of Technology, Shandong, China, in 2021. Now she is a master candidate with the faculty of applied sciences in Macao Polytechnic University, Macao, China. Her research interests include medical image processing and deep learning.

Yukun Li was born in Anhui, China, in 1997. He obtained his bachelor's degree from Guangzhou Huashang College of in 2022. He is currently a master's student in the Big Data and Internet of Things program at the Macao Polytechnic University, with a main research focus on medical patient data mining.

Dashun Zheng was born in Tianjin, China, in 1999. He obtained his Bachelor's degree from Xinjiang University of Finance and Economics in 2022. Currently, he is a master's student in the Big Data and IoT program at the Macao Polytechnic University, with a focus on natural language processing and medical AI.

Xiaohong Liu is Assistant Professor at the John Hopcroft Center for Computer Science at Shanghai Jiao Tong University. His research interests primarily lie in computer vision, including image and video enhancement, quality assessment, and segmentation. He received his Ph.D. degree in Electrical and Computer Engineering from McMaster University in November 2021.

Chan Tong Lam received the B.Sc. and M.Sc. degrees from Queen's University, Kingston, ON, Canada, in 1998 and 2000, respectively, and the Ph.D. degree from Carleton University, Ottawa, ON, Canada, in 2007. He is currently an Associate Professor with the Faculty of Applied Sciences, Macao Polytechnic University, Macao SAR, China. His research interests include mobile wireless communications, machine learning in communications, and computer vision in smart city.

Tao Tan is an Associate Professor at the Faculty of Applied Sciences at the Macao Polytechnic University. He is primarily involved in scientific research on precision prevention and treatment of breast cancer. He received his Ph.D. degree from Radboud University and has worked as a Senior AI Scientist at the global healthcare giant GE Healthcare.