

Research Paper

A deep neural network approach to heart murmur detection using spectrogram and peak interval features

Soyul Han^{a,b,1}, Taein Kang^{b,1}, Jungguk Lee^{b,1}, Narin Kim^b, Hyejin Won^b, Yeong-Hwa Kim^b, Wuming Gong^{c,*}, Il-Youp Kwak^{b,*}

^a Institute for Data Innovation in Science, Seoul National University, Seoul, 08826, South Korea

^b Department of Statistics and Data Science, Chung-Ang University, Seoul, 06974, South Korea

^c Lillehei heart institute, University of Minnesota, MN, 55455, United States

ARTICLE INFO

Dataset link: <https://moody-challenge.physionet.org/2022/>

Keywords:

Heart murmur detection
Heart anomaly detection
Multi-input classification
Deep learning
Convolutional neural network

ABSTRACT

Congenital heart disease affects about 1% of newborns, posing risks like heart failure and mortality. Developing countries often lack resources for diagnosis and treatment. The George B. Moody PhysioNet Challenge 2022 aims to develop systems for detecting murmurs and clinical outcome events using phonocardiogram (PCG) data, offering a cost-effective method for diagnosing cardiac diseases without invasive procedures. We proposed a deep learning model for this task and achieved the 5th rank out of 40 teams in both Track 1 (murmur detection) and Track 2 (clinical outcome prediction). This paper describes our methods and additional experiments. To extract features from the PCG, we employed techniques including Constant Q Transform (CQT) and Mel-scaled spectrogram (Mel-spectrogram) to generate two-dimensional representations in the frequency–time domain. Additionally, we extracted the Peak Interval (PI) feature, which measures the distance between peaks in the PCG data. This feature is useful because the peak intervals should be shorter in PCG recordings with murmurs. We also considered the sequence and mean of PI as additional features. Our proposed system, titled ‘Phonocardiogram-based Heart murmur Detection using Spectrogram and PI features (SpectroHeart),’ employs the Mel-spectrogram and PI features to detect heart murmurs and assess clinical outcomes. We believe that our deep learning-based system has great potential for automatically detecting heart signals from PCG.

1. Introduction

Cardiac auscultation is an essential process for the diagnosis of cardiovascular diseases, and recent advancements in digital phonocardiography have enabled more precise and efficient diagnostics. However, the interpretation of heart sounds and phonocardiograms (PCG) often relies on the expertise of clinicians, leading to a low concordance in diagnoses (Singh and Anand, 2007; Wen et al., 2014). In resource-limited settings, accurate screening becomes even more challenging (Wen et al., 2014). To address these issues, there is a pressing need for automated diagnostic systems that can provide rapid and accurate assessments, particularly in areas with limited medical resources.

The 2022 George B. Moody PhysioNet Challenge was conducted to develop algorithms capable of automatically detecting cardiac murmurs and abnormal heart functions using PCG recordings. As illustrated in Fig. 1, the challenge comprises two distinct tracks: the first aims to identify the presence of murmurs in PCG recordings, while the

second focuses on differentiating between normal and abnormal clinical outcomes based on various auscultation positions.

In this study, we introduce a novel approach that transforms PCG data into two-dimensional spectrograms, and we propose the use of a new concept, ‘Peak Interval (PI)’, inspired by the RR intervals utilized in Electrocardiography (ECG) to describe heart rate variability (HRV) (El-Segaier et al., 2005; Tsipouras et al., 2005; Faust et al., 2018). This leads to the development of the ‘SpectroHeart’ system, designed for the detection of cardiac murmurs through spectrogram and PI features.

The SpectroHeart system analyzes the variations in intervals between key peaks in the cardiac cycle by modeling spectrogram features, PI features, and patient demographic characteristics, thereby facilitating the detection of cardiac murmurs. The PI feature, a novel indicator that well reflects HRV, can be extracted from the PCG signal. PI represents the time interval between the major peaks (S1 and S2) of the cardiac cycle. In patients with heart murmurs, the abnormal

* Corresponding authors.

E-mail addresses: soyul5458@snu.ac.kr (S. Han), xodls4179@cau.ac.kr (T. Kang), daedoljg@cau.ac.kr (J. Lee), nrgolden@cau.ac.kr (N. Kim), whj9492@cau.ac.kr (H. Won), gogators@cau.ac.kr (Y.-H. Kim), gongx030@umn.edu (W. Gong), ikwak2@cau.ac.kr (I.-Y. Kwak).

¹ Authors contributed equally.

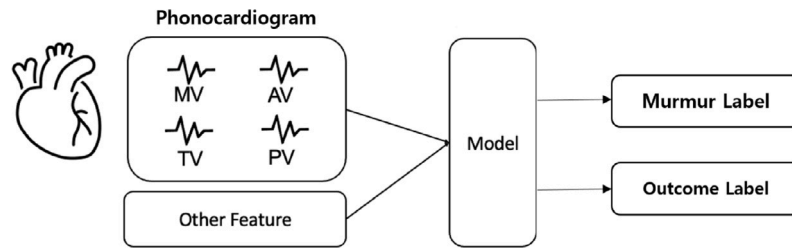


Fig. 1. Problem definition: Implementation of a model for heart murmur detection and clinical outcome identification using PCG data and demographic information.

blood flow patterns caused by valve problems generate new peak points and shorten the intervals between peak points. These changes in peak intervals play a crucial role in detecting heart murmurs.

To develop the SpectroHeart system, we considered LCNN and ResMax models (Wu et al., 2018; Kwak et al., 2021) for spectrogram features of PCG data and extracted PI features. Both the LCNN and ResMax models demonstrated improved performance in terms of weighted accuracy and cost evaluation metrics when incorporating PI features. In the PhysioNet Challenge, a model utilizing only the mean of PI was used (Lee et al., 2022), achieving 5th place in both subtracks that evaluated weighted accuracy and expected cost. Following the competition, additional experiments were conducted with models using the entire PI sequence, which showed enhanced performance compared to the model that only used the mean of the PI. We have uploaded the code for this paper to github.²

This study contributes to the development of a system for early diagnosis of heart diseases by analyzing cardiac sounds using spectrograms, PI, and AI algorithms. In particular, the SpectroHeart system aims to accurately identify the characteristics of heart murmurs, thereby aiding in the early detection of serious conditions such as congenital heart disease. However, this research requires validation across diverse regions and datasets, and future studies will need to address these limitations.

Section 2 of our study provides a comprehensive review of the literature relevant to our research. In Section 3, we introduce the SpectroHeart system and provide detailed explanations about its functioning. Moving on to Section 4, we describe the PCG dataset and outline the experimental setup that we employed. Furthermore, we discuss the experimental results, the final model, and the results that were submitted to the competition. In the Discussion section of Section 5, we compare our model with other models that participated in the competition. Lastly, Section 6 serves as a conclusion where we summarize our findings and discuss potential avenues for future research.

2. Related works

In the previous PhysioNet Challenge in 2016, diagnosis of cardiac conditions was determined using heart sound recordings from multiple locations, similar to task 2 of the George B. Moody PhysioNet Challenge 2022 (Liu et al., 2016; Clifford et al., 2016). The emphasis of the 2016 competition was on preprocessing and feature extraction from heart sound recordings. The use of Machine Learning models and shallow neural networks underscored the significance of feature extraction. Teams predominantly employed bandpass filtering and other one-dimensional signal preprocessing techniques for signal segmentation. Feature extraction methodologies included wavelet transforms and Mel-Frequency Cepstral Coefficients (MFCC), with the leading team utilizing discrete-time Fourier transform to analyze frequency components (Potes et al., 2016). To enhance model performance and robustness, sampling methods like K-fold cross-validation and bootstrapping were utilized, alongside an ensemble of machine learning

models and shallow Convolutional Neural Network (CNN)-based classifiers (Potes et al., 2016; Zabihi et al., 2016; Kay and Agarwal, 2016; Bobillo, 2016; Homsy et al., 2016).

However, the advent of Deep Neural Networks has marked a paradigm shift in preprocessing, feature extraction techniques, and model architectures, highlighting limitations in previous approaches (Reyna et al., 2022). While earlier studies primarily focused on the segmentation of one-dimensional signals, recent research has introduced advanced techniques that decompose the signals using a Gabor dictionary and reconstruct them as time–frequency 2D matrices (Fakhry and Gallardo-Antolín, 2024). Additionally, spectrograms and other visualization techniques are being utilized to analyze the signals (Han et al., 2023a). The model architectures have also evolved from traditional statistical models to hybrid models and deep learning models (Fakhry and Gallardo-Antolín, 2024; Han et al., 2023a; Kong et al., 2024).

The implementation of data augmentation techniques has emerged as a crucial method to increase dataset size and improve model robustness (Kim et al., 2022; Ballas et al., 2022; Knorr and Bremer, 2022; Lu et al., 2022). Efforts to enhance data quality have also been evident, with strategies such as re-labeling or integrating hierarchical features within the model (Xu et al., 2022), and adopting a two-step classification process for more accurate label differentiation (Bruoth et al., 2022).

This study distinguishes itself from previous research by leveraging these advancements in signal processing, data augmentation, and model architecture. By adopting a comprehensive approach towards feature extraction and employing state-of-the-art Deep Learning techniques, this research aims to overcome the limitations observed in prior studies.

