applied sciences

MDPI

Article

Heart Murmur Quality Detection Using Deep Neural Networks with Attention Mechanism

Tingwei Wu [1], Zhaohan Huang [1], Shilong Li [1], Qijun Zhao [2] and Fan Pan [1,*]

[1] College of Electronics and Information Engineering, Sichuan University, Chengdu 610065, China; 2021141450075@stu.scu.edu.cn (T.W.); 2022222050110@stu.scu.edu.cn (Z.H.); lishilong@stu.scu.edu.cn (S.L.)

[2] College of Computer Science, Sichuan University, Chengdu 610065, China; qjzhao@scu.edu.cn

* Correspondence: panfan@scu.edu.cn

Abstract: Heart murmurs play a critical role in assessing the condition of the heart. Murmur quality reflects the subjective human perception of heart murmurs and is an important characteristic strongly linked to cardiovascular diseases (CVDs). This study aims to use deep neural networks to classify the patients' murmur quality (i.e., harsh and blowing) from phonocardiogram (PCG) signals. The phonocardiogram recordings with murmurs used for this task are from the CirCor DigiScope Phonocardiogram dataset, which provides the murmur quality labels. The recordings were segmented, and a dataset of 1266 segments with average lengths of 4.1 s from 164 patients' recordings was obtained. Each patient usually has multiple segments. A deep neural network model based on convolutional neural networks (CNNs) with channel attention and gated recurrent unit (GRU) networks was first used to extract features from the log-Mel spectrograms of segments. Then, the features of different segments from one patient were weighted by the proposed "Feature Attention" module based on the attention mechanism. The "Feature Attention" module contains a layer of global pooling and two fully connected layers. Through it, the different features can learn their weight, which can help the deep learning model distinguish the importance of different features of one patient. Finally, the detection results were produced. The cross-entropy loss function was used to train the model, and five-fold cross-validation was employed to evaluate the performance of the proposed methods. The accuracy of detecting the quality of patients' murmurs is 73.6%. The F1-scores (precision and recall) for the murmurs of harsh and blowing are 76.8% (73.0%, 83.0%) and 67.8% (76.0%, 63.3%), respectively. The proposed methods have been thoroughly evaluated and have the potential to assist physicians with the diagnosis of cardiovascular diseases as well as explore the relationship between murmur quality and cardiovascular diseases in depth.

Keywords: cardiovascular diseases (CVDs); computer-aided diagnosis; heart murmur quality detection; phonocardiogram (PCG); deep learning; attention mechanism

check for updates

Citation: Wu, T.; Huang, Z.; Li, S.; Zhao, Q.; Pan, F. Heart Murmur Quality Detection Using Deep Neural Networks with Attention Mechanism. *Appl. Sci.* **2024**, *14*, 6825. https://doi.org/10.3390/app14156825

Academic Editor: Tobias Meisen

Received: 11 June 2024
Revised: 16 July 2024
Accepted: 1 August 2024
Published: 5 August 2024

1. Introduction

Cardiovascular diseases (CVDs) are one of the world's most serious diseases, killing more people each year than any other cause of death. According to World Health Organization estimates, 17.9 million people died from CVDs in 2019, representing 32% of all global deaths [1]. Most deaths caused by CVDs happen in low- and middle-income countries, where the majority of the population lacks access to an integrated primary healthcare system. Thus, the diagnosis and treatment of CVDs may be delayed, increasing the risk of early deaths [2].

Phonocardiograms (PCGs) are heart sound signals produced by the mechanical activity of the heart, containing information related to the heart condition [3]. Cardiac auscultation can provide insights into the PCGs and determine at a low cost whether more expensive testing should be ordered [4], thereby reducing deaths due to CVDs. Despite its benefits, cardiac auscultation is a difficult skill to acquire, requiring extensive training and clinical experience [2], which limits its popularization in most low- and middle-income countries that lack cardiologists [5]. Computer-aided auscultation offers a solution by significantly decreasing the cost related to cardiac auscultation, efficiently addressing this difficulty. Computer-aided heart sound classification and detection is a crucial part of computer-aided auscultation. Most studies in this field now focus on determining the presence or absence of heart murmurs and the normality or abnormality of heart sounds. An overview of the methods used in these studies can be found in [6–8]. Notably, except for [5,9], there has been limited research using deep learning methods to explore more detailed PCG signal characteristics such as the grading, pitch, shape, timing, and quality of heart sounds, though they are all important for assessing the condition of the heart.

Heart murmurs include systolic and diastolic murmurs, which occur during the systole period and the diastolic period, respectively. According to the classification of Leathem, systolic heart murmurs are majorly divided into the systolic ejection murmur and the systolic regurgitant murmur. The "harsh" quality is most commonly found in the systolic ejection murmur, and the presence of this type of murmur is usually associated with innocent systolic flow murmurs and valvular or vascular obstruction. The qualities of "blowing" and "harsh" are all likely to be found in the systolic regurgitant murmur, and the presence of these murmurs is associated with mitral or tricuspid regurgitation and ventricular septal defect [10]. So, the use of computer-aided detection of murmur quality would provide richer and more accurate information related to the CVDs for computer-aided auscultation. To the best of our knowledge, no studies have provided specific deep learning models for the specific task of murmur quality detection according to clinical criteria, and the related works can be found in Section 2 below.

The determination of murmur quality relies on human subjective judgment, which comes from the experience of cardiologists through extensive training, rather than from a recognized gold standard, leading to difficulties in computer detection of murmur quality. Deep learning is an effective method to address this problem because it can mimic highly skilled cardiologists, who possess considerable knowledge via ample training.

This work aims to detect the murmur quality (i.e., harsh and blowing) for patients and find the contribution of other murmur characteristics in this task using deep neural networks, which can help in the diagnosis of cardiac and the establishment of criteria for murmur quality detection. Specifically, the study makes the following contributions:

- Designing a deep learning model to extract features from the log-Mel spectrograms of PCG segments and proposing a new module to weight the features extracted from one patient for murmur quality detection;
- Thoroughly evaluating the advantages and inadequacies of deep learning methods in murmur quality classification;
- Exploring the relationship between murmur quality and other murmur characteristics by using deep learning models.

2. Related Works

There have been numerous algorithms and models developed to classify and detect heart sounds so far. The majority of them focus on traditional binary classification, which involves determining whether a murmur present or not and whether a heart sound is abnormal. For example, in the George B. Moody PhysioNet Challenge 2022 [11], the participating teams mainly strived to tackle the above two issues, as in [12–14]. In addition, a small portion of the work focuses on more novel classification problems, such as the diagnosis of cardiac diseases and murmur grading. The authors of [15–17] investigated the detection of cardiac valve disorders, whereas [5,9] graded murmurs. To the best of our knowledge, there are no deep learning models built for this specific task for the time being.

For the research methodology, part of the current research used machine learning algorithms to categorize the manually extracted features [18–20]. For example, [19] extracted two different features from heart sounds, discrete wavelet transform (DWT) and Mel-frequency cepstral coefficients (MFCCs), and they used three machine learning classifiers to categorize them. An important trend nowadays is to use deep learning algorithms for the study. Part of the deep learning algorithms directly uses deep neural networks to extract features from the waveforms of the PCGs (one-dimensional signals) and classify them, e.g., [21,22], while the other part pre-extracts the time–frequency feature maps (usually two-dimensional feature maps) and then uses the deep learning models to further extract the high-dimensional features and categorize them, e.g., [23–25]. Furthermore, Ref. [26] used a combination of deep learning models and traditional machine learning algorithms. All of the above studies address supervised learning. To improve model performance, supervised learning requires a huge amount of labeled training data, which increases the research costs. Self-supervised and unsupervised models can address this problem better, and some studies [27–29] have established unsupervised and self-supervised models for murmur detection and heart sound classification with good results. For example, Ref. [29] pretrained a wav2vec 2.0 model on the Circor DigiScope Phonocardiogram dataset [2,30]; then, they fine-tuned it on small-scale annotated data, and the model showed strong competitiveness and robustness.

In the field of computer-aided diagnosis of cardiovascular diseases, in addition to the use of PCGs as the basis for diagnosis, there are many studies that use an electrocardiogram (ECG) (e.g., [31–33]) as the main basis, and [34] conducted computer-aided diagnosis of heart disease by the method of feature selection. It is possible that combining PCGs with other cardiac parameters, such as ECGs, would provide a boost to the development of computer-aided diagnosis in the future.

3. Dataset

The dataset used in this study is the publicly available set of the George B. Moody PhysioNet Challenge 2022, namely the CirCor DigiScope Phonocardiogram dataset [2,30]. It was collected in northeast Brazil over the months of July–August 2014 and June–July 2015 [2] and contains 3163 PCG recordings from 942 patients at a sampling rate of 4000 Hz. The majority of patients have multiple PCG recordings, with most of them from the four auscultation locations: the aortic valve (AV), pulmonary valve (PV), tricuspid valve (TV), and mitral valve (MV), and a few are from other auscultation locations. As shown in Table 1, this dataset is balanced between males and females and was collected primarily from children and infants. The average lengths (±standard deviation) of the PCG recordings are 22.9 (±7.3) s with the shortest and longest lengths of 5.2 s and 64.5 s, respectively.

The dataset provides segmentation labels (S1, systolic period, S2, and diastolic period) for each PCG recording and murmur locations for each patient. Significantly, it is labeled with many characteristics of murmurs including the most audible location, timing, shape, pitch, grading, and quality manually by a cardiac physiologist [30]. Specifically, the murmur grading is described as "I/VI", "II/VI", and "III/VI" based on Levine's scale [35]; the murmur pitch is described as "High", "Medium", and "Low"; the murmur timing is described as "Early-", "Mid-", and "Late-" systolic; and the murmur shape is described as "Crescendo", "Decrescendo", "Diamond", and "Plateau" [2].

