electronics

MDPI

Article

Assisting Heart Valve Diseases Diagnosis via Transformer-Based Classification of Heart Sound Signals

Dongru Yang [1,2,3], Yi Lin [1,2,4], Jianwen Wei [1,2,4], Xiongwei Lin [5], Xiaobo Zhao [1,2,4], Yingbang Yao [1,2,4], Tao Tao [1,2,3], Bo Liang [1,2,3] and Sheng-Guo Lu [1,2,3,4,*]

1 Guangdong Provincial Research Center on Smart Materials and Energy Conversion Devices, Guangzhou 510006, China
2 Guangdong Provincial Key Laboratory of Functional Soft Condensed Matter, Guangzhou 510006, China
3 School of Integrated Circuits, Guangdong University of Technology, Guangzhou 510006, China
4 School of Materials and Energy, Guangdong University of Technology, Guangzhou 510006, China
5 School of Microelectronics, Shenzhen Institute of Information Technology, Shenzhen 518000, China
* Correspondence: sglu@gdut.edu.cn

Abstract: Background: In computer-aided medical diagnosis or prognosis, the automatic classification of heart valve diseases based on heart sound signals is of great importance since the heart sound signal contains a wealth of information that can reflect the heart status. Traditional binary classification algorithms (normal and abnormal) currently cannot comprehensively assess the heart valve diseases based on analyzing various heart sounds. The differences between heart sound signals are relatively subtle, but the reflected heart conditions differ significantly. Consequently, from a clinical point of view, it is of utmost importance to assist in the diagnosis of heart valve disease through the multiple classification of heart sound signals. Methods: We utilized a Transformer model for the multi-classification of heart sound signals. It has achieved results from four abnormal heart sound signals and the typical type. Results: According to 5-fold cross-validation strategy as well as 10-fold cross-validation strategy, e.g., in 5-fold cross-validation, the proposed method achieved a highest accuracy of 98.74% and a mean AUC of 0.99. Furthermore, the classification accuracy for Aortic Stenosis, Mitral Regurgitation, Mitral Stenosis, Mitral Valve Prolapse, and standard heart sound signals is 98.72%, 98.50%, 98.30%, 98.56%, and 99.61%, respectively. In 10-fold cross-validation, our model obtained the highest accuracy, sensitivity, specificity, precision, and F1 score all at 100%. Conclusion: The results indicate that the framework can precisely classify five classes of heart sound signals. Our method provides an effective tool for the ancillary detection of heart valve diseases in the clinical setting.

Keywords: heart sound signal; multi-classification; audio spectrogram; attention mechanism; transformer

Citation: Yang, D.; Lin, Y.; Wei, J.; Lin, X.; Zhao, X.; Yao, Y.; Tao, T.; Liang, B.; Lu, S.-G. Assisting Heart Valve Diseases Diagnosis via Transformer-Based Classification of Heart Sound Signals. *Electronics* **2023**, *12*, 2221. https://doi.org/10.3390/electronics12102221

Academic Editors: Phivos Mylonas, Katia Lida Kermanidis and Manolis Maragoudakis

Received: 9 March 2023
Revised: 24 April 2023
Accepted: 26 April 2023
Published: 13 May 2023

1. Introduction

Currently, heart valve diseases (HVDs), which include conditions such as mitral Aortic Stenosis (AS), Mitral Regurgitation (MR), Mitral Stenosis (MS), and Mitral Valve Prolapse (MVP) [1], are continually increasing in prevalence and mortality, and has become a great threat to the human health globally [2,3]. Early detection of underlying heart diseases is crucial for improving patient survival rates. While a variety of tests, such as blood tests [4], coronary angiograms [5], electrocardiograms (ECGs) [6], and various imaging techniques [7], have been effectively used in detecting heart diseases. These methods, however, often require specialized equipment and knowledge, as well as high costs. In contrast, auscultation, which is a low-cost, rapid, and straightforward diagnostic method for assessing a patient's heart conditions, has been widely employed by physicians. However, auscultation requires a significant amount of professional knowledge and experience, and training to

diagnose heart conditions based on the analysis of heart sounds (HS) takes considerable time [7].