Specifically, the introduction of two-dimensional spectrograms and the novel ‘PI’ concept for depicting HRV in PCG data represent significant advancements. These innovations not only enhance the performance of analysis but also contribute to the development of more accurate and robust diagnostic models, thereby addressing the critical gaps identified in earlier research efforts.

3. Methods

Fig. 2 illustrates the architecture of the SpectroHeart system, which was designed for the detection of murmurs and clinical outcomes. This system extracts two types of features from raw PCG signal data: 2D spectrogram and PI. In PI processing, either the PI mean or the PI sequence was utilized. Additionally, the use of demographic data was examined as an additional feature, and it showed performance improvements in clinical outcome classification. While demographic data is incorporated in the detection of clinical outcomes, it is not utilized in the detection of murmurs.

3.1. Spectrogram

The heartbeat sound data is a form of unstructured data that contains multiple frequencies. Thus, it is appropriate to use spectrogram features to represent the sound characteristics accurately. A spectrogram is a time–frequency representation that visually demonstrates the

² <https://github.com/ikwak2/SpectroHeart>.

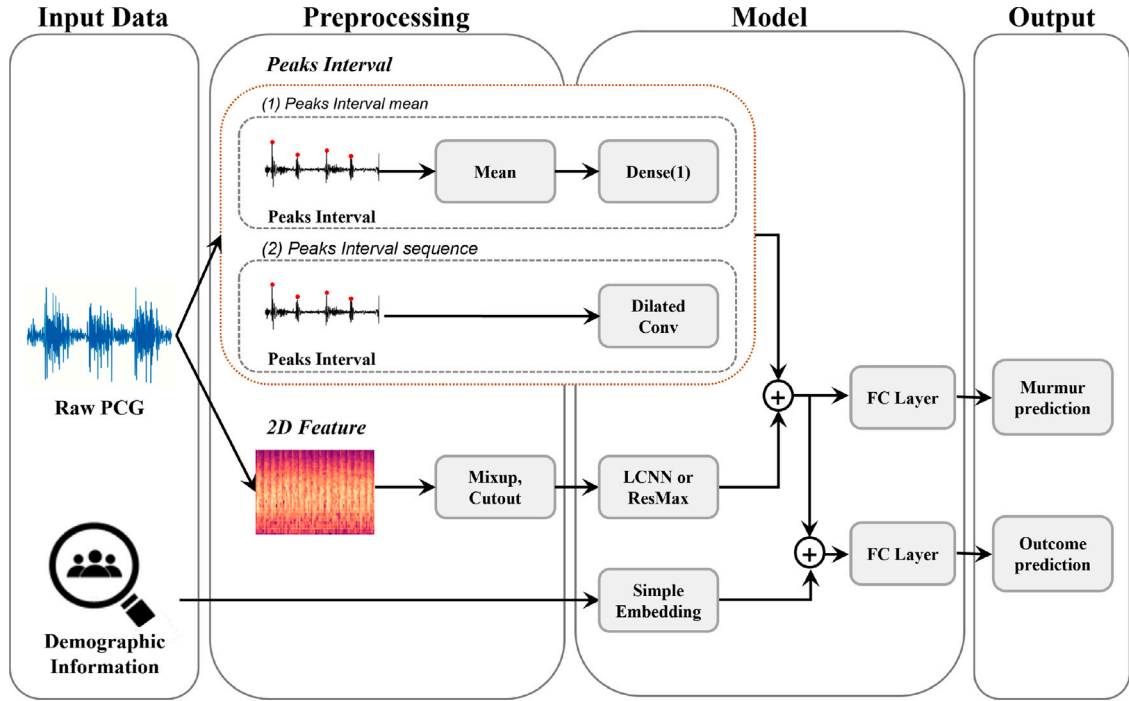


Fig. 2. SpectroHeart system architecture: Feature extraction from raw PCG signals includes 2D spectrogram and PI (mean or sequence). Demographic data, used for clinical outcome classification but not murmur detection.

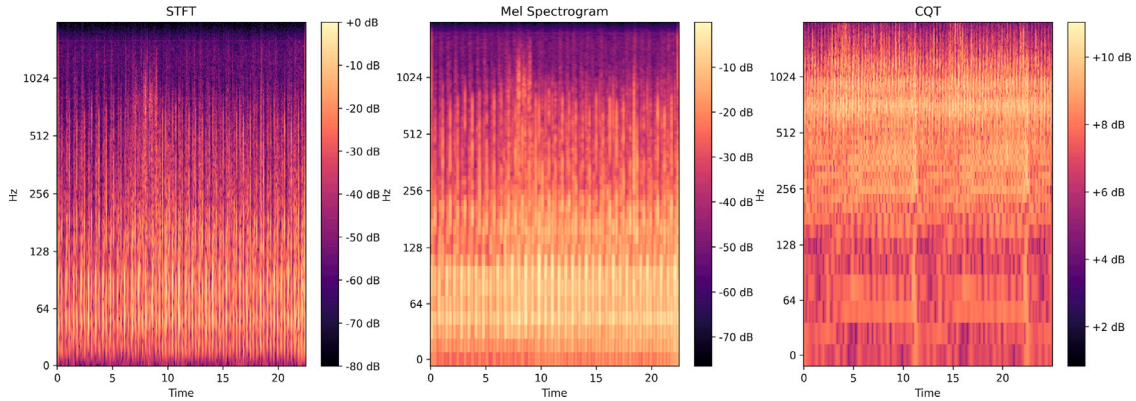


Fig. 3. Three methods for feature extraction from PCG signals: STFT, Mel-spectrogram, and CQT.

frequency variations of a signal. It is primarily used for analyzing audio signals such as speech and sound, providing an intuitive understanding and analysis of the signal's frequency characteristics. In a spectrogram, time is represented on the x -axis, frequency on the y -axis, and the third dimension, energy or decibels, is depicted through color or intensity. In this study, we employ three types of spectrograms: Short-Time Fourier Transform (STFT), Mel-scaled spectrogram (Mel-spectrogram), and Constant Q Transform (CQT). The visual comparison of these spectrograms is illustrated in Fig. 3.

3.1.1. Short-Time Fourier Transform (STFT)

The Short-Time Fourier Transform (STFT) is a powerful tool for analyzing signals in the time–frequency domain, making it useful for tracking the frequency components of a signal over time. The STFT divides the signal into short time segments and performs a Fourier transform on each segment, thereby providing both time and frequency information simultaneously. This allows for a detailed analysis of the

temporal variations of the signal from a frequency perspective. STFT can be represented mathematically as shown in Eq. (1).

$$X(t, \omega) = \int_{-\infty}^{\infty} x(\tau) w(\tau - t) e^{-j\omega\tau} d\tau \quad (1)$$

where $x(\tau)$ is the original signal, $w(\tau - t)$ is the window function centered at time t , and ω is the angular frequency. The window function $w(\tau)$ is commonly used in various forms such as Gaussian window, Hamming window, and Hanning window.

3.1.2. Mel-scaled spectrogram

The Mel-scaled spectrogram is a type of spectrogram that uses the mel scale to map the frequency spectrum of sound signals to a perceptual scale that more closely matches how humans hear sound. It is obtained by first performing the STFT and then applying a Mel-scale filter to convert the power of each frequency band into a logarithmic scale. The Mel-spectrogram is calculated by converting the power information of each frequency band from the generated STFT spectrogram

into the Mel scale as shown in Eq. (2).

$$Mel(f) = 2595 \log_{10} \left(1 + \frac{f}{700} \right) \quad (2)$$

$$LogMel(f) = 10 \log_{10} \left(\frac{Mel(f)}{ref} \right) \quad (3)$$

Here, f represents the frequency. The Mel-spectrogram in dB is obtained by taking the logarithm of the Mel-spectrogram's power as shown in Eq. (3). In this context, ref is a reference value to which $Mel(f)$ is relatively scaled.

3.1.3. Constant Q Transform (CQT)

The Constant Q Transform (CQT) was proposed for the purpose of music processing (Brown, 1991). CQT maintains a constant Q-factor (the ratio of the center frequency to bandwidth) across the entire frequency range. Unlike the STFT, which has a constant frequency resolution, the CQT's time resolution varies with frequency. The CQT has the following characteristics: First, the frequency resolution increases gradually from low to high frequencies. Second, it provides higher time resolution at higher frequencies, resulting in better performance in time–frequency analysis. Last, it is advantageous for analyzing signals that follow a logarithmic frequency scale, such as musical signals. Due to these characteristics, the CQT is widely used in music signal processing and analysis, as it allows for effective analysis of various frequency bands by adjusting the resolution along the frequency axis.

3.2. Peak detection method for peak interval

One of the representative methods for diagnosing the condition of the heart is the RR interval, which examines changes in the heart cycle. It looks at the intervals between R-peak points in ECG data, making it easy to observe changes in heart cycles. Patients with heart murmurs generate noise due to abnormal patterns of blood flow in the heart caused by problems with the heart valve. The creation of noise is likely to generate new peak points and the interval between peak points will be short. Based on this, peak points were identified and the interval between them, called the peak interval, was calculated and used as another feature.

To perform peak detection, two parameters are required: 'min distance' and 'threshold'. The 'min distance' is the minimum distance at which peaks can be detected. If multiple peaks are detected within this distance, they are not counted as separate peaks. The 'threshold' is a threshold parameter. Peaks that do not exceed the threshold are not detected, and typically the threshold is a positive real number greater than 0. Fig. 4(a) shows a typical PCG signal, while Fig. 4(b) displays the same signal post-application of the peak detection method, with parameters adjusted to capture all peak points. Fig. 4(c) displays the signal with parameters adjusted to detect only peak points that correspond to S1–S1 intervals.