The murmur quality is labeled as "Blowing", "Harsh", and "Musical", and all auscultation locations with murmurs in a single patient correspond to only one murmur quality. A "Harsh" murmur is described by a high-velocity blood flow from a higher to a lower pressure gradient, and a "Blowing" murmur is a sound caused by turbulent (rough) blood flow through the heart valves [2]. The murmur of aortic stenosis tends to be a harsh, grating murmur, whereas that of mitral regurgitation has a gentle, blowing quality [36]. However, the "Musical" quality is extremely rare, which may be related to some innocent murmurs [2]. Table 1 demonstrates that most murmurs are present in the systole period with very few in the diastolic period (2.8%); the majority of systolic murmur quality is determined as "Blowing" and "Harsh" with "Musical" murmurs being relatively rare (2.2%). Therefore, heart sounds with a systolic murmur quality described as "Musical" were excluded, and the study focused only on two types of heart sounds, i.e., systolic murmur quality described as "Blowing" and "Harsh". Also, not all heart sounds in all auscultation locations had murmurs for a patient diagnosed with murmurs, and the heart sounds in those auscultation locations where murmurs were present were studied only. After excluding the above, an analysis dataset of 174 patients with 604 recordings was included for analysis in the research. The details of the analysis dataset are displayed in Table 1. The average lengths (±standard deviation) of the PCG recordings in the analysis dataset are 22.2 (±7.9) s with the shortest and longest lengths of 6.4 s and 64.5 s, respectively.

Table 1. The percentage of patients in gender, age, and murmur quality (systolic and diastolic period) in the original dataset and analysis dataset.

	Original Dataset			**Analysis Dataset * (%)**
	Present (%)	**Absent and Unknown (%)**	**Total (%)**	
Gender				
Male	87 (48.6)	369 (48.4)	456 (48.4)	**84 (48.3)**
Female	92 (51.4)	394 (51.6)	486 (51.6)	**90 (51.7)**
Total	179 (100.0)	763 (100.0)	942 (100.0)	**174 (100.0)**
Age				
Adolescent	16 (8.9)	56 (7.3)	72 (7.6)	**16 (9.2)**
Child	132 (73.7)	532 (69.7)	664 (70.5)	**127 (73.0)**
Infant	25 (14.0)	101 (13.2)	126 (13.4)	**25 (14.4)**
Neonate	1 (0.6)	5 (0.7)	6 (0.6)	**1 (0.6)**
Nan	5 (2.8)	69 (9.0)	74 (7.9)	**5 (2.9)**
Total	179 (100.0)	763 (100.0)	942 (100.0)	**174 (100.0)**
Murmur quality (systolic)				
Blowing	78 (43.6)	-	78 (8.3)	**78 (44.8)**
Harsh	96 (53.6)	-	96 (10.2)	**96 (55.2)**
Musical	4 (2.2)	-	4 (0.4)	**-**
Nan	1 (0.6)	763 (100.0)	764 (81.1)	**-**
Total	179 (100.0)	763 (100.0)	942 (100.0)	**174 (100.0)**
Murmur quality (diastolic)				
Blowing	4 (2.2)	-	4 (0.4)	-
Harsh	1 (0.6)	-	1 (1.1)	-
Musical	0 (0.0)	-	0 (0.0)	-
Nan	174 (97.2)	763 (100.0)	937 (99.6)	-
Total	179 (100.0)	763 (100.0)	942 (100.0)	-

* The obtained dataset after the processing of excluding, which is the dataset used in the following methods.

According to the segmentation labels, each recording was cut into segments that contain about seven cardiac cycles to have equivalent information of heart sounds. The segments have an average length (±standard deviation) of 4.1 (±0.7) s and no overlap between them. Some patients' PCGs could not be used due to the missing of their segmentation labels. The percentage of patients in murmur quality, timing, shape, grading, and pitch of analysis dataset after segmentation is given in Table 2. The number of segments for segments' murmur quality detection is 1266, with 747 labeled as "Harsh" and 519 labeled as "Blowing", while the number of patients for patients' murmur quality detection is 164 with 90 labeled as "Harsh" and 74 labeled as "Blowing".

Table 2. The percentage of patients in murmur quality, timing, shape, grading, and pitch.

	Analysis Dataset after Segmentation (%)
Murmur quality	
Blowing	74 (45.1)
Harsh	90 (54.9)
Murmur timing	
Early-systolic	54 (32.9)

Table 2. *Cont.*

Analysis Dataset after Segmentation (%)	
Holosystolic	94 (57.3)
Mid-systolic	15 (9.1)
Late-systolic	1 (0.6)
Murmur shape	
Crescendo	2 (1.2)
Decrescendo	32 (19.5)
Diamond	28 (17.1)
Plateau	102 (62.2)
Murmur grading	
I	91 (55.5)
II	28 (17.1)
III	45 (27.4)
Murmur pitch	
High	39 (23.8)
Low	82 (0.5)
Medium	43 (26.2)
Total	164 (100.0)

4. Method

The segments' detection (the segment level) would be an intermediate step for model evaluation; thus, the study wants to detect the murmur qualities for patients (the patient level). As shown in Figure 1, the steps of murmur quality detection include data segmentation, log-Mel spectrogram feature extraction, deep neural network feature extraction and detection. The analysis dataset used for the proposed method and the analysis dataset after segmentation are defined in Section 3. The details are given below in this section.

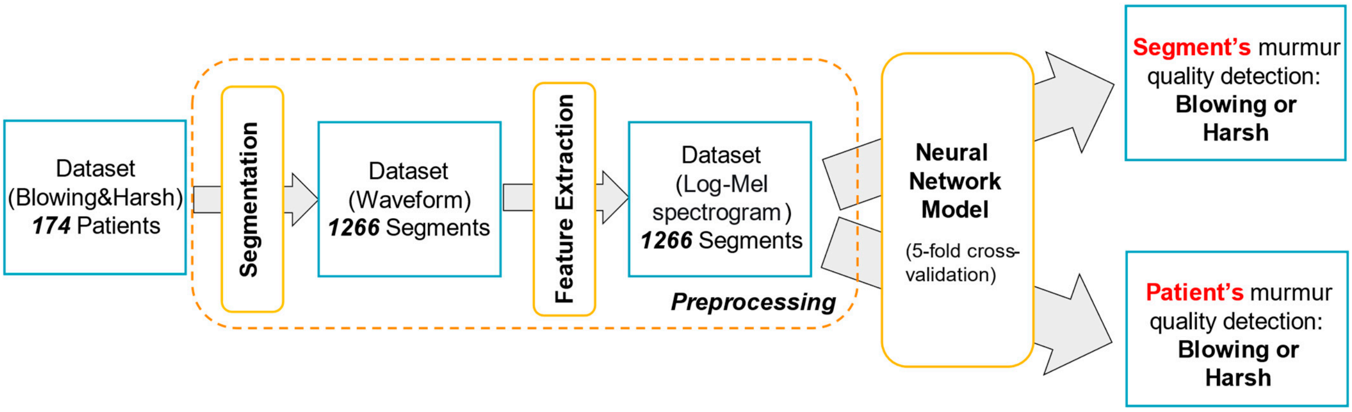

Figure 1. Overview of the methodology to detect murmur quality using deep neural network model. The dataset used is the analysis dataset defined in Section 3.

4.1. Feature Extraction

A 2D log-Mel spectrogram representation [37] was extracted for each PCG segment. In deep learning, this representation is widely used in the preprocessing of acoustic signals, which analyzes the spectrum of sound based on the mechanism of human hearing. Using them as inputs to the neural network model can more successfully characterize the sound compared to waveforms. For the extraction of log-Mel spectrograms, a frame length of 25 ms, a frame shift of 10 ms, and a window type of "Povery" were chosen, and heart sounds were analyzed in the frequency range of 0 to 2000 Hz using 128 Mel filters. Figure 2 shows the waveforms and log-Mel spectrograms of two typical heart sound segments (two with systolic murmurs described as "Harsh" and "Blowing"). Since the average length of the segments is 4.1 s, the length of the log-Mel spectrogram was determined to be 400 values, and the log-Mel spectrograms were cut (zero-padded) for those longer (shorter) than 400 values. This preparation is necessary before inputting it into the deep neural network. Finally, a 128 × 400 log-Mel spectrogram representation matrix was extracted for each segment. The log-Mel spectrograms were normalized before being used as the following method:

$$X = \frac{X - \text{mean\_value}}{2 * \text{std\_value}} \tag{1}$$

where X denotes each value of the log-Mel spectrogram, mean_value denotes the mean value of the log-Mel spectrogram, and std_value denotes the standard deviation of the log-Mel spectrogram.