To address this issue, many researchers proposed the use of artificial intelligence (AI) algorithms to automatically analyze and diagnose HVDs from the HS signals [8–21]. Previous researches were mainly focusing on the dichotomous classification of HSs as normal and abnormal. However, for physicians to make a more detailed and in-depth diagnosis of a patient's heart conditions, it is necessary to pursue a quantitative assessment of HS signals. Therefore, there is a need to further improve the accuracy and efficiency of the multi-classification systems. In addition, most of the HS multi-classification systems proposed in the existing work use traditional neural network models, such as convolutional neural networks (CNN), long-short-term memories (LSTM), recurrent neural networks (RNN), and their variants, all of them have some limitations such as gradient disappearance and explosion, and long training or prediction times [22–24].

To achieve this goal, we proposed a self-attention mechanism-based method for multi-classification of HS signals, along with a data-augmentation technique to improve the model robustness. The proposed model is fine-tuned on HS signals via a pretrained Transformer encoder. Our method offers several advantages over traditional neural network algorithms. Firstly, the self-attention mechanism allows the network to automatically learn important features from HS signals and reduces the dependence on large training datasets. Secondly, the self-attention mechanism helps the network better handle the incomplete or noisy HS signals. Finally, our data augmentation technique takes into account various situations that may arise in practical applications and improves model generalization. We applied our method to a five-category dataset [20] including AS, MR, MS, MVP, and normal HS signals (N), and procured excellent results, as described in Section 4.

The paper is organized as follows: Section 2 summarizes the existing work on HS classification. Section 3 describes the dataset, data preprocessing, data augmentation, model architecture, and experiments. Section 4 presents the experimental results and comparison with other state-of-the-art methods. Section 5 discusses the results, and Section 6 is the conclusion.

2. Literature Review

Over the past several years, a variety of methods have been proposed to detect and diagnose HVDs, in terms of the HS signals and artificial intelligence (AI) methods [8–21]. One common method for detecting or diagnosing HVDs using HS signals involves recording the sounds produced by the heart using a specialized microphone or stethoscope, and analyzing the recording for abnormalities or changes in the sounds that may indicate the presence of a HVD. In 2012, Jia et al. [8] used fuzzy neural networks, which were initially extracted using features such as discrete wavelet transform and Shannon Entropy, to classify the normal and abnormal HS signals collected by themselves. Subsequently, more researchers began to focus on the research and analyze of open-source HS datasets such as the PASCAL dataset [9] as well as the 2016 PhysioNet/CinC Challenge database (PNC) [25], both of which are well-known and popular datasets in the field of HS classification. In 2017, Zhang et al. [10] proposed a binary HS classification (normal or abnormal) using spectrograms and support vector machine (SVM) algorithm, working on PASCAL dataset. Afterwards, these researchers [11] extended the work to PhysioNet dataset, and improved feature extraction using the tensor decomposition method. They obtained the accuracy, sensitivity, and specificity rates of 93.85%, 97% and 92%, respectively. In 2018, Hamidi et al. [12] proposed a curve fitting feature extraction combined with k-nearest neighbors (kNN) algorithm for HS classification and reached the highest accuracy rate of 98%. Recently, Krishnan et al. [16] obtained a balanced accuracy of 85.7% with a sensitivity of 86.73% by use of the deep neural networks (DNN). Based on the same dataset, Deng et al. [15] developed a framework for HS signal classification using the Mel-frequency Cepstral Coefficients (MFCC) combined with convolutional recurrent neural networks and produced an accuracy of 98%. Deng's model took only 0.9–1.2 s to predict a

single HS signal. With the maturity of the transformer structure, Dahr et al. [21] suggested that using the AlexNet to classify the HS signals, all the accuracy, precision, sensitivity, and specificity were achieved with the same accuracy, i.e., 98%. Zeinali et al. [13] made further far-reaching contributions to PASCAL dataset as well as the PhysioNet dataset. They used special filters to eliminate the noise and improve the sound quality and accuracy of feature mining. After that, they achieved an accuracy rate of 87.5% and 95% for multi-class data (3 classes) and 98% for binary classification (normal or abnormal) using the machine learning (ML) algorithms.