3.3. Data augmentation

Our research involved experimenting with various data augmentation techniques such as Mixup, Cutout, CutMix, FFM, and SpecAugment. Mixup and CutMix involve altering or combining specific regions of the data, while Cutout, FFM, and SpecAugment are masking techniques that hide specific regions. The technique depends on which parts of the data are selected for manipulation. For our final submission, we selected Mixup and Cutout as the data augmentation techniques. Fig. 5(a) and (b) show Mixup and Cutout, respectively, both of which were implemented as online training.

3.3.1. Mixup

Mixup (Zhang et al., 2017) is an augmentation technique that creates new samples by mixing the inputs and labels of two different samples, as shown in Fig. 5(a). When generating a new sample, two existing samples are multiplied by λ and $1-\lambda$, respectively, where λ is determined by drawing from a Beta distribution $\lambda \sim Beta(\alpha, \alpha)$, $\alpha \in (0, \infty)$ with hyperparameter α set to 0.7. With the chosen λ , one sample is multiplied by λ and the other sample is multiplied by $1-\lambda$, and the two are added to create a new input and label as a sample. This can be represented mathematically as shown in Eq. (4).

$$\begin{aligned} \tilde{X} &= \lambda X_i + (1 - \lambda) X_j \\ \tilde{y} &= \lambda y_i + (1 - \lambda) y_j \end{aligned} \quad (4)$$

3.3.2. Cutout

Fig. 5(b) shows the augmented data with Cutout, which is a data augmentation technique that masks out a randomly selected rectangular patch of the 2D feature by setting its value to 0. DeVries and Taylor (2017) used a square-shaped patch and its size was set differently for each dataset by grid search as a hyperparameter. In our case, we used a rectangular patch with the same aspect ratio as the 2D feature, and the size was determined by setting a hyperparameter λ to $\sqrt{1 - \lambda}$ proportion of the patch's size. Like Mixup, we determined λ from the Beta distribution with a hyperparameter α of 0.7.

3.4. Model

We experimented with two model architectures, LCNN and ResMax, for processing 2D spectrogram features. These models have shown high performance in Voice Spoofing Detection (Lavrentyeva et al., 2017, 2019; Kwak et al., 2021, 2022, 2023; Han et al., 2023b), and we hypothesized their effectiveness would translate to PCG data due to their shared characteristic as sound data. The role of the LCNN and ResMax models are to capture heart murmur-related embedding from the spectrogram features. During the training process, the LCNN model will learn to identify features associated with heart disease.

The proposed system architecture extracts Mel-spectrogram and PI features from PCG data. For the PI feature, we consider both the average and sequence information. The average information is modeled with a single-dense layer, similar to the processing of demographic information. For sequence information, we propose a new Dilated Convolution structure for 1D signals. Patients with murmurs are expected to exhibit more peaks due to the murmur, and there may be local patterns within the peak interval sequences for these cases. By utilizing Dilated Convolution, we anticipated that the dependencies within peak interval sequences could be effectively modeled.

3.4.1. LCNN

As one of the methods to model the spectrogram in the SpectroHeart model, LCNN was considered. In contrast to the Light CNN-9 model proposed by Wu et al. (2018), this study employs a more extensive LCNN model that involves nine LCNN blocks. Each LCNN block comprises convolution, Max Feature Map (MFM), and an optional batch normalization layer, as depicted in Fig. 6(a). Fig. 6(b) presents the overall architecture of the LCNN model for spectrogram feature. The implemented LCNN model used 32, 32, 48, 48, 64, 64, 32, 32, and 32 convolution filters, where the kernel size of the first convolution layer was set to 5, while the other convolution layers were set to 1 or 3.

3.4.2. ResMax

To model the spectrogram in the SpectroHeart model, in addition to LCNN, the use of the ResMax model was also considered. The ResMax model proposed by Kwak et al. (2021, 2023) constructs the model by defining ResMax blocks, as shown in Fig. 7(a). Similar to the LCNN model, MFM activation is used instead of relu activation after the convolution layer, and residual connections are additionally considered. The overall model architecture is defined using seven ResMax blocks with 16, 16, 24, 32, 32, 48, 48, 64, and 64 filter size (f) parameters, as shown in Fig. 7(b).

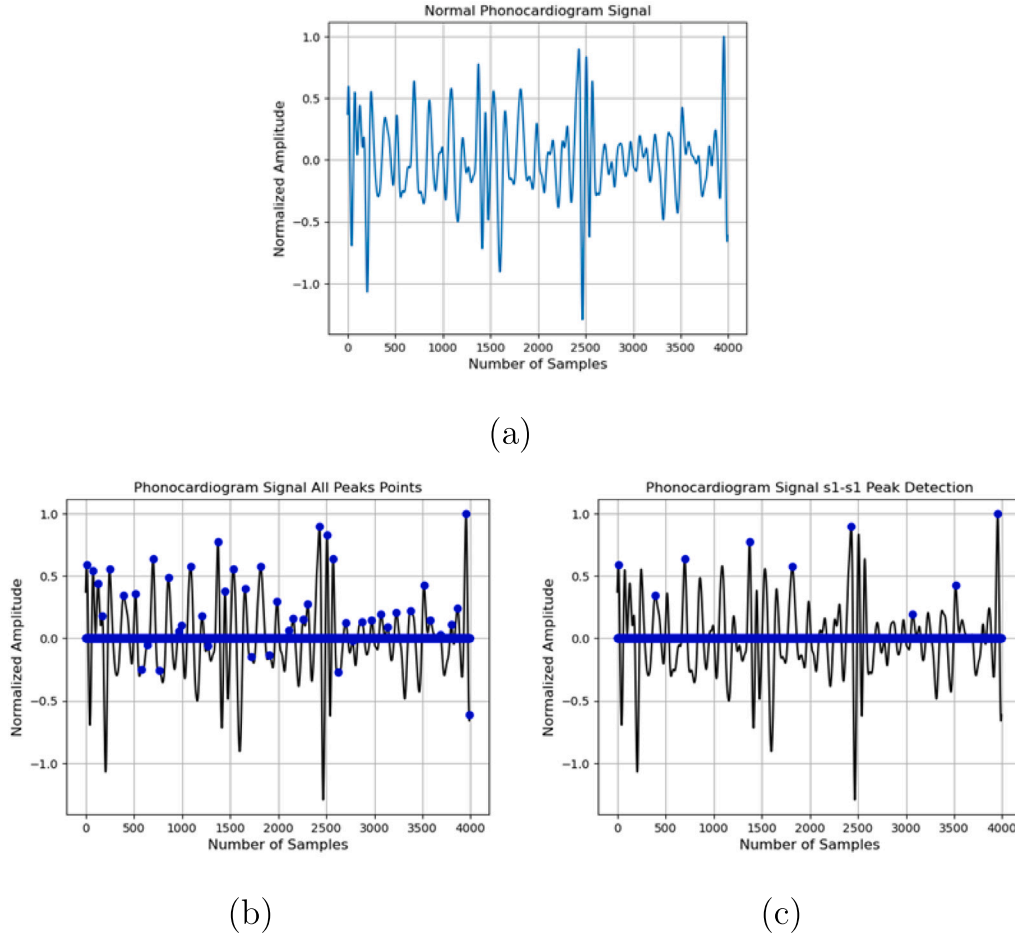


Fig. 4. Detection of peak points in PCG signals: (a) PCG signal, (b) peak detection capturing all peak points, and (c) Peak detection adjusted to identify only the peak points corresponding to S1-S1 intervals.

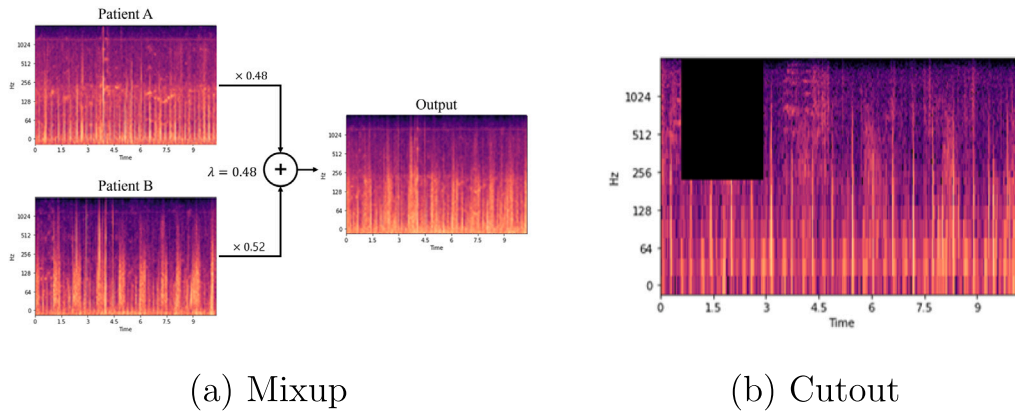


Fig. 5. Data augmentation techniques.