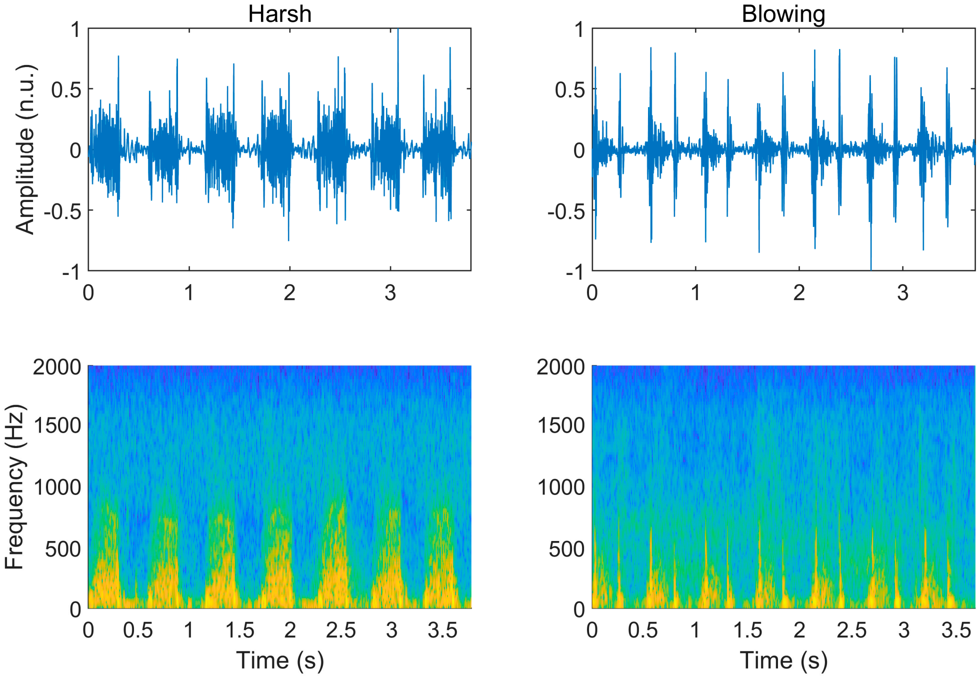

Figure 2. The waveforms and log-Mel spectrograms of two typical heart sound segments (two with systolic murmurs described as "Harsh" and "Blowing"). n.u. refers to normalized units.

4.2. Neural Network Model for Segments' Detection

The designed neural network model is based on two-dimensional convolutional neural networks (2D-CNNs), channel attention, and bidirectional gated recurrent units (Bi-GRU). The CNN-Block was used for initial image feature extraction from the log-Mel spectrograms. Subsequently, the feature maps were flattened into a long feature sequence to be fed into the Bi-GRU for temporal feature extraction. In particular, a Squeeze-and-Excitation (SE)-Block was added between the CNN-Block and the Bi-GRU for giving different weights to different channels of the feature map, since the information in different channels may have different importance for the detection of the murmur quality (the results verified this). Figure 3 shows the structure of the designed neural network model, which included the following:

- CNN-Block: It contains three layers of 2D convolution; each layer of 2D convolution follows an activation function (ReLU) and a batch-normalization (BN) layer. The first layer of 2D convolution has a convolution kernel size of 5 × 5, a stride of 2, and a padding of 2; the second and third layers of 2D convolution both have a convolution kernel size of 3 × 3, a stride of 1, and a padding of 1. Bias is added to all convolutions, and the padding is zero padding. Through the CNN-Block, the 1-channel log-Mel spectrogram became a 32-channel feature map.
- SE-Block: This kind of block was proposed by Hu et al. [38]. It can model the interdependencies between channels by the information of different channels of the feature map, that is, to generate corresponding weights for each channel of the feature map, so as to recalibrate the features by the importance of different channels. Now, the

structure of the added SE-Block was explained. Firstly, there is global average pooling, which can change the feature map from a C × H × W matrix to a C × 1 × 1 matrix, which is called squeeze. Then, there are two fully connected (FC) layers (between which is the ReLU activation function) with the output sizes of C/2 and C and the Sigmoid activation function, which is called excitation. Finally, a C × 1 × 1 matrix was obtained, which means that each channel received a weight, and then the different channels were weighted by doing a channel-wise multiplication with the input feature map (Scale).

- Bi-GRU: It is a bidirectional GRU module for extracting features from long sequences. The size of the hidden state in this module is 64, and the module is added bias.
- Linear Prediction Head: It contains an FC layer and a Sigmoid activation function that produces the final predicted labels for each segment.

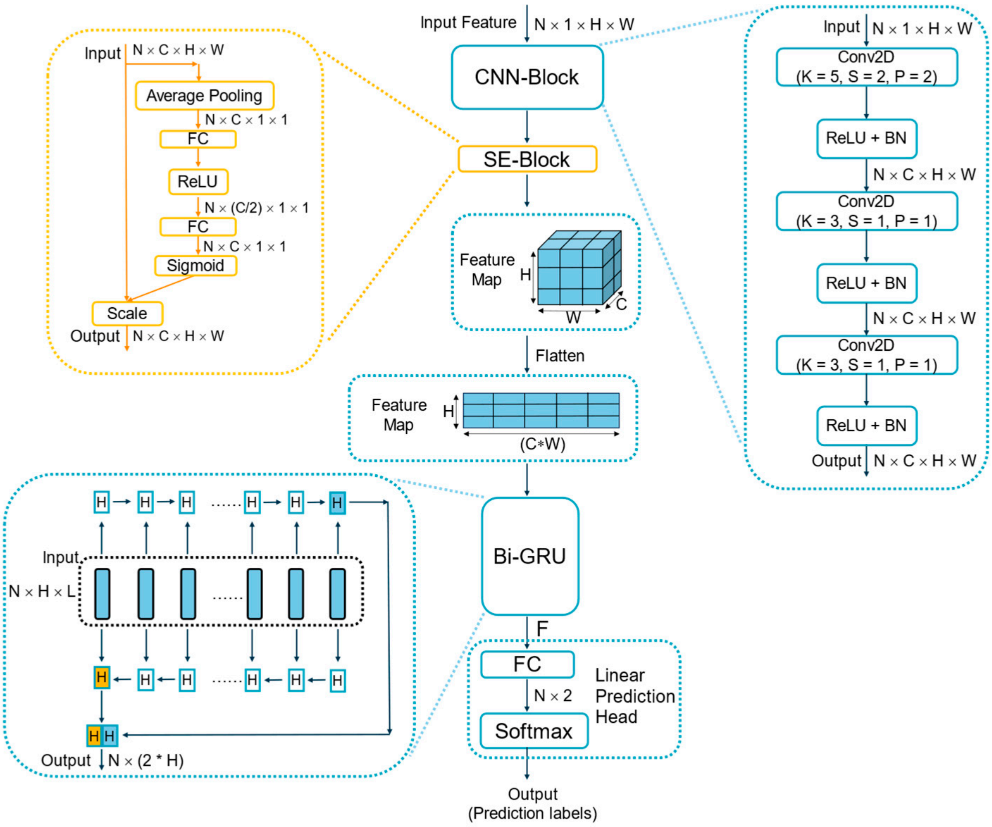

Figure 3. Structure of the proposed neural network model. N corresponds to the batch size. K, S, and P correspond to the kernel size, stride, and padding, respectively, in the CNN-Block. C, H, and W correspond to the channel, height, and width of the input feature maps, respectively, in the CNN-Block and SE-Block. H and L in Bi-GRU correspond to the feature size and sequence length of the input feature sequences, respectively.

When training the neural network, the Adam optimizer was used, the learning rate was set to $3 * 10^{-5}$, $\beta_1 = 0.9$, $\beta_2 = 0.98$, and the weight decay was 0.01. β_1 refers to the exponential decay rate of the first moment estimation, and β_2 refers to the exponential decay rate of the second moment estimation. The weight decay setting allowed for the L2 regularization to be added to the loss function, thus reducing overfitting to some extent. The loss function used was the cross-entropy loss function. The batch size was set to 64.

4.3. Method of Feature Weighting for Patients' Detection

The method of feature weighting is inspired by the SE-Block and based on the attention mechanism, which is called Feature Weighting. In this method, the linear prediction head of the neural network was eliminated, resulting in the neural network outputting (i.e., the output of Bi-GRU in the designed neural network model) $\mathbf{F}_n \in \mathbb{R}^{1\times128}$ (where $n = 1, \ldots, N$) for N segments of each patient. By using the attention mechanism, different weights were assigned to the different $\mathbf{F}_n$. Specifically, this was accomplished in the following way:

As shown in Figure 4a, $\mathbf{F}_n$ (where $n = 1, \ldots, N$) is concatenated into a new feature $\mathbf{F}' \in \mathbb{R}^{N\times128}$, which is then fed into the Feature Attention module (Figure 4b). In the Feature Attention module, $\mathbf{F}'$ performed both global maximum pooling (*MaxPool*) and global average pooling (*AvgPool*), which was then concatenated. The concatenated features were passed through two fully connected layers with a ReLU activation function between them. Finally, the weights $\mathbf{W} \in \mathbb{R}^{N\times1}$ were obtained from this process. The Feature Attention module can be described as

$$\mathbf{W} = FC(\sigma(FC(MaxPool(\mathbf{F}') + AvgPool(\mathbf{F}')))) = \mathbf{W_1}(\sigma(\mathbf{W_0}(\mathbf{F}'_{\mathbf{max}} + \mathbf{F}'_{\mathbf{avg}}))) \tag{2}$$

where σ denotes ReLU, + denotes concatenation, $\mathbf{W_0} \in \mathbb{R}^{16\times2}$, $\mathbf{W_1} \in \mathbb{R}^{1\times16}$, and $\mathbf{F}'_{\mathbf{max}}$ denotes the feature after $\mathbf{F}'$ performing global maximum pooling, and $\mathbf{F}'_{\mathbf{avg}}$ denotes the feature after $\mathbf{F}'$ performing global average pooling.