Admittedly, previous researchers have contributed greatly to the automatic HS classification based on the AI algorithms, but most of their work has focused on the simple dichotomous work of classifying the HSs into normal or abnormal categories. It is necessary to carry out a quantitative assessment of HS signals for physician to make a more detailed and in-depth diagnosis on the patient's heart condition. Yaseen et al. [20] proposed a five categories dataset containing four categories for abnormal HSs (AS, MR, MS, MVP) and one category for normal HSs. Furthermore, they proposed a multi-classification method based on the DNN and the ML techniques, and obtained 97% in accuracy, 94.5% in sensitivity, and 98.2% in specificity. Oh et al. [17] put forward a Deep WaveNet-based framework for categorizing HSs, and obtained 97% in accuracy, 92.5% in sensitivity, and 98.1% in specificity. Based on the same dataset, Turker et al. [19] further improved the average accuracy and produced an excellent result of 98.38% in accuracy. However, the average computation time for predicting one HS signal was 9.69 seconds, which was a long computing process. Khan [14] implemented further segmentation work on Yaseen's dataset. They subdivided the HS classification problem into five classes', four classes', three classes', and two classes' problems and achieved accuracies of 98.53%, 98.84%, 99.07% and 99.70% respectively for these four problems using Fourier-Bessel Series Expansion-based Empirical Wavelet Transform (FBSE-EWT) and Salp Swarm Optimization Algorithm.

In summary, the field of HS classification using AI algorithms has made significant progresses in recent years. Various methods have been proposed to detect and diagnose the HVDs, by utilizing HS signals and AI methods. Researchers have worked on open-source HS datasets and proposed different classification methods based on features extraction, deep neural networks, and machine learning algorithms. While most of the previous work focused on the simple dichotomous classification, recent studies have proposed more detailed and in-depth classification methods with multi-class datasets, achieving high accuracy, sensitivity, and specificity rates. However, there is still a need for further research and development to improve the accuracy and efficiency of HS multi-classification systems. The use of advanced techniques such as transformers and self-attention mechanism-based models may be the promising directions for obtaining even more accurate and efficient HS classification systems in the future.

3. Materials and Methods

3.1. Dataset Introduction

In this dataset, Yaseen et al. [20] proposed four types of abnormal HS signals, namely AS, MR, MS, MVP, and the normal signals, with 200 recordings for each type. The actual sampling rate is 8 kHz. In addition, all signal amplitudes are normalized between −1 and 1. The length of HS signals is from 1.16 to 3.99 s. Representative examples of each type of HS signals are shown in Figure 1a–e. Figure 1a shows the waveform of AS, Figure 1b the waveform of MR, Figure 1c the waveform of MS, Figure 1d the waveform of MVP, Figure 1e the waveform of standard HS signal, and Figure 1f how the HS signal is collected.

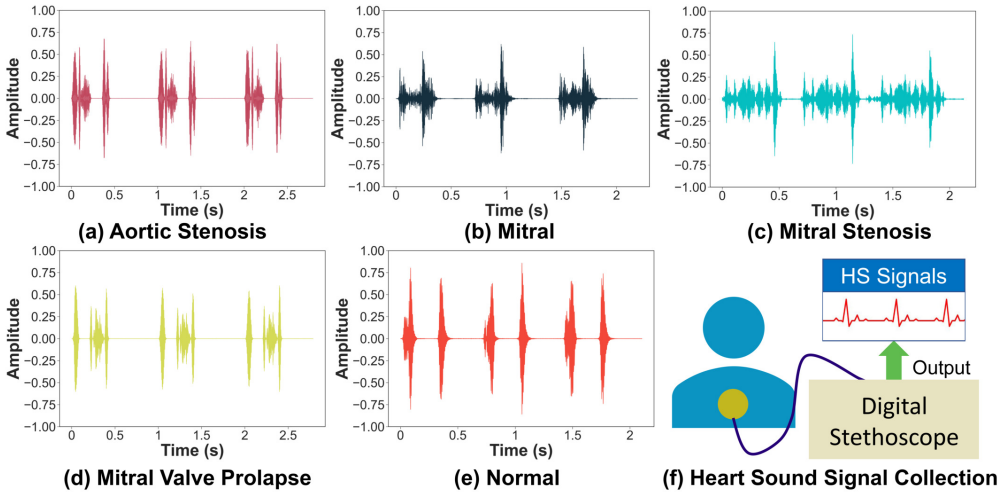


Figure 1. Different types of HS signals. (**a**) shows the waveform of AS, (**b**) the waveform of MR, (**c**) the waveform of MS, (**d**) the waveform of MVP, (**e**) the waveform of standard HS signal, and (**f**) how the HS signal is collected.