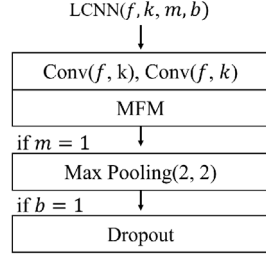
3.4.3. Simple embedding for demographic information

The demographic information included Age, Sex, Height, Weight, and Pregnancy status. Age had six categories (10s, 20s, ..., 60s+), we connected a Dense layer with an output size of 2. Height and Weight were combined as a 2-dimensional vector, and a Dense layer with an output size of 1 was connected. The Sex and Pregnancy status variables each had a Dense layer connected with an output size of 1. Then, all demographic information (outputs of Dense layers) was concatenated into an embedding vector of dimension 5. We refer to this as a Simple Embedding for demographic information features. Because

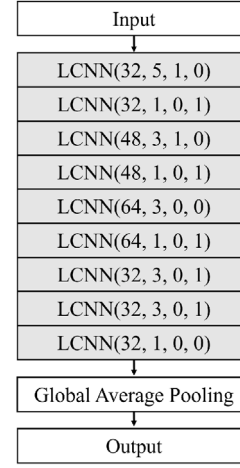
the data dimensionality was not high. By employing a straightforward embedding method, we aimed to enhance the accuracy of inference without the need for complex models. This straightforward embedding technique allows us to effectively incorporate these features into our models without adding significant complexity.

3.4.4. Dilated convolution

In this study, we expanded upon the architecture submitted to the George B. Moody PhysioNet Challenge 2022 by investigating the inclusion of the Peak Interval sequence within the model. While the original

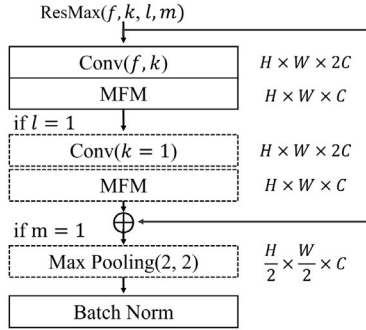


(a) LCNN block

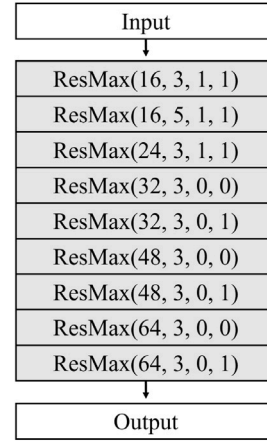


(b) LCNN model

Fig. 6. LCNN architecture.

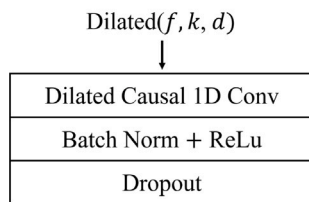


(a) ResMax block

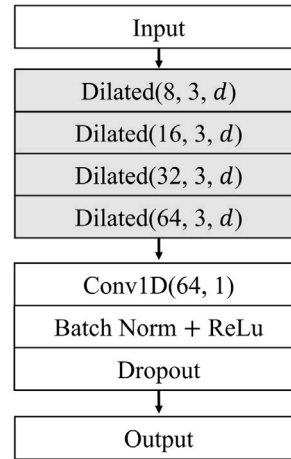


(b) ResMax model

Fig. 7. ResMax architectures.



(a) Dilated Conv block



(b) Dilated Conv model

Fig. 8. Dilated convolution architecture.

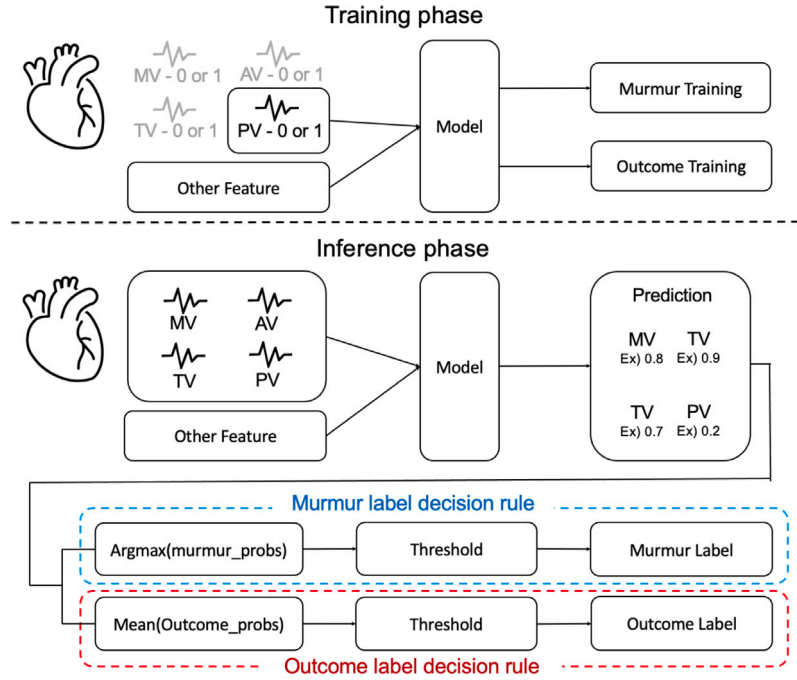


Fig. 9. Overview of training and inference processes, Including decision rules for murmur and outcome labels based on PCG data.

model considered the mean value of PI intervals, we posited that the incorporation of the PI interval sequence could provide additional information. This led us to propose the use of a Dilated Convolution (Dilated Conv) model (Yu et al., 2017).

In our SpectroHeart system, the Dilated Conv model is specifically utilized for the PI sequence feature. The block structure and the overall architecture of this model are illustrated in Fig. 8(a) and (b), respectively. Each block applies a Dilated Causal 1D convolution layer, followed by a Batch Normalization and a ReLU activation function, and then employs dropout. The overall model structure comprises four dilated conv blocks, each with filter sizes (f) of 8, 16, 32, and 64. All kernel sizes were set to 3, and the dilation factors (d) were set to various sizes. Following the passage through the four dilated conv blocks, a Conv 1D with a filter size of 64 was employed to construct the model.

The Dilated Conv distinguishes itself from traditional convolution by introducing a dilation factor that spaces out each element within the kernel. In this model, the PI sequence is utilized as a feature, and a Conv1D with an added dilation option is employed. The characteristic of dilation provides the possibility of using wide-range kernels with fewer resources by expanding the space between the original kernel components. This method allows for a broader coverage of input data without increasing the kernel size or computational complexity, proving particularly beneficial for effectively capturing information from data with long-term dependencies such as time series data. Owing to these attributes, Dilated Conv plays a crucial role in accurately understanding complex temporal features and enhancing model performance.

3.5. Training & inference

The distinction between the training and inference stages is depicted in Fig. 9. The provided data includes a maximum of four phonocardiograms and labels for tasks 1 and 2 per patient. While this scenario might necessitate a multiple instance learning approach, we opted to train the model by considering each cardiac signal as a single instance during the training stage, thereby maximizing the utilization of multiple phonocardiograms from a single patient and assigning the provided label to each of the patient's phonocardiograms. Although this methodology aligns with conventional supervised learning, it not only

augments the dataset within the PhysioNet Challenge but also enhances the model's learning across all four phonocardiograms.

During the inference phase where we predict the patient's heart sound and cardiac condition, we adopted a strategy of combining all the patient's multiple heart sounds as a multiple instance to make a decision about the patient's condition. We applied different decision rules for the murmur detection task and clinical outcome to make a final label determination. For murmur detection, we determined the murmur label by aggregating the output of the murmur classifier for each heart sound, taking the maximum value. For clinical outcome prediction, we aggregated the heart sounds using the mean value to make the final determination.

In the inference process of determining the patient's heart sounds and cardiac condition, we adopted a strategy of combining the results for all the patient's recorded auscultations to make a final decision about their condition. For the murmur detection and clinical outcome detection, we applied different decision rules to determine the final label. For murmur detection, we determined the murmur label by aggregating the output of the murmur classifier for each auscultation and taking the maximum value, while for clinical outcome prediction, we aggregated the results by taking the average value to make the final decision.

4. Experiments

4.1. Data set

The dataset used in this study was compiled from two large-scale screening campaigns conducted in northeastern Brazil in July–August 2014 and June–July 2015 (Oliveira et al., 2021). During the campaigns, auscultation was performed and audio samples were collected from four auscultation points by the same operator in a real clinical environment. The anatomical locations of the four heart valves are as follows: (1) Aortic Valve (AV), situated at the second intercostal space and right sternal border; (2) Pulmonic Valve (PV), positioned at the second intercostal space and left sternal border; (3) Tricuspid Valve (TV), found at the left lower sternum; and (4) Mitral Valve (MV), located at the fifth intercostal space and midclavicular line (apex of the heart). The PCG audio files generated from these auscultations were evaluated

Table 1

A confusion matrix for murmurs.

		Murmur expert		
		Present	Unknown	Absent
Murmur classifier	Present	m_{PP}	m_{PU}	m_{PA}
	Unknown	m_{UP}	m_{UU}	m_{UA}
	Absent	m_{AP}	m_{AU}	m_{AA}

for signal quality and subsequently segmented by a cardiac physiologist. Signal quality evaluation and segmentation were performed by different experts, while heart noise annotation was performed by the same annotator for the entire dataset.