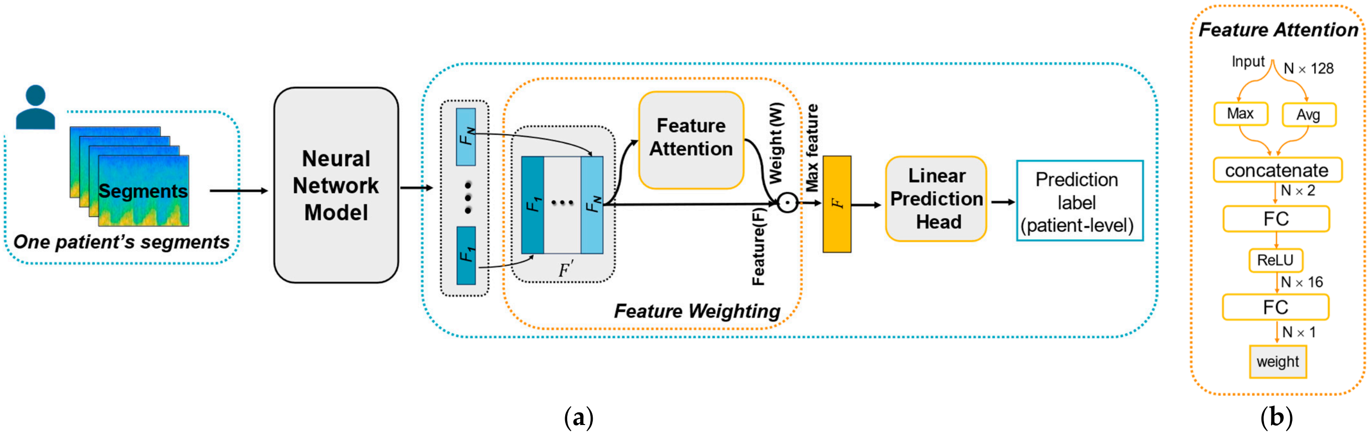

Figure 4. (**a**) Structure of the Feature Weighting method. (**b**) Structure of the Feature Attention module. N corresponds to the number of segments for one patient. The two figures do not contain the batch dimension.

After the above processing, the feature weighting was performed, and the maximum value of each matrix column was taken, which can be described as

$$\mathbf{F} = \max(\mathbf{F}' \odot \mathbf{W}) \tag{3}$$

where $\odot$ denotes the Hadamard product, i.e., feature weighting by elemental multiplication, and max denotes taking the maximum value of each matrix column.

Finally, the new feature $\mathbf{F} \in \mathbb{R}^{1\times128}$ was fed into the linear prediction head to obtain the prediction label of the patient.

To investigate whether the proposed method of feature weighting is effective, the proposed method was compared with three other methods. In these three methods, the linear prediction head of the neural network as in Figure 3 was not eliminated, so each segment had its own detection likelihood. These three methods are now described in detail:

1. Method of "Voting": It is a traditional method. When a patient had only one segment, the predicted class for that segment was considered as the predicted class for that patient; when a patient had several segments, the predicted class for that patient was determined by the majority of the segments.
2. Method of "Max": It takes the maximum value of the detection likelihoods as the final likelihoods. Specifically, for the output of each patient, i.e., $p_i = \left\{p_{i,\text{harsh}}, p_{i,\text{blowing}}\right\}$

for $i = 1, 2, \ldots, N$, where N is the number of segments, and $p_{i,\text{harsh}}$ ($p_{i,\text{blowing}}$) is the likelihood of the specific segment predicted to be "Harsh" ("Blowing"), the patient's detection result is

$$p_{\text{patient}} = \left\{ \max_{1 \le i \le N} \{p_{i,\text{harsh}}\}, \max_{1 \le i \le N} \left\{ p_{i,\text{blowing}} \right\} \right\} \quad (4)$$

3. Method of "Average": It takes the average value of the detection likelihoods as the final likelihoods. Specifically, for $p_i = \left\{ p_{i,\text{harsh}}, p_{i,\text{bowing}} \right\}$ as above, the patient's detection result is:

$$p_{\text{patient}} = \left\{ \frac{1}{N} \sum_{i=1}^{N} p_{i,\text{harsh}}, \frac{1}{N} \sum_{i=1}^{N} p_{i,\text{blowing}} \right\} \quad (5)$$

At the patient-level detection, the optimizer, learning rate, β_1, β_2, weight decay, and loss function were the same as the segment-level detection. The batch size was set to eight patients, since one patient may have multiple segments.

4.4. Method of Finding the Contribution of Other Murmur Characteristics

In this section, the method to find the contribution of other murmur characteristics is given. At the segment-level, the Grading, Pitch, Timing, and Shape labels described in Section 2 were sequentially fed into the neural network model as a priori information with the same conditions as mentioned in Section 4.2. Since these characteristics are not easily accessible, they can only be used to explore the connection between other characteristics and murmur quality and were not used for patient-level detection above. As shown in Figure 5, the labels are firstly encoded as One-Hot Encoding; then, the discrete labels are changed into continuous embedding vectors by the Embedding layer, and finally, they concatenate with the features (i.e., F as the output of Bi-GRU in the designed model) extracted by the neural network model and were fed into the linear prediction head to produce the prediction labels.

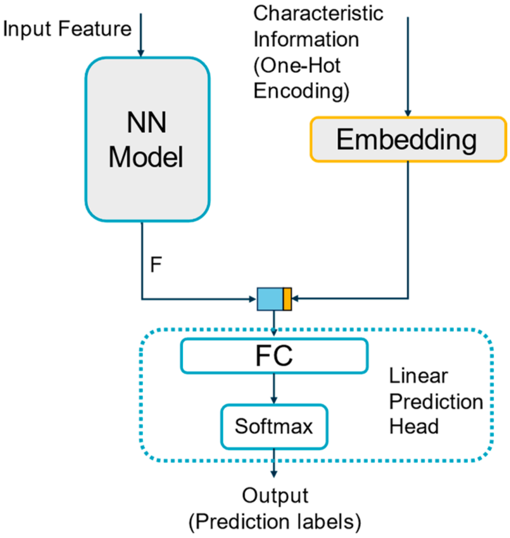
Input Feature
Characteristic Information (One-Hot Encoding)
NN Model
Embedding
F
FC
Softmax
Linear Prediction Head
Output (Prediction labels)

Figure 5. Methods for adding characteristic information.

4.5. Method of Data Augmentation

The limitations of the data volume in the analysis dataset—it only contained 1266 segments with an average length of 4.1 s, whereas the entire dataset had 7338 segments with an average length of 4.2 s (if the same segmentation was applied) greatly hampered the effectiveness of neural networks. Thus, data augmentation was tried in the analysis dataset.

The data augmentation methods included speed increasing, speed decreasing, frequency shifting, time and frequency masking, and speed decreasing and increasing. Specifically, the method of speed increasing speeded up the original heart sounds to 1.5× and 2.0× speed. The method of speed decreasing slowed down the original heart sounds to 0.8× and 0.6× speed. The method of frequency shifting shifted the values of the log-Mel spectrogram up by 50 (from approximately 0–900 Hz to 490–2000 Hz) and 25 (from approximately 0–1380 Hz to 220–2000 Hz) in the frequency dimension. The method of time and frequency masking randomly masked 15% and 30% of values in both the time and frequency dimensions [39]. Each of these methods doubled the amount of original data (1266 segments). The method of speed decreasing and increasing contained both speed decreasing and increasing as mentioned above, which quadrupled the amount of original data. After performing data augmentation on the training set of each fold, the data were fed into the neural network model for five-fold cross-validation as above in Section 4.2 to evaluate the model performance.

4.6. Evaluation Metrics

The metrics of accuracy, precision, recall, and F1-score were used to evaluate all the algorithms above, and the exact calculation methods are described below.

The special values were explained as follows:

- TH (True Harsh): number of correctly detected "Harsh".
- TB (True Blowing): number of correctly detected "Blowing".
- FH (False Harsh): number of incorrectly detected "Harsh".
- FB (False Blowing): number of incorrectly detected "Blowing".

In addition, H is for "Harsh" and B is for "Blowing".

The evaluation metrics were calculated as follows:

1. Accuracy: percentage of the number of correctly detected segments (patients) to the total number of segments (patients).

$$Accuracy = \frac{TH + TB}{TH + TB + FH + FB} \tag{6}$$

2. Precision: percentage of the number of correctly detected "Harsh" ("Blowing") segments (patients) to the total detected "Harsh" ("Blowing") segments (patients).

$$Precision_H = \frac{TH}{TH + FH}, Precision_B = \frac{TB}{TB + FB}. \tag{7}$$

3. Recall: percentage of the number of correctly detected "Harsh" ("Blowing") segments (patients) to the total real "Harsh" ("Blowing") segments (patients).

$$Recall_H = \frac{TH}{TH + FB}, Recall_B = \frac{TB}{TB + FH}. \tag{8}$$

4. F1-score: weighted average of precision and recall.

$$F1_H = 2 * \frac{Precision_H * Recall_H}{Precision_H + Recall_H}, F1_B = 2 * \frac{Precision_B * Recall_B}{Precision_B + Recall_B}. \tag{9}$$

A comprehensive and thorough understanding of the model's performance is given by these metrics, as they indicate the model's ability to correctly detect the murmur quality (accuracy), the extent to which it produces misdetections (precision) as well as misses of the correct class (recall), and the different performances in detecting the two classes (F1-score).

5. Results

This section gives the hyperparameters and conditions of the five-fold cross-validation; then, it shows the results. In the tables of results in this section, the values are given as the mean ± standard deviation of the metrics for the five folds of the five-fold cross-validation, and all the metrics correspond to the definition in Section 4.6. The number of segments and patients corresponds to the description in Section 3.