3.2. Data Preprocessing

Data Preprocessing is an essential pre-work in deep learning. In this work, to ensure the consistency of the HS signal and to avoid the effects of noise interference on subsequent work, we first denoised the audio signal to reduce the noise interference using adaptive filters. Next, we resampled the HS signals to 16 kHz to ensure that the HS transformer model could learn as more details as possible from the input data while minimizing the amount of computation. To ensure that HS signals of any length can be input into the model, the signals should be trimmed or supplemented. The excess part of the original HS signals must be trimmed if their lengths exceed the input threshold (400 ms). When the original signals were shorter than the threshold, a blank signal should be added at the end to fill the length.

3.3. Data Augmentation

To improve the model's generalization and to prevent over-fitting, we pursued data augmentation processing on the existing dataset. Due to the nature of the original audio data, we mainly augmented the dataset in three domains—the frequency domain, time domain, and time-frequency domain. In our work, eight data augmentation methods, i.e., noise adding, loudness adjustment, amplitude clipping, volume amplification, waveform displacement, pitch adjustment, partial erasure, and speed adjustment. During the data augmentation process, each HS underwent processing using the aforementioned eight methods, with each method applied for ten times. Especially, the noise addition method employed two types of noise-white noise and ambient noise. Furthermore, each of the eight methods randomly selected a value from a specified range during the processing. After the data augmentation, the number of HS signals increased from 1000 to 90,000. The follows are the reasons for using these eight data augmentation methods.

- Noise adding. Adding artificial noise to the HS data can simulate variations that might occur due to the external factors, namely the white noise and environmental conditions. The signal-to-noise ratio is between 5–15 dB. This can help make the model more robust and able to deal with the real-world variability.
- Volume reduction. Refers to reducing the overall volume or loudness of HSs, which can be achieved by reducing the amplitudes of the waveforms. The volume reduction range should be 0.5–1.0 times the original volume. This helps the model learn to recognize HSs at low volumes. The reason for the low volume of HSs may be affected by factors such as the distance between the microphone and the sound source or the level of ambient noise.

- Volume amplification. This involves increasing the overall volume of the HSs, either by increasing the amplitudes of the waveforms, or by applying a gain to the audio signal. The volume reduction range should be 1.0–1.5 times the original volume. This can help our model to learn to recognize HSs that are louder or more dominant in the audio mix.
- Waveform displacement. Waveform displacement involves shifting the waveform of the HSs either forward or backward in time. Waveform displacement ranges from −0.75–1.25. This can help simulate the effects of delays or latency in the recording equipment, or create examples of HSs that are out of synchronization with other audio sources.
- Pitch adjustment. Adjusting the pitch of the HS data can simulate the variations in the heart rate or other physiological processes that might affect the frequency of the HS. Pitch adjustment ranges −0.8–1.2. This can help the model to better handle the changes in heart rates or other physiological conditions.
- Partial erasure. This involves removing or erasing part of the original HS data, either by deleting a section of the waveform or applying a mask to the audio signal. Partial erasure ranges 0–50% of the HS. This can help simulate the effects of missing or incomplete data, which might be caused by issues with the recording equipment or by other factors.
- Speed adjustment. This involves changing the speed at which the HSs were recorded, either by speeding up or slowing down the waveform or by applying a time stretch to the audio signals. The range of speed adjustment is 0.5–1.5 of the original speed. This can help the machine learning model to learn to recognize HSs that are recorded at different speeds, which might be affected by factors such as the age or health or exercise status of the person producing the sounds.

Overall, these techniques can be used to create a large and diverse dataset of artificially generated HS samples that includes a wide range of variations in the data. This can be useful for training our model to recognize and classify different types of HSs, as the models will be exposed to a wider range of examples and will be better able to handle real variations in the data in the future.