60% of the dataset was publicly shared as the training set for the George B. Moody PhysioNet Challenge 2022. The remaining 40% of the data was held as validation and test sets to evaluate the model's performance in the PhysioNet Challenge 2022. The total number of patients of shared training set is 942, and the total number of auscultation samples is 3162.

The data provided by the competition consists of structured and unstructured data for each patient. For the structured data, it is demographic information and contains information on age, sex, height, weight, and pregnancy status. The unstructured data is recorded heart-beat sound data. There are a total of four stethoscope locations (PV, MV, TV, AV), and files exist for each stethoscope location. However, the location of the stethoscope for each patient is not the same. For example, assuming that there are patients A and B, if patient A has a total of four stethoscopes, PV, MV, TV, and AV, there are four samples. If patient B had only auscultation in PV and MV positions, there were two samples. Label is divided into three classes: Present, Absent, and Unknown for heart murmur detect, and outcome detect is divided into two classes: Normal and Abnormal.

4.2. Experimental setup

We implement SpectroHeart and all the baseline methods with Tensorflow. All the experiments are conducted on an Nvidia RTX A6000 and 8000 GPU with 48 GB memory.

For the Mel-spectrogram and STFT features, we set the window length to 512 and the hop length to 256. For the CQT features, we set the number of bins to 80 and the minimum frequency to 10. Manually written Python code was used to extract the STFT and Mel-spectrogram features, while the Python librosa package was used for CQT feature extraction. A sampling rate of 4000 was used for the recordings. For peak point detection, the peakutils module was employed with the threshold parameter set to 0.5/max(points) and the min_dist parameter set to 630. Detailed codes are publicly available from our github.

For the purpose of learning and evaluation, we partitioned the dataset consisting of 942 patients into a training set and a test set, with a ratio of 8:2. We emphasized the distribution of the 'Present' murmur class due to its critical role in the evaluation of both murmur and outcome. Consequently, we equally distributed the most audible locations between the training and test sets. The 'Unknown' and 'Absent' categories of murmur were randomly assigned at a ratio of 8:2. The similar class distribution of 'outcome' led us to use the same train/test split for both murmur and outcome evaluations. The training set contained 2532 samples from 751 patients, while the test set included 631 samples from 191 patients.

The competition evaluated the model based on two aspects: weighted accuracy and cost. Weighted accuracy gave more importance to correctly identifying murmurs, while the cost metric was assessed from the perspective of expenses incurred when actually using the model. The cost metric is calculated by estimating costs in a real clinical scenario (Reyna et al., 2023). Weighted accuracy is calculated with the Eq. (5) using Table 1.

$$S_{murmur} = \frac{5m_{PP}+3m_{UU}+m_{AA}}{5(m_{PP}+m_{UP}+m_{AP})+3(m_{PU}+m_{UU}+m_{AU})+(m_{PA}+m_{UA}+m_{AA})} \quad (5)$$

Table 2

Hyperparameter selection for optimal model exploration: Including 2D features, data augmentation, and more.

Hyperparameters	Contents
Spectrogram features	STFT, CQT, Mel-spectrogram
Data augmentation	Mixup, Cutout, CutMix, FFM, SpecAug
Cost-sensitive learning	1,3, or 5 on murmur or clinical outcome groups
Peak Interval (PI)	mean, not use
Inference	argmax, mean

The cost calculation assumes a clinical application scenario. Fig. 10 is a flowchart illustrating the application of the developed heart murmur detection model in a clinical setting. The proposed model is used in the algorithmic pre-screening stage, with the model's usage cost estimated at \$10. If the algorithm incorrectly identifies a case as normal when it is actually abnormal, the resulting cost is estimated at \$50,000. Additionally, if the algorithm identifies a case as abnormal, it undergoes a reevaluation by a human expert, which incurs additional costs. If the human expert also identifies the case as abnormal, the cost of further testing is \$10,000. When the model is applied to the evaluation data based on this flowchart, the total cost is calculated. Therefore, minimizing the cost metric equates to finding the optimal model for the clinical scenario depicted in the flowchart in Fig. 10. More detailed equations for calculating the cost metric are provided on the PhysioNet Challenge 2022 website.³

Each experiment was conducted five times, and the performance metrics of each model were derived by calculating the average. We conducted experiments for model development with training data and evaluated in terms of weighted accuracy and cost in test data for model advancement.

We employed cost-sensitive learning due to the varying significance of the murmur class and the outcome class in evaluation metrics (Elkan, 2001). The model was trained by merging the Unknown class with the Absent class, as the distribution of values for the Unknown class did not converge satisfactorily in murmur detection. During evaluation, the superior performance of weighted accuracy was observed when the Unknown class was not detected. Consequently, the system was configured to detect only the 'Absent' or 'Present' classes.

4.3. Random search for the challenge submission

To determine the hyperparameters in the LCNN and ResMax models, random search was employed, as detailed in Table 2. Our exploration encompassed various hyperparameters, including the selection of 2D features and data augmentation techniques. Regarding the 2D features, we investigated options such as STFT, CQT, and Mel-spectrogram among others. Regarding the data augmentation, we experimented with techniques like Mixup and Cutout, alongside others mentioned previously. Furthermore, careful consideration was given to the weight for cost-sensitive learning, which addresses class imbalance issues and is pivotal for achieving balanced performance across different classes.

Additionally, we assessed the impact of incorporating the PI mean feature into our model, deliberating on its relevance and contribution to overall performance. Lastly, deliberations were made regarding the inference type to be utilized during evaluation and prediction phases, ensuring the selection of an appropriate method or strategy to obtain predictions from our trained model, thereby ensuring efficient and accurate inference. By conducting 100 random runs of the model and allowing parameters to be chosen randomly, the best parameters were identified.

The initial competition model only took into account the simple PI mean feature, and the experimentation with the PI sequence feature was carried out in a later model. The results of the extended experiments on the PI sequence feature will be presented in the extended experiments on PI sequence feature section.

³ <https://moody-challenge.physionet.org/2022/>.

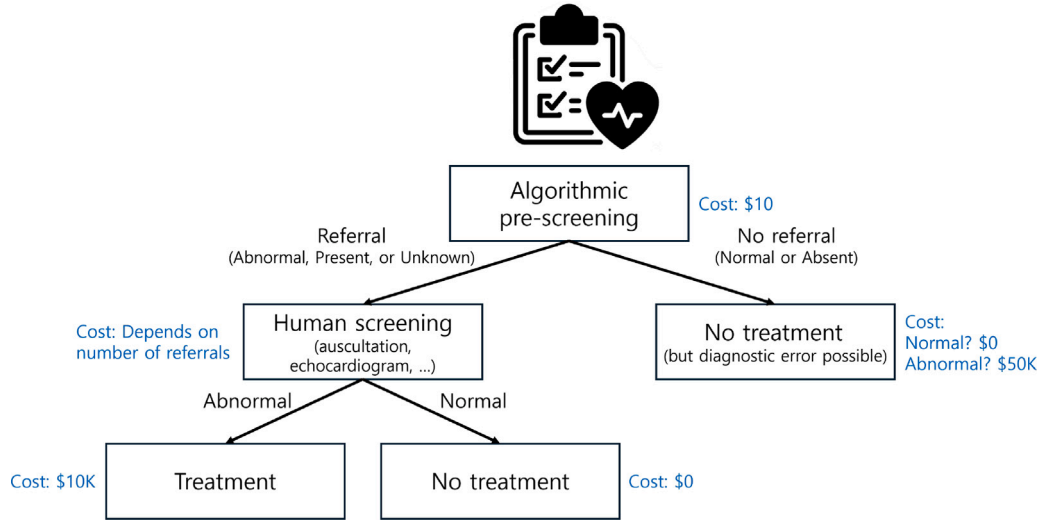


Fig. 10. Flowchart of decision making and cost outcomes associated with algorithmic prescreening.

Table 3

Test set results from the exploration of spectrogram hyperparameters ('-' indicates Mel-spectrogram).

Model	Added features	Weighted Acc (95% CI)	Cost (95% CI)
LCNN	-	0.798 (0.781, 0.816)	11,113 (10,608, 11,618)
LCNN	STFT	0.789 (0.764, 0.814)	11,600 (11,297, 11,903)
LCNN	CQT	0.703 (0.676, 0.730)	11,629 (11,382, 11,875)
ResMax	-	0.773 (0.734, 0.812)	11,354 (10,557, 12,151)
ResMax	STFT	0.769 (0.760, 0.779)	12,074 (11,450, 12,698)
ResMax	CQT	0.765 (0.732, 0.798)	11,523 (11,057, 11,989)

Abbreviation: Acc, Accuracy; CI, Confidence Interval; STFT, Short-Time Fourier Transform; CQT, Constant Q Transform.

Table 4

Evaluation on test set for the additional use of other variables.

Model	Added features	Weighted Acc (95% CI)	Cost (95% CI)
LCNN	-	0.798 (0.781, 0.816)	11,113 (10,608, 11,618)
LCNN	Demo	0.801 (0.788, 0.814)	11,208 (10,711, 11,704)
LCNN	PI mean	0.803 (0.789, 0.817)	11,295 (10,890, 11,700)
LCNN	Demo, PI mean	0.796 (0.780, 0.812)	11,081 (10,779, 11,384)
ResMax	-	0.773 (0.734, 0.812)	11,354 (10,557, 12,151)
ResMax	Demo	0.771 (0.727, 0.814)	12,812 (12,645, 12,980)
ResMax	PI mean	0.783 (0.755, 0.811)	12,519 (11,997, 13,041)
ResMax	Democ, PI mean	0.779 (0.764, 0.794)	11,528 (10,979, 12,076)

Abbreviation: Acc, Accuracy; CI, Confidence Interval; Demo, Demographic; PI, Peak Interval.