5.1. Settings

As described in Section 4, when training the neural network at the segment-level, the Adam optimizer was used, the learning rate was set to $3 * 10^{-5}$, β_1 = 0.9, β_2 = 0.98, the weight decay was 0.01, and the loss function was the cross-entropy loss function; while at the training at the patient-level, the optimizer, learning rate, β_1, β_2, weight decay, and loss function were the same as the segment-level. The batch size was set to eight patients, since one patient may have multiple segments. At the five-fold cross-validation, the separation of five folds was based on patients, so segments from one patient are not included simultaneously in the training and validation process. The conditions for the implementation of the five-fold cross-validation are listed in Table 3.

Table 3. Conditions.

Designation	Parameters
CPU	Intel(R) Xeon(R) E5-2680 v4
Memory	32 GB
GPU	NVIDIA GeForce RTX 3080Ti
Software Environment	Python 3.11 PyTorch 2.2.1 Cuda 12.1.1

5.2. The Results of the Ablation Analysis at the Segment-Level

In this ablation analysis, the SE-Block was removed to explore the effect of the channel attention. Table 4 shows the results. The accuracy of the model had a significant improvement of 2.4% after adding the SE-Block between the CNN-Block and Bi-GRU. Not only that, but all other metrics (precision, recall, and F1-score) were improved. It indicated that 32 channels of the feature map have different importance for the detection of the murmur quality, and the SE-Block can effectively distinguish the importance, which helps the Bi-GRU for feature extraction.

Table 4. Results of ablation analysis. "Without SE" refers to no SE-Block between the CNN-Block and Bi-GRU, and "SE" refers to the opposite.

Model	**Accuracy (%)**	**Harsh**			**Blowing**		
		Precision (%)	**Recall (%)**	**F1-Score (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Without SE	66.4 ± 6.5	70.2 ± 3.7	74.4 ± 11.7	71.9 ± 7.3	61.5 ± 9.6	54.9 ± 6.3	57.5 ± 5.7
SE	**68.8 ± 5.0**	**72.4 ± 2.6**	**75.9 ± 8.9**	**73.9 ± 5.5**	**63.8 ± 7.5**	**58.6 ± 4.7**	**60.8 ± 4.4**

5.3. The Results of Comparison between the Proposed Model and Other Models at the Segment-Level

Since there is no relevant deep learning model to detect the murmur quality to the best of our knowledge, the results of the designed model were compared with the results of some well-known models without pretrained, such as SqueezeNet [40], ECAPA-TDNN [41], EfficientNet B0 [42], MobileNet V3 [43], Resnet50 [44], GoogleNet [45], and DenseNet [46]. In addition, the results of all these models were obtained by five-fold cross-validation in the same conditions following the same data preprocessing.

Table 5 demonstrates that the F1-score for the "Harsh" murmur is all higher than that of the "Blowing" murmur (e.g., 73.9% vs. 60.8%), which indicates that all the models were less effective at detecting "Blowing" murmurs compared to "Harsh" murmurs. This may be due to the features of "Blowing" murmurs being less obvious than those of "Harsh" murmurs, as in the log-Mel spectrograms in Figure 2.

The proposed model exhibited higher performance compared to other models. Specifically, the accuracy was 68.8%, the F1-score for the "Harsh" murmur was 73.9%, and the precision for the "Blowing" murmur was 63.8%, all of which were the highest among the models. This result highlights the advantages of the combination of the CNN and gated recurrent neural network (RNN) with the SE-block, since none of the other models employed the structure associated with RNN. In the designed model, the CNN with the SE-block focuses on the extraction of image features for the acoustic spectrograms, whereas GRU can compensate for the CNN by extracting temporal features. In addition, the designed model uses a smaller number of CNN layers and feature map channels compared to other models, which can reduce the parameters of the model and mitigate the overfitting in this dataset.

Table 5. Performance of different models at the segment-level.

Model	Accuracy (%)	Harsh			Blowing		
		Precision (%)	Recall (%)	F1-Score (%)	Precision (%)	Recall (%)	F1-Score (%)
SqueezeNet [40]	65.9 ± 5.3	71.7 ± 3.0	70.0 ± 12.4	70.2 ± 7.1	59.6 ± 8.2	59.9 ± 8.9	59.0 ± 4.7
ECAPA-TDNN [41]	66.9 ± 5.3	74.8 ± 6.1	66.7 ± 4.3	70.4 ± 4.5	58.3 ± 5.3	67.3 ± 9.1	62.3 ± 6.7
EfficientNet B0 [42]	66.5 ± 5.0	71.1 ± 3.3	72.7 ± 8.7	71.7 ± 5.4	60.3 ± 7.2	57.6 ± 6.1	58.5 ± 4.8
MobileNet V3 [43]	66.7 ± 3.0	69.3 ± 2.5	**78.4 ± 5.2**	73.5 ± 2.8	61.9 ± 4.6	49.7 ± 6.5	54.9 ± 4.6
Resnet50 [44]	67.1 ± 3.5	73.3 ± 8.5	75.0 ± 17.4	71.8 ± 7.9	63.8 ± 7.9	55.8 ± 19.5	57.0 ± 6.6
GoogleNet [45]	67.5 ± 3.1	**75.4 ± 3.2**	67.3 ± 9.1	70.7 ± 4.8	59.7 ± 4.1	**67.8 ± 8.9**	**63.0 ± 3.4**
DenseNet [46]	67.5 ± 3.6	71.7 ± 3.4	74.9 ± 9	72.8 ± 4.5	61.9 ± 5.0	56.9 ± 10.1	58.5 ± 5.4
Proposed model	**68.8 ± 5.0**	72.4 ± 2.6	75.9 ± 8.9	**73.9 ± 5.5**	**63.8 ± 7.5**	58.6 ± 4.7	60.8 ± 4.4

5.4. The Results of the Proposed Method at the Patient-Level

5.4.1. Comparison between Different Methods

This section shows the results of comparison methods in Section 4.3. As shown in Table 6, the method of "Feature Weighting" is the most effective (highest accuracy (73.6%) and all other metrics). The method of "Max" follows as the second most effective, whereas the methods of "Voting" and "Average" are found to be the least successful.

Table 6. Performance of different methods at the patient-level. The number of patients for this result is 164 with 90 labeled as "Harsh" and 74 labeled as "Blowing".

Model	Accuracy (%)	Harsh			Blowing		
		Precision (%)	Recall (%)	F1-Score (%)	Precision (%)	Recall (%)	F1-Score (%)
Voting	69.5 ± 6.8	68.0 ± 11.2	82.7 ± 5.8	74.1 ± 8.1	71.5 ± 8.2	53.2 ± 8.6	60.5 ± 6.6
Max	72.7 ± 4.3	72.4 ± 9.9	82.2 ± 4.0	76.4 ± 5.1	72.4 ± 10.8	61.6 ± 12.1	65.6 ± 9.5
Average	68.1 ± 3.9	69.2 ± 8.1	77.4 ± 9.2	72.2 ± 3.6	67.0 ± 15.7	58.8 ± 4.3	61.6 ± 6.7
Feature Weighting	**73.6 ± 6.3**	**73.0 ± 9.8**	**83.0 ± 11.3**	**76.8 ± 7.1**	**76.0 ± 17.0**	**63.3 ± 7.9**	**67.8 ± 7.8**

5.4.2. The Model's Performance under Different Murmur Characteristics

To thoroughly evaluate the proposed algorithms at the patient-level, a statistic of the detection accuracy across different types of PCGs was conducted. Note that the two patients with "Crescendo" (a type of "Shape") and the one patient with "Late-systolic" (a type of "Timing") were ignored in the statistics due to their small numbers. Figure 6 illustrates the accuracy of different types of PCGs (different types of "Timing", "Shape", "Grading" and "Pitch"). For the types of "Timing" and "Shape", the detection accuracy of "Early-systolic" and "Decrescendo" was significantly lower than the other two classes (most of the Early-systolic murmurs are Decrescendo murmurs in this dataset). For the types of "Grading", the PCGs of II/VI were detected significantly less accurately than the other two types. In addition, the accuracy of murmur quality detection of the "High" murmur was much higher than that of the "Low" and "Medium" murmur.

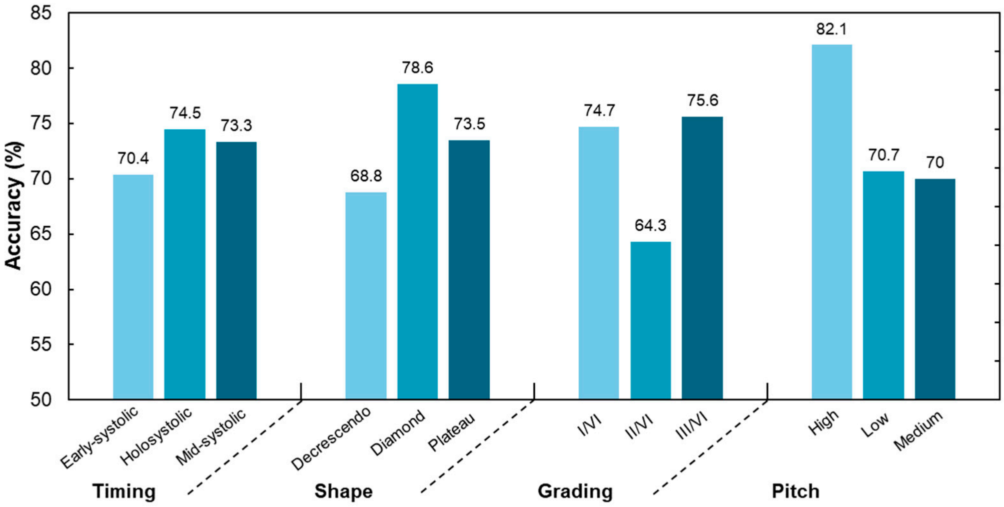

Figure 6. The accuracy of different types of PCGs, including different types of "Timing", "Shape", "Grading", and "Pitch". Specific values for each type of PCG can be found in Table 2.