The representative HS signals after data augmentation are shown in Figure 2a–i. Figure 2a shows the waveform of the original standard HS signal; Figure 2b the waveform of the standard HS signal after adding noise; Figure 2c the waveform of the standard HS signal after amplitude clipping; Figure 2d the waveform of standard HS signal after volumn aplification; Figure 2e the waveform of standard HS signal after partial erasing; Figure 2f the waveform of standard HS signal after speed adjustment; Figure 2g the waveform of standard HS signal after the pitch adjustment; Figure 2h the waveform of standard HS signal after volume reduction; Figure 2i the waveform of standard HS signal after waveform displacement. These eight methods are used for data enhancement for all HS signals. As can be seen from Figure 2b–i, the red part represents the original HS signals. In contrast, the blue part represents the HS signals after the data augmentation, which introduces differences without changing their essences.

3.4. Model Architecture

In this work, we abandoned the decoder of the encoder-decoder architecture in the ViT architecture [26] and used only the original encoder for multi-classifying HS signals. Also, we added a SoftMax layer at the end of the classification model. To avoid overfitting, we introduced a dropout mechanism (the dropout rate was set as 0.5) before the output of the transformer unit. In the end, we randomly froze 25% of the parameters in each training period to speed up the convergence of the network. Figure 3 shows the workflow of the model we proposed.

In Figure 3, first, to get the input of our model, the HS signals were converted with a 128-dimensional log Mel filter bank each 10 ms, with a Hamming Window of 25 ms to prevent the spectral leakage. In detail, the method of Mel frequency analysis is closer to the

auditory characteristics of human ears. By converting the HS signals into spectrograms, we were able to represent the audio data in a 2D matrix form, which allowed us to leverage the power of Transformer-based models for audio processing. Several studies utilized spectrograms for HS classification, including literatures [15,27]. Our work extended on these studies by utilizing a transformer-based architecture that is capable of modeling long-range dependencies and capturing complex patterns in the HS signals. By considering the advantages and potential effects of using the spectrograms, we provided a more comprehensive discussion of the core ideas and contributions of our study.

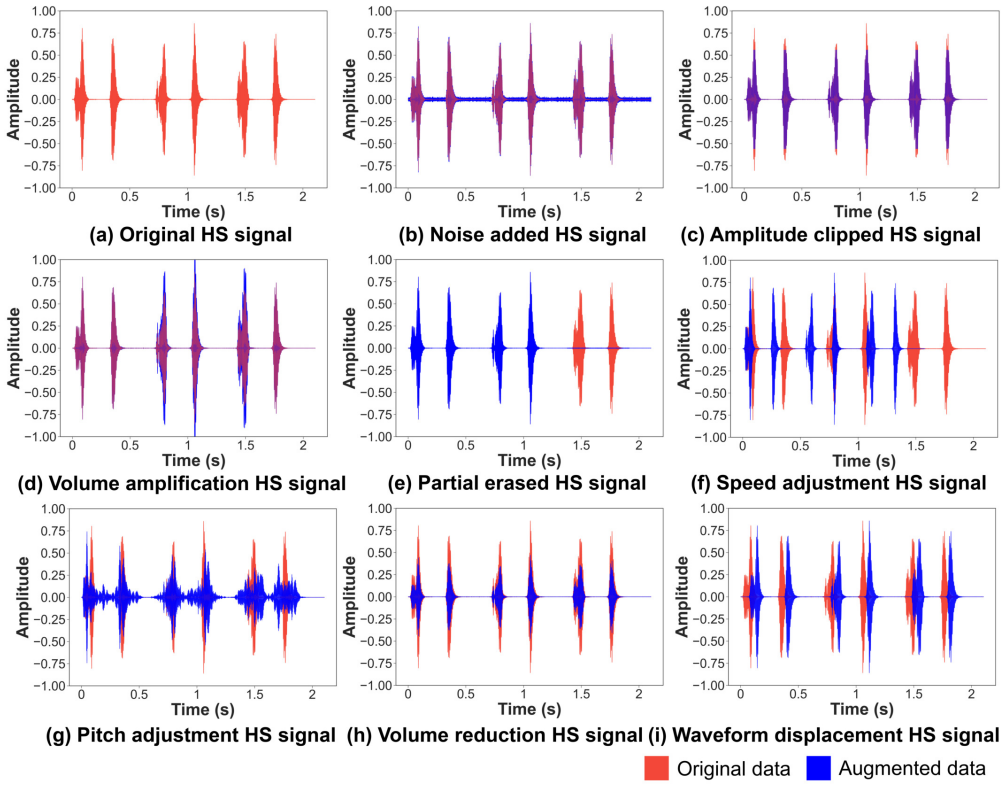


Figure 2. Comparison between the data-augmented and the original HS signal.