4.3.1. Additional use of STFT and CQT

Early trials showed that Mel-spectrogram feature is effective, and we experimented with performance changes while taking additional STFT and CQT features into account. In Table 3, both LCNN and ResMax models have similar or worse performance considering additional features. This led to the usage of solely Mel-spectrogram features.

The role of Mel filter banks is to represent the frequency scale with exponential increases, similar to how humans perceive sound. Therefore, the frequencies in the low range that humans can perceive well are represented in detail, while the other ranges are relatively less detailed. Since the heartbeat is also located in the low frequency range, the Mel-spectrogram has shown better performance than STFT and CQT in practice, making it a feature used in the final model.

4.3.2. Additional use of other variables

We evaluated the impact of demographic information (age, sex, height, weight, pregnancy, recording location) and PI mean on the

basic LCNN model. Regarding the 2D feature, we only used the Mel-spectrogram, as was also done in the results of Section 4.3.1. As seen in Table 4, the PI mean feature helped the LCNN model perform better. The model using both demographic information and PI mean features in the LCNN model demonstrated the best performance in terms of cost, whereas the model applying solely PI mean features to LCNN demonstrated the best performance in terms of weighted accuracy.

4.4. Extended experiments on PI sequence feature

We delve into the extended experiments conducted on the PI sequence feature, exploring the use of dilated convolution for processing time series features, such as the PI sequence, which can extract more information than solely relying on the average PI value.

A series of ablation experiments were performed to investigate the optimal hyperparameters for the Dilated Convolution model. The dilation factor d , which determines the extent of local dependency to be considered, was varied across four levels (1, 2, 4, and 8), resulting in a total of 14 different combinations in the Table 5.

In Table 5, the PI mean is succinctly denoted as PI-M, and the PI sequence as PI-S. This table presents the experimental results of models using only the PI sequence and those incorporating both the PI sequence and PI mean features. Despite variations in dilation combinations, both models exhibited similar performance, with the most effective model demonstrating a weighted accuracy of 0.816. However, in terms of the cost metric, the model using only the PI sequence feature achieved the highest score of 10,335.

Interestingly, the model that utilized only the PI-S feature exhibited the highest weighted accuracy of 0.816 when the dilation was set to a combination of (1,1,1,1). Consequently, with other values fixed, additional experiments were conducted on the Conv1D model after changing the padding to the default value. The experimental results of model #15 demonstrated a decrease in performance when dilation was not used. This suggests that dilated convolution effectively handles long-term dependencies in time-series data, accurately captures complex patterns within the data, and thus contributes to enhancing the performance of the model.

Table 6 presents the results of our proposed method, demonstrating the efficacy of PI features in both heart murmur detection and clinical outcome prediction tasks. The experiments were conducted on a total of six models, with Models 1 and 2 being the ones submitted to the competition. In the table, 'Demo' denotes 'Demography'. Notably, there is a significant performance improvement in models incorporating PI sequences compared to the competition-submitted models. Overall,

Table 5

Random search for dilation hyperparameter optimization ('-' indicates no dilated convolution network applied).

Model	Dilated	Weighted accuracy (95% CI)		Cost (95% CI)	
		PI-S	PI-S, PI-M	PI-S	PI-S, PI-M
#1	1111	0.816 (0.795, 0.836)	0.809 (0.794, 0.823)	10,574 (10,156, 10,992)	10,844 (10,507, 11,180)
#2	1112	0.812 (0.790, 0.835)	0.804 (0.782, 0.825)	10,640 (10,239, 11,041)	10,625 (10,284, 10,966)
#3	1122	0.808 (0.789, 0.828)	0.816 (0.802, 0.829)	10,669 (10,333, 11,005)	10,717 (10,291, 11,143)
#4	1222	0.799 (0.782, 0.817)	0.803 (0.774, 0.831)	10,564 (10,233, 10,895)	10,609 (10,275, 10,943)
#5	1224	0.813 (0.805, 0.821)	0.802 (0.787, 0.816)	10,580 (10,231, 10,928)	10,695 (10,137, 11,253)
#6	1244	0.801 (0.792, 0.809)	0.806 (0.791, 0.821)	10,496 (9,885, 11,106)	10,659 (10,167, 11,151)
#7	1248	0.806 (0.794, 0.819)	0.815 (0.785, 0.845)	10,648 (10,341, 10,955)	10,664 (10,282, 11,046)
#8	1444	0.813 (0.800, 0.826)	0.813 (0.795, 0.831)	10,466 (10,153, 10,778)	10,626 (10,369, 10,882)
#9	1144	0.798 (0.786, 0.810)	0.805 (0.788, 0.821)	10,439 (10,144, 10,734)	10,518 (10,123, 10,912)
#10	1114	0.810 (0.784, 0.835)	0.808 (0.796, 0.819)	10,335 (9,951, 10,719)	10,769 (10,410, 11,128)
#11	2222	0.808 (0.801, 0.815)	0.805 (0.784, 0.827)	10,613 (10,277, 10,950)	10,792 (10,512, 11,072)
#12	2244	0.811 (0.787, 0.834)	0.811 (0.788, 0.833)	10,510 (10,401, 10,618)	10,651 (10,024, 11,279)
#13	2444	0.799 (0.776, 0.823)	0.813 (0.804, 0.822)	10,689 (10,430, 10,948)	10,510 (10,056, 10,964)
#14	4444	0.804 (0.782, 0.827)	0.801 (0.785, 0.816)	10,744 (10,513, 10,974)	10,551 (10,008, 11,095)
#15	-	0.789 (0.769, 0.809)	0.790 (0.774, 0.805)	10,780 (10,515, 11,045)	10,766 (10,335, 11,197)

Abbreviation: CI, Confidence Interval; PI-S, Peak Interval sequence; PI-M, Peak Interval mean.

Table 6

Evaluation on test set for the extended use of peak interval.

Num	Extended features	Weighted accuracy	Cost	Accuracy	AUC
1 ^a	PI-M	0.803	-	0.859	0.768
2 ^a	PI-M, Demo	-	11,081	0.833	0.777
3	PI-S	0.816	-	0.822	0.772
4	PI-S, Demo	-	10,335	0.880	0.774
5	PI-S, PI-M	0.816	-	0.874	0.766
6	PI-S, PI-M, Demo	-	10,510	0.817	0.774

^a Challenge submission.**Table 7**

Comparison of average training time per epoch, average evaluation time per patient, and number of parameters for different models.

Num	Average training time per epoch (s)	Average evaluation time per patient (s)	# Params
1 ^a	4.840	3.055	141,460
2 ^a	4.718	3.092	141,522
3	7.245	3.597	736,224
4	12.825	3.618	545,248
5	8.237	3.620	736,224
6	13.504	3.672	545,248

^a Challenge submission.

these results suggest that additional information obtained through PI sequence plays a crucial role in enhancing the accuracy of heart disease diagnosis, and therefore can be an effective approach to improve model performance. The model training process was oriented towards maximizing weighted accuracy and minimizing cost, achieved through meticulous parameter tuning. Consequently, other evaluation metrics may exhibit relatively lower performance.

Table 7 compares the average training time per epoch, evaluation time per patient, and the number of parameters for each model. As observed from the table, the challenge submission models (Model 1 and Model 2) have relatively fewer parameters, resulting in shorter average training and evaluation times compared to other models. In contrast, the extended models incorporate sequence information from the Peak interval, leading to an increased number of parameters and consequently longer training and evaluation times.

Additionally, Models 1, 3, and 5 are designed to predict murmur labels, whereas Models 2, 4, and 6 are intended to predict outcomes. The differences in the characteristics and prediction goals of each model may contribute to variations in training and evaluation times.

5. Discussion

5.1. Comparison with other top performing methods from the PhysioNet challenge 2022

Table 8 presents the preprocessing methods, networks employed, and scores on the private test set for the top 10 teams in both the George B. Moody PhysioNet Challenge 2022 murmur detection task and clinical outcome prediction task, in which our research team participated. In contrast to previous competitions (Clifford et al., 2016) where preprocessing and feature extraction played significant roles, the rapid advancement of neural networks has led to a notable shift. As the speed of progress in neural networks has increased, the emphasis on preprocessing methods has diminished, focusing on a few key techniques, while the advancements in network architectures have become

more prominent. In the top-performing teams of the murmur detection task, it was common to achieve good results by employing frequency-time domain features based on Fourier transformation as the primary method of feature extraction (Lu et al., 2022; McDonald et al., 2022; Xu et al., 2022; Walker et al., 2022). On the other hand, in the clinical outcome prediction task, various methods such as 2D features, MFCC, and raw PCG were observed to be successful approaches among the top-performing teams (McDonald et al., 2022; Chang et al., 2022; Imran et al., 2022; Bruoth et al., 2022). These diverse techniques showcased their effectiveness in extracting informative features for predicting clinical outcomes.