5.5. The Results after the Other Characteristics Feeding

Table 7 shows the results after feeding other characteristics to the model using the method in Section 4.4. The results show that the accuracy of the model is improved (1.1%) most significantly when fed with the "Timing" label, whereas there is no improvement when the model was fed with the "Grading" or "Pitch" label. In addition, the performance gain of the neural network model was greater for the "Blowing" murmur than for the "Harsh" murmur when the "Timing" or "Shape" labels were fed into it.

Table 7. Model performance when other characteristic information was used.

Murmur Characteristics	Accuracy (%)	Harsh			Blowing		
		Precision (%)	Recall (%)	F1-Score (%)	Precision (%)	Recall (%)	F1-Score (%)
Not used *	68.8 ± 5.0	72.4 ± 2.6	**75.9 ± 8.9**	**73.9 ± 5.5**	63.8 ± 7.5	58.6 ± 4.7	60.8 ± 4.4
Grading	67.1 ± 3.4	74.7 ± 3.4	67.3 ± 6.9	70.6 ± 4.2	59.0 ± 3.9	66.9 ± 7.1	62.4 ± 3.7
Pitch	68.7 ± 5.3	73.7 ± 4.1	72.9 ± 6.5	73.3 ± 5.0	62.0 ± 6.9	62.6 ± 5.9	62.2 ± 5.8
Timing	**69.9 ± 4.8**	75.6 ± 5.1	74.0 ± 14.3	73.6 ± 7.1	**65.4 ± 8.7**	64.0 ± 12.3	63.4 ± 4.1
Shape	69.1 ± 4.0	**76.4 ± 3.8**	69.5 ± 8.6	72.4 ± 4.9	61.6 ± 5.5	**68.6 ± 7.9**	**64.5 ± 4.0**

* "Not used" means no characteristic information was used.

5.6. The Effect of Data Augmentation

The data augmentation was tried following the method in Section 4.5; however, it had minimal impact. The results are shown in Table 8. For the metric of accuracy, there was only a maximum improvement of 0.5% compared to the original data with some metrics decreased, so the data augmentation was not used for segment- and patient-level detection above. It is evident that the traditional data augmentation method is ineffective in mitigating the negative impact of insufficient data volume. Therefore, it is crucial to collect more data in the future to adequately support the training of neural network models.

Table 8. Results of data augmentation. The number of segments used for data augmentation is 1266, with 747 labeled as “Harsh” and 519 labeled as “Blowing”.

Data Augmentation Method	Accuracy (%)	Harsh			Blowing		
		Precision (%)	Recall (%)	F1-Score (%)	Precision (%)	Recall (%)	F1-Score (%)
- *	68.8 ± 5.0	72.4 ± 2.6	75.9 ± 8.9	73.9 ± 5.5	63.8 ± 7.5	58.6 ± 4.7	60.8 ± 4.4
Speed decreasing	67.9 ± 4.5	**78.9 ± 5.9**	64.1 ± 13.7	69.3 ± 7.9	59.8 ± 6.2	**73.2 ± 12.0**	**65.0 ± 3.9**
Speed increasing	**69.3 ± 4.8**	75.8 ± 3.3	70.4 ± 9.6	72.7 ± 5.8	62.2 ± 6.3	67.6 ± 6.9	64.4 ± 4.2
Frequency shifting	68.9 ± 4.1	71.8 ± 2.7	**78.2 ± 9.8**	**74.5 ± 5.0**	**65.2 ± 7.1**	55.5 ± 8.5	59.2 ± 4.5
Frequency and time masking [39]	68.7 ± 5.0	73.2 ± 1.8	74.2 ± 12.4	73.1 ± 6.8	64.1 ± 8.7	60.9 ± 7.1	61.6 ± 2.6
Speed decreasing and increasing	68.4 ± 5.2	76.0 ± 3.0	68.4 ± 12.7	71.2 ± 7.4	61.6 ± 7.7	68.4 ± 8.1	64.1 ± 3.3

* “-” means no data augmentation method was used.

6. Discussion

The proposed deep neural network algorithms for the detection of murmur quality have the potential to be employed in computer-aided auscultation devices to assist the automatic auscultation. This section discusses the method’s ability to choose the feature of the important segment and conducts a performance analysis under different murmur characteristics; then, it expresses the findings about other characteristics’ effects on quality detection and the limitations of this study.

6.1. The Significance of the Model’s Ability to Choose the Important Segment

The difficulty of detecting the murmur quality varies in different auscultation locations and cardiac cycles (i.e., different segments) for a given patient, although the murmur quality remains consistent for that patient. For instance, a “Harsh” murmur may be present in patients with pulmonary valve or arterial stenosis [16], and the AV and PV are closer to the aortic and pulmonary valves, respectively, resulting in a stronger and more easily recognizable murmur in the AV and PV heart sounds in these patients. Similarly, both types of murmurs may be present in patients with ventricular septal defect and mitral or tricuspid regurgitation [16], and the MV and TV are closer to the mitral and tricuspid valves, respectively, leading to a stronger and more readily detectable murmur in the MV and TV heart sounds in these patients. In addition to these physiologic reasons, the possible presence of ambient noise contamination in a patient’s segments can also lead to variations in the difficulty of recognizing the murmur quality for different segments.

The method of “Feature Weighting” took this view into account. Differences in the difficulty of segments from a single patient in detecting the murmur quality lead to differences in the importance of the features extracted from the different segments. Features of easily detectable segments outweigh features of segments that are hard to detect. This method used “Feature Attention” to assign different weights to the features of different segments, enabling distinction of the importance of features from different segments.

The method of “Max” considered this view as well. When a segment is easier to detect, the greater the confidence the neural network has, and the higher the likelihood of the correct class; and vice versa, the likelihood is relatively low, although it is correct. This method could select the segment with the highest likelihood value from all segments of a single patient; that is, it considered the likelihood of easily detectable segments as the likelihood of the patient while ignoring the likelihood of segments that were hard to detect.

However, both the method of “Voting” and the method of “Average” failed to consider this view. When using the method of “Voting”, it is possible that a patient had a greater number of tough segments than easy segments, and the results would favor the tough segments over the easy ones. The “Average” method just averaged the likelihood values and did not consider the importance of the different segments.

The results in Table 4 confirmed this view and show the significance of the model’s ability to choose the important segment. The detection result was better (73.6% and 72.7%) using the method of “Feature Weighting” and “Max”, and it was poorer (69.5% and 68.1%) using the method of “Voting” and “Average”, which was only comparable to the segment-level.

The results also show that the "Feature Weighting" method outperforms the "Max" method, which was presumably because it can distinguish the importance of different segments at the feature level, whereas the "Max" method cannot. In summary, the "Feature Weighting" method shows promising potential and can be used for similar problems with any new model in the future.

6.2. Performance Analysis of the Proposed Model under Different Murmur Characteristics

Figure 6 in Section 5.4.2 shows the different performances of the proposed model under different murmur characteristics. This section conducts a performance analysis based on it.

The significantly low accuracy of "Early-systolic" and "Decrescendo" was attributed to the fact that the early-systolic murmur disappears shortly before the mid-systole period [2]; thus, it has a short duration, which makes it difficult for the model to extract features.

In this dataset, grade labeling may have deviated from the original definition [2]. The murmurs were classified as I/VI by default when not all of a patient's auscultation locations were recorded. So, the PCGs of I/VI may be louder while still being classified as I/VI. Therefore, among the three grading types of PCGs, II/VI may have the highest proportion of low-intensity PCGs in fact. Based on this fact, due to the model's poor ability to detect the quality of low-intensity murmurs, the PCGs of II/VI had a lower accuracy than the other two types.

In addition, the accuracy of the "High" murmur was much higher than that of the "Low" and "Medium" murmur. This discrepancy may indicate that the "Low" and "Medium" murmurs are more likely to overlap with the normal heart sounds (the first (S1) and second (S2) heart sounds) in the frequency domain (because the frequency range of normal heart sounds is relatively low (mainly 20–150 Hz [47])), making feature extraction and detection more challenging.

To visualize the above discussion, Figure 7 shows the waveforms and log-Mel spectrograms of different types of PCGs for 0–1.5 s (about 3 cardiac cycles) to succinctly present the PCGs. Figure 7a is a correctly categorized log-Mel spectrogram of PCG, which has the timing of "Holosystolic", the grading of "III", and the pitch of "High". Figure 7b–d shows three misclassified log-Mel spectrograms of "Early-systolic" timing, "II" grading, and "Low" pitch, respectively. As Figure 7, the duration of the systolic murmur is shorter in "Early-systolic" PCG compared to "Holosystolic" PCG; the color of the systolic log-Mel spectrogram is lighter in PCG with II grading compared to PCG with III grading, which means that the systolic murmur is less loud for the PCG with II grading; and "Low" PCG has a lower frequency distribution than "High" PCG. This is consistent with the discussion above.

In conclusion, the model shows inadequacy in detecting the murmur quality for murmurs of short duration, low intensity, and low pitch. These findings can point the way to future research.