The spectrogram sequence was divided into 16 × 16 sizes in the time domain and frequency domain, with an overlap of 6. However, the transformer architecture could not capture the spatial and temporal relationships as the traditional neural networks do. The trainable positional embedding of the same size as the input sequence was added to each patch embedding to ensure that the model could handle the spatial information of the audio spectrogram. At last, we added a class token at the beginning of the input sequence.

Furthermore, the preprocessed HS signal was further processed into the input sequence of the proposed model. The transformer encoder contained an embedding dimension of 768, 12 layers, and 12 heads [26,28]. As a final step, the transformer encoder's output was subjected to a linear layer with a Sigmoid activation function to facilitate the classification.

Figure 4 explains how our workflow works in detail. Firstly, the model was pretrained based on the AudioSet dataset [29], which collected over 2 million audio clips (10 s each) from YouTube videos and tagged the clips with 527 labels. Then we transferred it to the five-classification tasks for HSs, where parameters were randomly frozen and discarded. After that, we validate the model and continuously adjusted the parameters until the model achieved a satisfactory classification level. In the final step, we selected the best trained-model to test on testing dataset (independent from training dataset and validation dataset) to evaluate the performance. Specifically, to begin with, the pretrained model was loaded after being trained on a very large amount of audio data spectrograms.

Then, the network was modified to adapt to the specific needs of the 5-classification task for HSs. During training, HSs were loaded according to the prescribed cross-validation strategy and given parameters, such as batch size. The measures to prevent overfitting, such as dropout and parameter freezing, were implemented during the training process. Following each round of training, the divided validation set was utilized to verify the classification performance of the model, and the training was adjusted as needed to fine-tune the parameters. Once the training and verification process was completed, the optimal model was selected and used for the final testing, which characterized the actual results of the model and potential practical applications.

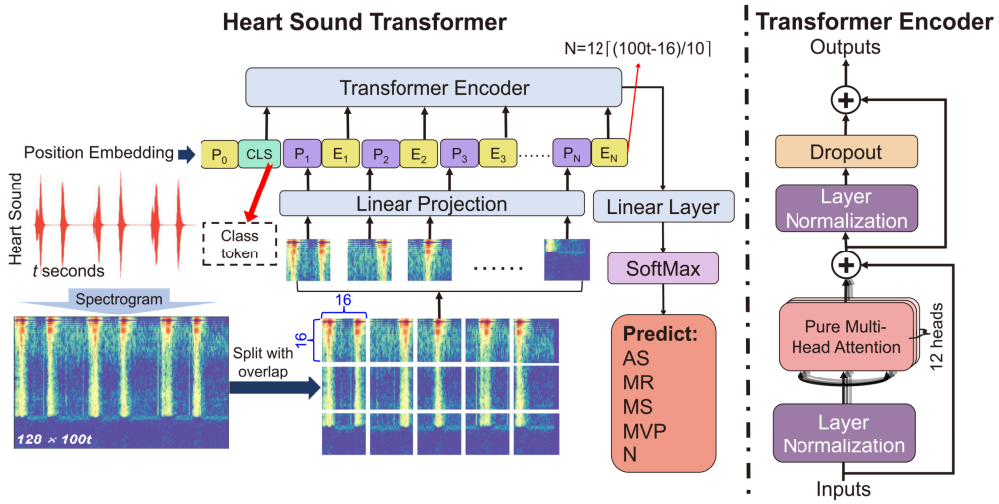


Figure 3. Schematic diagram of the proposed model architecture.

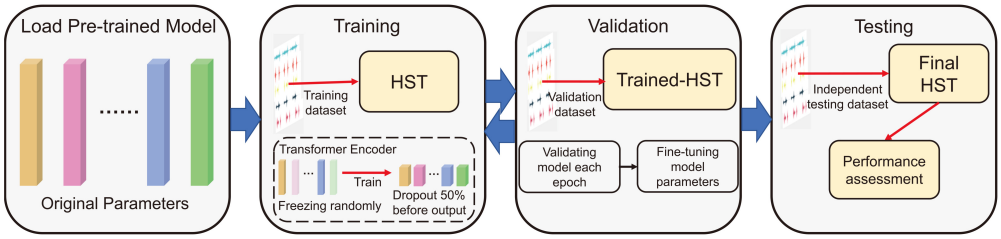


Figure 4. Overview of the integrated workflow.