In the competition, various network architectures were employed, including those from the CNN family, Recurrent Neural Network (RNN) family, Transformers, and traditional machine learning methods. Among these, the CNN-based approaches were most commonly used, and they demonstrated strong performance, particularly in the task of murmur detection (Lu et al., 2022; Xu et al., 2022; Walker et al., 2022). On the other hand, machine learning methods were found to be more effective in the clinical outcome prediction task (McDonald et al., 2022; Imran et al., 2022). One notable aspect of the competition was the deep exploration of labels by some teams, where they performed label correction to enhance the quality of the labels used (Xu et al., 2022). Additionally, there were several innovative network architectures observed, such as hierarchical networks and parallel network structures (Xu et al., 2022). Furthermore, while previous competitions often relied on K-fold cross-validation and resampling methods to make models more robust, the current trend leaned towards strengthening models through various data augmentation techniques (Lu et al., 2022; Kim et al., 2022; Ballas et al., 2022; Knorr and Bremer, 2022). This shift indicates a focus on improving model robustness through data augmentation rather than relying solely on traditional validation and resampling techniques.

Our team has experimented with a variety of 2D features such as STFT, CQT, and Mel-spectrogram. A unique distinction from other

Table 8
Top 10 Team of Task1,2 summarization (Murmur Detection/Clinical Outcome Prediction).

Team	Rank	Preprocessing	Network	Test score
HearHeart	1/10	Mel-spectrogram	CNN, RandomForest	0.78/12,110
CUED_Acoustics	2/1	Mel-spectrogram	RNN, HSMM, Catboost	0.77/11,144
HearTech+	2/9	Mel-spectrogram	Hierarchical multi-scale CNN	0.776/12,069
PathToMyHeart	4/17	Mel-spectrogram	Bayesian ResNet, XGBoost	0.771/12,637
CAU_UMN (ours)	5/5	Mel-spectrogram, PI	LCNN	0.767/11,933
SmartBeatIT	6/25	4 envelopes	Bi-LSTM	0.757/13,815
Care4MyHeart	6/29	Wavelet	Transformer	0.757/14,410
CeZIS	8/4	Raw PCG	ResNet50-1D	0.756/11,916
ISIBrno-AIMT	9/12	STFT	CNN-GRU	0.755/12,313
Murmur Mia!	9/28	MFCC	CNN, Gradient Boosting, Logistic	0.755/14,228
prna	20/2	MFCC	CNN	0.694/11,403
Melbourne_Kangas	29/3	MFCC	CNN-Bi-LSTM with Attention, RUSBoost, RobustBoost	0.632/11,735
HCCL	21/6	Mel-spectrogram	Vision Transformer	0.69/11,943
Listen2YourHeart	13/7	Raw PCG	CNN	0.737/11,946
uke-cardio	15/8	Mel-spectrogram	EfficientNet	0.735/11,990

Table 9
Weighted accuracy & Cost metric scores (official Challenge score) for our final selected entry (team CAU_UMN) for the murmur detection & clinical outcome task, including the ranking of our team on the hidden test set.

Task	Metric	Training	Validation	Test	Ranking
Murmur detection	Weighted accuracy	0.828	0.734	0.767	5/40
Clinical outcome prediction	Cost	8097	9493	11,933	5/40

teams lay in the utilization of the PI feature. We considered both PI sequence and PI average interval features, significantly advancing the performance of our model. We have introduced a distinctive element to the network architecture by utilizing the MFM activation function, setting our approach apart from other teams. Additionally, to create a robust model, we have explored various data augmentation techniques such as Mixup, Cutout, CutMix, FFM, and SpecAug. Ultimately, our submission have employed Mixup and Cutout as data augmentation methods to enhance the performance of model.

Table 9 summarizes our result. Weighted accuracy metric scores for our submitted models were 0.828, 0.734, and 0.767, respectively, on the training, validation, and test sets. Cost metric scores for our submitted models were 8097, 9493, and 11,933, respectively, on the training, validation, and test sets.

While other teams made attempts to capture the variability of heart rate through statistical measures on HRV, our team took a different approach by focusing on the PI feature. In the competition, we achieved the 5th position using the PI mean feature. However, in this paper, by extending the usage of PI, the sequence of PI itself was passed through a Dilated Conv structure and combined with the output of the existing LCNN. This resulted in a about 1.62%p improvement in the weighted accuracy metric compared to the submitted model on our team's created test set. The performance of the cost metric also improved significantly compared to the previous model. Although these results are not specific to the private test set of the competition, it is expected that the enhanced performance will carry over to the private test set as well.

5.2. Performance of LCNN and ResMax in other domain tasks

The LCNN and ResMax models used in this study are widely utilized deep learning models beyond the field of heart murmur detection. The LCNN, proposed by Wu et al. (2018), was initially introduced for face recognition tasks and has also been extensively used in voice spoofing detection (Lavrentyeva et al., 2017, 2019). The ResMax model also has demonstrated excellent performance in the spoofing detection domain (Kwak et al., 2023, 2022; Han et al., 2023b). Both LCNN and ResMax models performed well in voice spoofing detection problems. However, LCNN, initially proposed for face recognition and used in various fields, appears to perform better in more general scenarios compared to the ResMax model, which was specifically proposed for spoofing detection problems. This might explain why LCNN outperformed the ResMax model in our experiments.

5.3. Performance regarding data augmentation methods

When the evaluation data distribution differs significantly from the training data distribution, using data augmentation techniques can help train the model to be more robust, often resulting in better performance on the evaluation data. However, caution is needed as the performance may decline if the augmented data distribution deviates too much from the real data distribution.

In this competition, the evaluation data had a different distribution from the train and development sets, making it a more challenging task. Since the evaluation data was not publicly available, we could not include detailed experimental results for specific data augmentation techniques for the evaluation data set. Nonetheless, we found that applying techniques like cutout and Mixup significantly improved performance on the evaluation data, whereas techniques such as SpecAugment and FFM did not show substantial performance differences and were therefore not applied for our final models.

6. Conclusion

This paper delves into the models employed by our team in the George B. Moody PhysioNet Challenge 2022 and the extension of our proposed model. To distinguish between murmurs and outcomes from PCG, the 2022 competition utilized weighted accuracy and cost as evaluation metrics. We introduced an innovative deep learning model based on spectrograms, PI, and demographic features. Our LCNN model incorporating the PI feature achieved a weighted accuracy of 0.767 (ranking 5th out of 40 teams), while our LCNN model for outcome detection attained a cost of 11,933 (also ranking 5th out of 40 teams) during the official phase of the George B. Moody PhysioNet Challenge. Furthermore, we conducted experiments on the use of the PI feature, considering the PI sequence itself using dilated 1D convolution. Subsequent research revealed a 1.3% improvement in the performance of the murmur detection task, underscoring the efficacy of extending the PI Sequence feature.

The early diagnosis of cardiovascular diseases, including heart murmurs, is crucial for effective management. However, as outlined in the research background, diagnosing heart murmurs poses various challenges. In this study, our research team participated in the George B. Moody PhysioNet Challenge 2022 with the objective of addressing these challenges using end-to-end deep learning techniques. Through

additional experiments, we demonstrated the potential for advancement in this field.

Future research will focus on further refining our model and exploring frameworks that can be easily interpreted by clinicians. The potential applications of our research extend to the area of telemedicine. By integrating our model into wearable devices or smartphone applications, it is possible to monitor cardiac health effectively, as well as to provide continuous management of cardiovascular conditions in elderly patients, thereby facilitating preventive measures.

Our ultimate objective is to leverage the implemented artificial intelligence model to enhance accessibility to healthcare services, ensuring that no individual is marginalized. Through these efforts, we aim to promote public health and improve overall healthcare outcomes.

CRediT authorship contribution statement

Soyul Han: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Taein Kang:** Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Junggu Lee:** Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Narin Kim:** Software, Methodology, Investigation. **Hyejin Won:** Software, Investigation. **Yeong-Hwa Kim:** Writing – review & editing, Supervision, Investigation, Conceptualization. **Wuming Gong:** Writing – review & editing, Supervision, Methodology, Investigation, Funding acquisition, Conceptualization. **Il-Youp Kwak:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of competing 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.

Data availability

Data sets are available from the challenge website: <https://moody-challenge.physionet.org/2022/>.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science and ICT (RS-2023-00208284) and the Global-LAMP Program of the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education (No. RS-2023-00301976). During the preparation of this work, the authors used ChatGPT 3.5 in order to improve language and readability. After using this tool, the authors reviewed and edited the content as needed and took full responsibility for the content of the publication.

References

Ballas, A., Papapanagiotou, V., Delopoulos, A., Diou, C., 2022. Listen2YourHeart: A self-supervised approach for detecting murmur in heart-beat sounds. *arXiv preprint arXiv:2208.14845*.
 Bobillo, I.J.D., 2016. A tensor approach to heart sound classification. In: 2016 Computing in Cardiology Conference (CinC). IEEE, pp. 629–632.
 Brown, J.C., 1991. Calculation of a constant Q spectral transform. *J. Acoust. Soc. Am.* 89 (1), 425–434.
 Bruoth, E., Bugata, P., Gajdoš, D., Hudák, D., Kmečová, V., Staňková, M., Szabari, A., Vozáriková, G., et al., 2022. Murmur identification using supervised contrastive learning. In: 2022 Computing in Cardiology (CinC). Vol. 498, IEEE, pp. 1–4.
 Chang, Y., Liu, L., Antonescu, C., 2022. Multi-task prediction of murmur and outcome from heart sound recordings. In: 2022 Computing in Cardiology (CinC). Vol. 498, IEEE, pp. 1–4.