6.3. Findings about Other Characteristics' Effect

In this section, some findings regarding the association between other labels and murmur quality through the result in Section 5.5 are discussed. The lower usefulness of "Pitch" may be because the frequency-domain features reflected by "Pitch" are already embodied in the log-Mel spectrograms and are easily extracted by the neural networks. And the "Grading" label represents the loudness features of the murmur, which are less relevant to the murmur quality, leading to the lower usefulness of "Grading". However, extracting time-domain features from log-Mel spectrograms is harder for neural networks compared to extracting frequency-domain features, and the "Timing" label reflecting time-domain features can make up for this, thus improving the model performance. The "Shape" label proved to be less effective than the "Timing" label when it comes to fitting the newly extracted features of the neural network, despite it also reflecting temporal characteristics, resulting in little improvement of only 0.3%. Also, the greater performance gain for the

"Blowing" murmur than for the "Harsh" murmur with "Timing" or "Shape" labels feeding might suggest that the "Blowing" murmur is more related to time-domain features than the "Harsh" murmur.

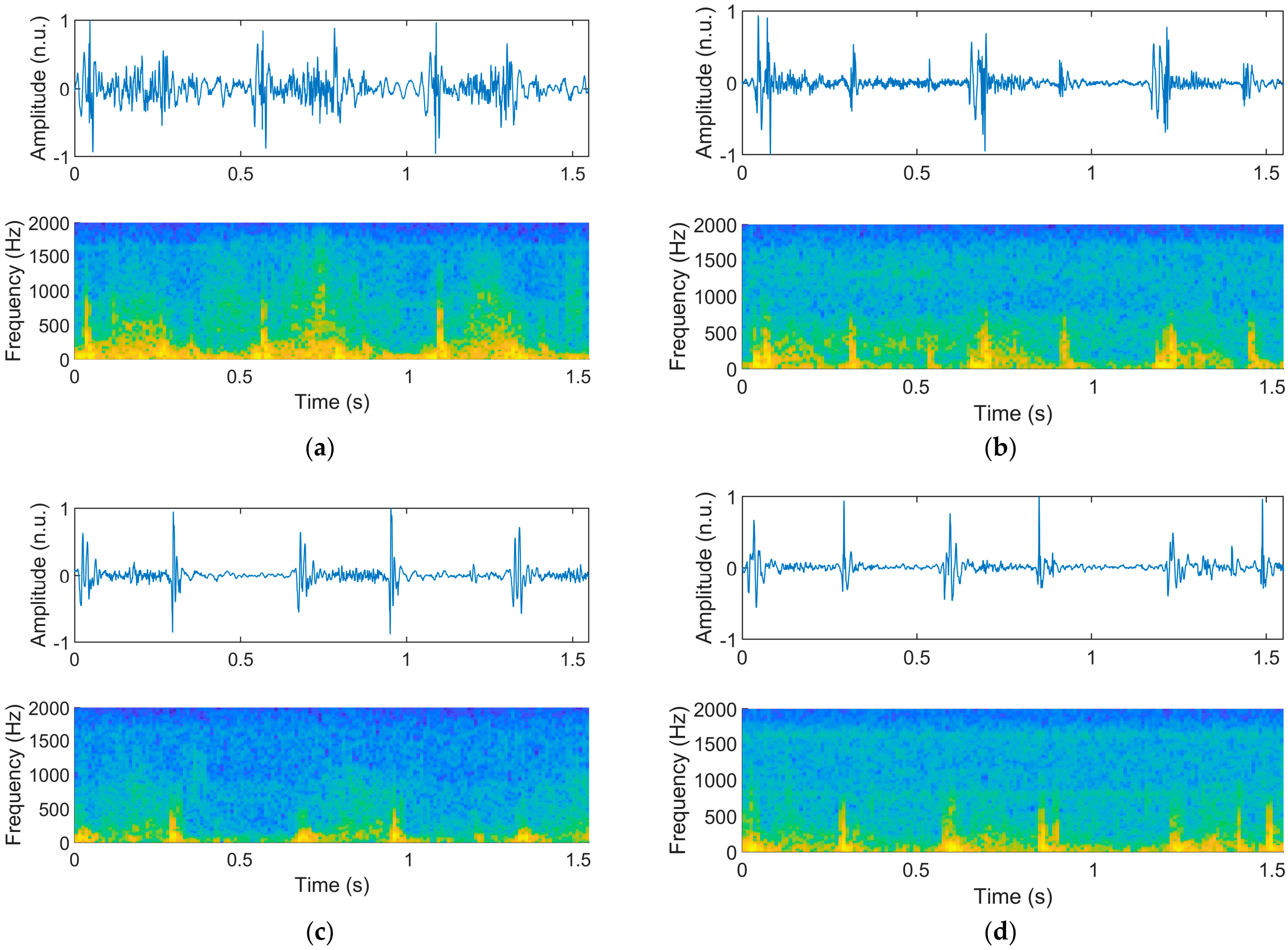

Figure 7. The waveforms and log-Mel spectrograms of different types of PCGs: (**a**) the correctly categorized PCG with the timing of "Holosystolic", the grading of "III", and the pitch of "High"; (**b**) the misclassified PCG with the timing of "Early-systolic"; (**c**) the misclassified PCG with the grading of "II"; (**d**) the misclassified PCG with the pitch of "Low". For succinctness, the PCGs only show 0–1.5 s (about 3 cardiac cycles), not all the segments.

This phenomenon illustrated that for the murmur quality detection problem, the joint use of different classes of labels may have a facilitating effect. It inspired us to explore multi-label classification tasks in future work, which may help the model extract heart sound features better than single-label detection tasks. Meanwhile, when designing the model, correctly using both time-domain representations and spectrograms may have better results than using only spectrograms.

6.4. Limitations of the Study

For segments' detection, besides using the designed model, other models were used to conduct the five-fold cross-validation, the lowest accuracy was 65.9%, and the highest accuracy was only 68.8% (the designed model). The proposed model offered small improvements over other models. For patients' detection, although there was an improvement, the highest accuracy was only 73.7%.

This is mainly due to three reasons. Firstly, the small sample size of the dataset limits the performance of the deep neural network models, as mentioned above. Secondly, detecting murmur quality is difficult because it primarily relies on the subjective feelings of annotators without a specific standard or a related representation. There is a lot of uncertainty with the labeling of murmur quality. Establishing a standardized approach for this task through deep learning models and human judgment would be meaningful in the future. The third reason is the presence of ambient noise. The sounds were recorded in an ambulatory environment with speaking, crying, laughing, or stethoscope rubbing noise [30], which is a great challenge for model training.

7. Conclusions

In this study, the deep neural network algorithms for the detection of murmur quality were proposed, which could be applied to larger datasets and employed in computer-aided auscultation devices to assist the automatic auscultation, find the relationship between murmur quality and CVDs in depth, and help to establish a standardized approach for the murmur quality detection task. The use of the proposed "Feature Attention" module significantly improves the model performance at the patient-level (73.6% vs. 69.5%). But for short-duration, low-intensity, and low-pitch murmurs, the model is not effective, which is an issue that needs to be worked on in future studies. At the same time, traditional data augmentation methods do not help much in classifying, and it would be meaningful that more data can be collected to support the training of neural network models in future work. Moreover, the performance of the neural network can be improved to a certain extent with other labels (e.g., "Timing") added into the inputs of the neural network, which is an inspiration for exploring the design of multi-input or multi-label neural networks in the future.

Author Contributions: Conceptualization, T.W. and F.P.; Methodology, T.W.; Software, T.W., Z.H. and S.L.; Validation, T.W. and F.P.; Formal analysis, T.W., Z.H. and S.L.; Investigation, T.W., Z.H. and S.L.; Resources, F.P.; Data curation, T.W.; Writing—original draft preparation, T.W.; Writing—review and editing, T.W. and F.P.; Visualization, T.W.; Supervision, F.P. and Q.Z.; Project administration, F.P.; Funding acquisition, F.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (grant number: 62066042).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data presented in the study are openly available in the Circor DigiScope Phonocardiogram Dataset at https://physionet.org/content/circor-heart-sound/1.0.3/.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Cardiovascular Diseases (CVDs). Available online: https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (accessed on 11 June 2021).
2. Oliveira, J.; Renna, F.; Costa, P.D.; Nogueira, M.; Oliveira, C.; Ferreira, C.; Jorge, A.; Mattos, S.; Hatem, T.; Tavares, T.; et al. The CirCor DigiScope dataset: From murmur detection to murmur classification. *IEEE J. Biomed. Health Inform.* **2021**, *26*, 2524–2535. [CrossRef]
3. Kumar Roy, A.; Misal, A.; Sinha, G.R. Classification of PCG Signals: A Survey. In Proceedings of the National Conference on Recent Advances in Information Technology, Solapur, India, 15–16 February 2014; pp. 22–26.
4. Hanna, I.R.; Silverman, M.E. A history of cardiac auscultation and some of its contributors. *Am. J. Cardiol.* **2002**, *90*, 259–267. [CrossRef] [PubMed]
5. Elola, A.; Aramendi, E.; Oliveira, J.; Renna, F.; Coimbra, M.T.; Reyna, M.A.; Sameni, R.; Clifford, G.D.; Rad, A.B. Beyond heart murmur detection: Automatic murmur grading from phonocardiogram. *IEEE J. Biomed. Health Inform.* **2023**, *27*, 3856–3866. [CrossRef]
6. Dwivedi, A.K.; Imtiaz, S.A.; Rodriguez-Villegas, E. Algorithms for automatic analysis and classification of heart sounds–a systematic review. *IEEE Access* **2018**, *7*, 8316–8345. [CrossRef]
7. Li, S.; Li, F.; Tang, S.; Xiong, W. A review of computer-aided heart sound detection techniques. *BioMed Res. Int.* **2020**, *2020*, 5846191. [CrossRef]