3.5. Experiment

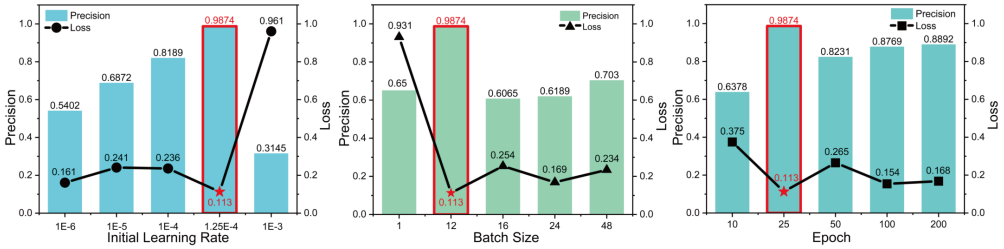
First, we selected a random portion of the orginal dataset as the testing set, i.e., 10% of the data from each category. After the selection, the number of testing set was 20 for each category and the number of training set was 180 for each category. Then, each HS signal was processed using the data augmentation methods mentioned above, and the final number of dataset was 1800 for each category in the testing set and 16,200 for each category in the training set.

We ran a series of simulations with different sets of hyperparameters. For each simulation, 80% of the 1000 original HS signals were randomly selected for training and 20% for validation before data augmentation. The data used for training and validation were independent of each other. The hyperparameter set included the initial learning rate (η), batch size, and training epochs. Figure 5a–c show the relationship among the three hyperparameters—precision, loss, and epoch. Formulas (1) and (2) show the loss function. In addition, we used the Adam optimizer [30] in the proposed model. The final initial learning rate, batch size, and training epoch were 0.000125, 12, and 25, respectively.

$$Loss = \{l_1, ..., l_N\} \quad (1)$$

$$l_n = -[(y_n \cdot \log(\frac{1}{1+\exp(-x)})) + (1-y_n) \cdot \log(1-\frac{1}{1+\exp(-x)})] \quad (2)$$

where, 'N ' denotes the number of batches; 'n' the number of labels.



(a) Initial learning rate with PREC and loss. (b) Batch size with PREC and loss. (c) Epoch with PREC and loss.

Figure 5. The selection process of different hyperparameters with precision & loss.

After finding out the most appropriate set of hyperparameters, the original HS and augmented mixed HS datasets were used for the final model training. Besides, the 5-fold cross-validation method as well as the 10-fold cross-validation method were used in this work.

4. Experimental Results

4.1. Evaluation Criteria

In order to validate the performance of the proposed model, we employed five commonly used metrics, namely accuracy (ACC), precision (PRE), macro precision (macro PRE), specificity (SPEC), macro recall, sensitivity (SENS), and F1 score. Formulas (3)–(9) show how to calculate these five indicators. ACC is the percentage of correctly inferenced heart sounds. PRE is a measure of the model's ability to correctly classify the HS signals. It is calculated as the number of correctly classified signals divided by the total number of objects classified by the model. Macro PRE is calculated as the average precision of the model across all classes. Recall, also known as SENS or the true positive rate, is a metric used to measure the proportion of actual positive instances that the model is able to correctly identify. Macro recall is a measure of the proportion of actual positive instances that the model is able to correctly identify, averaged across all classes. SPEC is a measure of the proportion of actual negative instances that the model is able to correctly identify. F1 score is a balance between the PRE and recall, taking into account both the proportion of the true positive predictions made by the model and the proportion of actual positive instances that the model is able to identify.

$$\text{Accuracy} = \frac{\text{Number of correctly classified samples}}{\text{Total number of samples}} \quad (3)$$

$$\text{PRE} = \frac{TP}{TP+FP} \quad (4)$$

$$\text{Macro PRE} = \frac{\sum_{k=1}^{K} \text{PRE}_k}{K} \quad (5)$$

$$\text{Macro Recall} = \frac{\sum_{k=1}^{K} \text{Recall}_k}{K} \quad (6)$$

$$\text{SENS} = \text{Recall} = \frac{TP}{TP+FN} \quad (7)$$

$$\text{SPEC} = \frac{TN}{TN + FP} \tag{8}$$

$$\text{F1-score} = 2\frac{\text{PRE} * \text{SENS}}{\text{PRE} + \text{SENS}} \tag{9}$$

where TP (True Positive) means the number of HSs correctly classified to the target label. TN (True Negative) means the number of HSs correctly classified to the non-target label. FP (False Positive) means the number of HSs with the non-target label wrongly detected as the target label. FN (False Negative) means the number of HSs with the target label detected as the non-target label. K represents the number of categories of data.