Clifford, G.D., Liu, C., Moody, B., Springer, D., Silva, I., Li, Q., Mark, R.G., 2016. Classification of normal/abnormal heart sound recordings: The PhysioNet/Computing in cardiology challenge 2016. In: 2016 Computing in Cardiology Conference (CinC). IEEE, pp. 609–612.
 DeVries, T., Taylor, G.W., 2017. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*.
 El-Segaier, M., Lilja, O., Lukkarinen, S., Sörnmo, L., Sepponen, R., Pesonen, E., 2005. Computer-based detection and analysis of heart sound and murmur. *Ann. Biomed. Eng.* 33 (7), 937–942.
 Elkan, C., 2001. The foundations of cost-sensitive learning. In: Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2. IJCAI '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 973–978.
 Fakhry, M., Gallardo-Antolín, A., 2024. Elastic net regularization and gabor dictionary for classification of heart sound signals using deep learning. *Eng. Appl. Artif. Intell.* 127, 107406.
 Faust, O., Shenfield, A., Kareem, M., San, T.R., Fujita, H., Acharya, U.R., 2018. Automated detection of atrial fibrillation using long short-term memory network with RR interval signals. *Comput. Biol. Med.* 102, 327–335.
 Han, S., Jeon, W., Gong, W., Kwak, I.-Y., 2023a. MCHear: multi-channel-based heart signal processing scheme for heart noise detection using deep learning. *Biology* 12 (10), 1291.
 Han, S., Kang, T., Choi, S., Seo, J., Chung, S., Lee, S., Oh, S., Kwak, I.-Y., 2023b. CAU KU deep fake detection system for ADD 2023 challenge. In: IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis (DADA 2023). pp. 23–30.
 Homsí, M.N., Medina, N., Hernandez, M., Quintero, N., Perpiñan, G., Quintana, A., Warrick, P., 2016. Automatic heart sound recording classification using a nested set of ensemble algorithms. In: 2016 Computing in Cardiology Conference (CinC). IEEE, pp. 817–820.
 Imran, Z., Grooby, E., Malgi, V.V., Sitaula, C., Aryal, S., Marzbanrad, F., 2022. A fusion of handcrafted feature-based and deep learning classifiers for heart murmur detection. In: 2022 Computing in Cardiology (CinC). Vol. 498, IEEE, pp. 1–4.
 Kay, E., Agarwal, A., 2016. Dropconnected neural network trained with diverse features for classifying heart sounds. In: 2016 Computing in Cardiology Conference (CinC). IEEE, pp. 617–620.
 Kim, J., Park, G., Suh, B., 2022. Classification of phonocardiogram recordings using vision transformer architecture. In: 2022 Computing in Cardiology (CinC). Vol. 498, IEEE, pp. 1–4.
 Knorr, M.S., Bremer, J.P., 2022. Using mel-spectrograms and 2D-CNNs to detect murmurs in variable length phonocardiograms. In: 2022 Computing in Cardiology (CinC). Vol. 498, IEEE, pp. 1–4.
 Kong, L., Barnova, K., Jaros, R., Mirjalili, S., Snasel, V., Pan, J.-S., Martinek, R., 2024. Analysis on fetal phonocardiography segmentation problem by hybridized classifier. *Eng. Appl. Artif. Intell.* 135, 108621.
 Kwak, I.-Y., Choi, S., Yang, J., Lee, Y., Han, S., Oh, S., 2022. Low-quality fake audio detection through frequency feature masking. In: Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia. DDAM '22, Association for Computing Machinery, New York, NY, USA, pp. 9–17.
 Kwak, I.-Y., Kwag, S., Lee, J., Huh, J.H., Lee, C.-H., Jeon, Y., Hwang, J., Yoon, J.W., 2021. ResMax: Detecting voice spoofing attacks with residual network and max feature map. In: 25th International Conference on Pattern Recognition. ICPR, IEEE Computer Society, Milan, pp. 4837–4844.
 Kwak, I.-Y., Kwag, S., Lee, J., Jeon, Y., Hwang, J., Choi, H.-J., Yang, J.-H., Han, S.-Y., Huh, J.H., Lee, C.-H., Yoon, J.W., 2023. Voice spoofing detection through residual network, max feature map, and depthwise separable convolution. *IEEE Access* 11, 49140–49152.
 Lavrentyeva, G., Novoselov, S., Malykh, E., Kozlov, A., Kudashev, O., Shchemelinin, V., 2017. Audio replay attack detection with deep learning frameworks. In: Proc. Interspeech 2017. ISCA, Stockholm, pp. 82–86.
 Lavrentyeva, G., Novoselov, S., Tseren, A., Volkova, M., Gorlanov, A., Kozlov, A., 2019. STC antispoofing systems for the asvspoof2019 challenge. In: Proc. Interspeech 2019. ISCA, Graz, pp. 1033–1037.
 Lee, J., Kang, T., Kim, N., Han, S., Won, H., Gong, W., Kwak, I.-Y., 2022. Deep learning based heart murmur detection using frequency-time domain features of heartbeat sounds. In: 2022 Computing in Cardiology (CinC). Vol. 498, pp. 1–4.
 Liu, C., Springer, D., Li, Q., Moody, B., Juan, R.A., Chorro, F.J., Castells, F., Roig, J.M., Silva, I., Johnson, A.E., et al., 2016. An open access database for the evaluation of heart sound algorithms. *Physiol. Meas.* 37 (12), 2181.
 Lu, H., Yip, J.B., Steigleder, T., Griefhammer, S., Heckel, M., Jami, N.V.S.J., Eskofier, B., Ostgathe, C., Koelpin, A., 2022. A lightweight robust approach for automatic heart murmurs and clinical outcomes classification from phonocardiogram recordings. In: 2022 Computing in Cardiology (CinC). Vol. 498, IEEE, pp. 1–4.
 McDonald, A., Gales, M.J., Agarwal, A., 2022. Detection of heart murmurs in phonocardiograms with parallel hidden semi-Markov models. In: 2022 Computing in Cardiology (CinC). Vol. 498, IEEE, pp. 1–4.
 Oliveira, J., Renna, F., Costa, P.D., Nogueira, M., Oliveira, C., Ferreira, C., Jorge, A., Mattos, S., Hatem, T., Tavares, T., Elola, A., Rad, A.B., Sameni, R., Clifford, G.D., Coimbra, M.T., 2021. The CirCor DigiScope dataset: From murmur detection to murmur classification. *IEEE J. Biomed. Health Inform.* 26 (6), 2524–2535.

- Potes, C., Parvaneh, S., Rahman, A., Conroy, B., 2016. Ensemble of feature-based and deep learning-based classifiers for detection of abnormal heart sounds. In: 2016 Computing in Cardiology Conference (CinC). IEEE, pp. 621–624.
- Reyna, M.A., Kiarashi, Y., Elola, A., Oliveira, J., Renna, F., Gu, A., Perez Alday, E.A., Sadr, N., Sharma, A., Kpodonu, J., et al., 2023. Heart murmur detection from phonocardiogram recordings: The george b. moody physionet challenge 2022. *PLOS Digit. Health* 2 (9), e0000324.
- Reyna, M.A., Kiarashi, Y., Elola, A., Oliveira, J., Renna, F., Gu, A., Perez-Alday, E.A., Sadr, N., Sharma, A., Mattos, S., Coimbra, M.T., Sameni, R., Rad, A.B., Clifford, G.D., 2022. Heart murmur detection from phonocardiogram recordings: The George B. Moody PhysioNet challenge 2022. <http://dx.doi.org/10.1101/2022.08.11.22278688>, medRxiv.
- Singh, J., Anand, R.S., 2007. Computer aided analysis of phonocardiogram. *J. Med. Eng. Technol.* 31 (5), 319–323.
- Tsipouras, M.G., Fotiadis, D.I., Sideris, D., 2005. An arrhythmia classification system based on the RR-interval signal. *Artif. Intell. Med.* 33 (3), 237–250.
- Walker, B., Krones, F., Kiskin, I., Parsons, G., Lyons, T., Mahdi, A., 2022. Dual Bayesian ResNet: A deep learning approach to heart murmur detection. *Comput. Cardiol.*
- Wen, Y.-N., Lee, A.P.-W., Fang, F., Jin, C.-N., Yu, C.-M., 2014. Beyond auscultation: acoustic cardiography in clinical practice. *Int. J. Cardiol.* 172 (3), 548–560.
- Wu, X., He, R., Sun, Z., Tan, T., 2018. A light CNN for deep face representation with noisy labels. *IEEE Trans. Inf. Forensics Secur.* 13 (11), 2884–2896.
- Xu, Y., Bao, X., Lam, H.-K., Kamavuako, E.N., 2022. Hierarchical multi-scale convolutional network for murmurs detection on pcg signals. In: 2022 Computing in Cardiology (CinC). Vol. 498, IEEE, pp. 1–4.
- Yu, F., Koltun, V., Funkhouser, T., 2017. Dilated residual networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 472–480.
- Zabihi, M., Rad, A.B., Kiranyaz, S., Gabbouj, M., Katsaggelos, A.K., 2016. Heart sound anomaly and quality detection using ensemble of neural networks without segmentation. In: 2016 Computing in Cardiology Conference (CinC). IEEE, pp. 613–616.
- Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D., 2017. Mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.