8. Chen, W.; Sun, Q.; Chen, X.; Xie, G.; Wu, H.; Xu, C. Deep learning methods for heart sounds classification: A systematic review. *Entropy* **2021**, *23*, 667. [CrossRef]
9. Xu, C.; Li, X.; Zhang, X.; Wu, R.; Zhou, Y.; Zhao, Q.; Zhang, Y.; Geng, S.; Gu, Y.; Hong, S. Cardiac murmur grading and risk analysis of cardiac diseases based on adaptable heterogeneous-modality multi-task learning. *Health Inf. Sci. Syst.* **2023**, *12*, 2. [CrossRef] [PubMed]
10. Rosenthal, A. How to distinguish between innocent and pathologic murmurs in childhood. *Pediatr. Clin. N. Am.* **1984**, *31*, 1229–1240. [CrossRef]
11. Reyna, M.A.; Kiarashi, Y.; Elola, A.; Oliveira, J.; Renna, F.; Gu, A.; Perez Alday, E.A.; Sadr, N.; Sharma, A.; Kpodonu, J. Heart murmur detection from phonocardiogram recordings: The george b. moody physionet challenge 2022. *PLoS Digit. Health* **2023**, *2*, e0000324. [CrossRef]
12. Lu, H.; Yip, J.B.; Steigleder, T.; Grießhammer, S.; Heckel, M.; Jami, N.V.S.J.; Eskofier, B.; Ostgathe, C.; Koelpin, A. A lightweight robust approach for automatic heart murmurs and clinical outcomes classification from phonocardiogram recordings. In Proceedings of the 2022 Computing in Cardiology (CinC), Tampere, Finland, 4–7 September 2022; pp. 1–4.
13. Walker, B.; Krones, F.; Kiskin, I.; Parsons, G.; Lyons, T.; Mahdi, A. Dual Bayesian ResNet: A deep learning approach to heart murmur detection. In Proceedings of the 2022 Computing in Cardiology (CinC), Tampere, Finland, 4–7 September 2022; pp. 1–4.
14. Wen, H.; Kang, J. Searching for effective neural network architectures for heart murmur detection from phonocardiogram. In Proceedings of the 2022 Computing in Cardiology (CinC), Tampere, Finland, 4–7 September 2022; pp. 1–4.
15. Safara, F.; Doraisamy, S.; Azman, A.; Jantan, A.; Ramaiah, A.R.A. Multi-level basis selection of wavelet packet decomposition tree for heart sound classification. *Comput. Biol. Med.* **2013**, *43*, 1407–1414. [CrossRef]
16. Oh, S.L.; Jahmunah, V.; Ooi, C.P.; Tan, R.-S.; Ciaccio, E.J.; Yamakawa, T.; Tanabe, M.; Kobayashi, M.; Acharya, U.R. Classification of heart sound signals using a novel deep WaveNet model. *Comput. Methods Programs Biomed.* **2020**, *196*, 105604. [CrossRef] [PubMed]
17. Baghel, N.; Dutta, M.K.; Burget, R. Automatic diagnosis of multiple cardiac diseases from PCG signals using convolutional neural network. *Comput. Methods Programs Biomed.* **2020**, *197*, 105750. [CrossRef] [PubMed]
18. Nogueira, D.M.; Ferreira, C.A.; Gomes, E.F.; Jorge, A.M. Classifying heart sounds using images of motifs, MFCC and temporal features. *J. Med. Syst.* **2019**, *43*, 168. [CrossRef] [PubMed]
19. Rath, A.; Mishra, D.; Panda, G.; Pal, M. Development and assessment of machine learning based heart disease detection using imbalanced heart sound signal. *Biomed. Signal Process. Control* **2022**, *76*, 103730. [CrossRef]
20. Zeinali, Y.; Niaki, S.T.A. Heart sound classification using signal processing and machine learning algorithms. *Mach. Learn. Appl.* **2022**, *7*, 100206. [CrossRef]
21. Chorba, J.S.; Shapiro, A.M.; Le, L.; Maidens, J.; Prince, J.; Pham, S.; Kanzawa, M.M.; Barbosa, D.N.; Currie, C.; Brooks, C.; et al. Deep Learning Algorithm for Automated Cardiac Murmur Detection via a Digital Stethoscope Platform. *J. Am. Heart Assoc.* **2021**, *10*, e019905. [CrossRef]
22. Singstad, B.-J.; Gitau, A.M.; Johnsen, M.K.; Ravn, J.; Bongo, L.A.; Schirmer, H. Phonocardiogram classification using 1-dimensional inception time convolutional neural networks. In Proceedings of the 2022 Computing in Cardiology (CinC), Tampere, Finland, 4–7 September 2022; pp. 1–4.
23. Chen, W.; Sun, Q.; Wang, J.; Wu, H.; Zhou, H.; Li, H.; Shen, H.; Xu, C. Phonocardiogram classification using deep convolutional neural networks with majority vote strategy. *J. Med. Imaging Health Inform.* **2019**, *9*, 1692–1704. [CrossRef]
24. Li, J.; Ke, L.; Du, Q.; Ding, X.; Chen, X. Research on the classification of ecg and pcg signals based on bilstm-googlenet-ds. *Appl. Sci.* **2022**, *12*, 11762. [CrossRef]
25. Patwa, A.; Rahman, M.M.U.; Al-Naffouri, T.Y. Heart murmur and abnormal pcg detection via wavelet scattering transform & a 1d-cnn. *arXiv* **2023**. [CrossRef]
26. Gündüz, A.F.; Talu, F. Pcg frame classification by classical machine learning methods using spectral features and mfcc based features. *Avrupa Bilim Teknol. Derg.* **2022**, *42*, 77–82. [CrossRef]
27. Hu, W.; Lv, J.; Liu, D.; Chen, Y. Unsupervised feature learning for heart sounds classification using autoencoder. *J. Phys. Conf. Ser.* **2018**, *1004*, 012002. [CrossRef]
28. Ballas, A.; Papapanagiotou, V.; Delopoulos, A.; Diou, C. Listen2yourheart: A self-supervised approach for detecting murmur in heart-beat sounds. In Proceedings of the 2022 Computing in Cardiology (CinC), Tampere, Finland, 4–7 September 2022; pp. 1–4.
29. Panah, D.S.; Hines, A.; McKeever, S. Exploring wav2vec 2.0 model for heart murmur detection. In Proceedings of the 2023 31st European Signal Processing Conference (EUSIPCO), Helsinki, Finland, 4–8 September 2023; pp. 1010–1014.
30. Oliveira, J.; Renna, F.; Costa, P.; Nogueira, M.; Oliveira, A.C.; Elola, A.; Ferreira, C.; Jorge, A.; Bahrami Rad, A.; Reyna, M.; et al. The CirCor DigiScope Phonocardiogram Dataset. Available online: https://physionet.org/content/circor-heart-sound/1.0.3/ (accessed on 10 May 2022).
31. Hu, Y.; Zhao, Y.; Liu, J.; Pang, J.; Zhang, C.; Li, P. An effective frequency-domain feature of atrial fibrillation based on time–frequency analysis. *BMC Med. Inform. Decis. Mak.* **2020**, *20*, 308. [CrossRef] [PubMed]
32. Avanzato, R.; Beritelli, F. Automatic ECG diagnosis using convolutional neural network. *Electronics* **2020**, *9*, 951. [CrossRef]
33. Minic, A.; Jovanovic, L.; Bacanin, N.; Stoean, C.; Zivkovic, M.; Spalevic, P.; Petrovic, A.; Dobrojevic, M.; Stoean, R. Applying recurrent neural networks for anomaly detection in electrocardiogram sensor data. *Sensors* **2023**, *23*, 9878. [CrossRef]

34. Noroozi, Z.; Orooji, A.; Erfannia, L. Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction. *Sci. Rep.* **2023**, *13*, 22588. [CrossRef] [PubMed]
35. Freeman, A.R.; Levine, S.A. The clinical significance of the systolic murmur: A study of 1000 consecutive non-cardiac cases. *Ann. Intern. Med.* **1933**, *6*, 1371–1385. [CrossRef]
36. Baggish, A.L.; Sabatine, M.S. Systolic murmurs. In *Decision Making in Medicine: An Algorithmic Approach*, 3rd ed.; Mushlin, S.B., Greene, H.L., Eds.; Elsevier Health Sciences: Philadelphia, PA, USA, 2009; p. 70. ISBN 978-0-323-04107-2.
37. Rabiner, L.R.; Schafer, R.W. *Theory and Applications of Digital Speech Processing*, 1st ed.; Prentice Hall Press: Upper Saddle River, NJ, USA, 2010; ISBN 0-136-03428-4.
38. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
39. Park, D.S.; Chan, W.; Zhang, Y.; Chiu, C.-C.; Zoph, B.; Cubuk, E.D.; Le, Q.V. SpecAugment: A simple data augmentation method for automatic speech recognition. *arXiv* **2019**, arXiv:1904.08779.
40. Iandola, F.N.; Han, S.; Moskewicz, M.W.; Ashraf, K.; Dally, W.J.; Keutzer, K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. *arXiv* **2016**, arXiv:1602.07360.
41. Desplanques, B.; Thienpondt, J.; Demuynck, K. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. *arXiv* **2020**, arXiv:2005.07143.
42. Tan, M.; Le, Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 10–15 June 2019; pp. 6105–6114.
43. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
44. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
45. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
46. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 22–25 July 2017; pp. 4700–4708.
47. Tsai, K.-H.; Wang, W.-C.; Cheng, C.-H.; Tsai, C.-Y.; Wang, J.-K.; Lin, T.-H.; Fang, S.-H.; Chen, L.-C.; Tsao, Y. Blind monaural source separation on heart and lung sounds based on periodic-coded deep autoencoder. *IEEE J. Biomed. Health Inform.* **2020**, *24*, 3203–3214. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.