4.2. Results

4.2.1. Confusion Matrices

Figures 6 and 7 show the confusion matrices for each fold of the 5-fold cross-validation and the 10-fold cross-validation, respectively. The X-axis represents the predicted labels exported from the system, while the Y-axis represents the actual labels of the HS signals. Figure 6a–e show the confusion matrix for the best model in each fold of the 5-fold cross-validation on the independent testing set, in order. Figure 7a–j show the confusion matrix of the best model in each fold of the 10-fold cross-validation on the independent testing set, in order. Tables 1 and 2 show the performance of the 5-fold cross-validation and 10-fold cross-validation respectively.

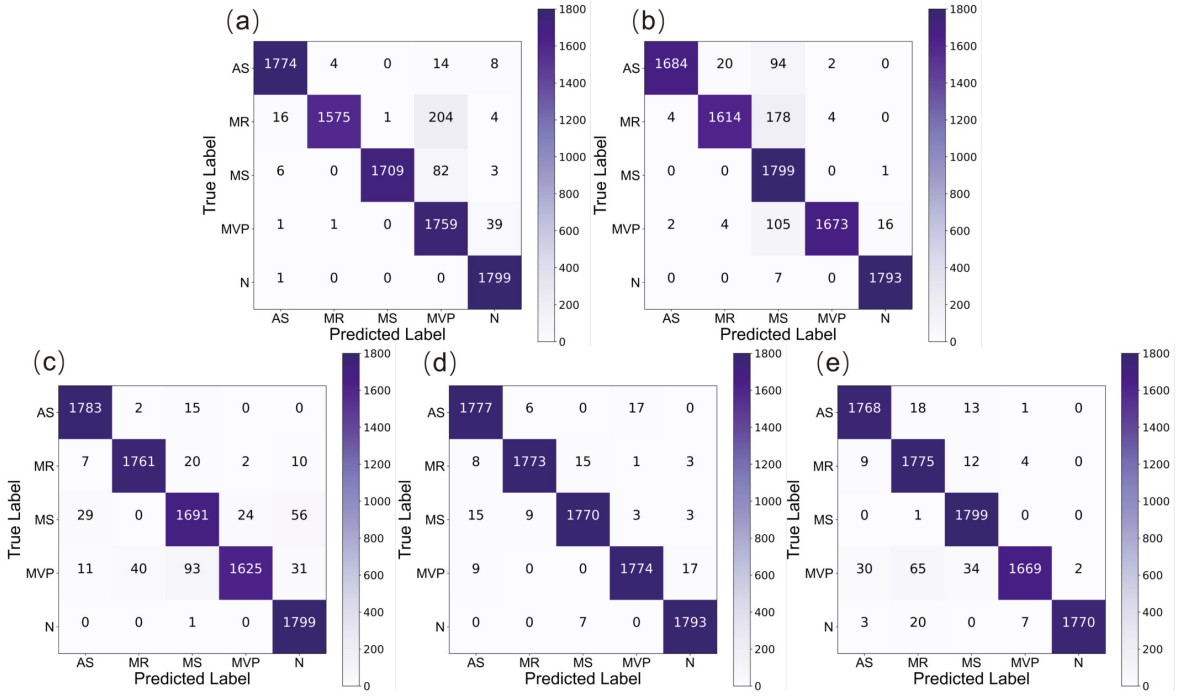


Figure 6. Confusion matrices for the 5-fold cross-validation of the proposed model.

For the 5-fold cross-validation, the model has an average ACC of about 96.68%, an average macro PRE of about 96.34%, an average macro recall of about 96.68%, and an average F1 score of about 96.81%. Through examining the confusion matrix for each fold in more detail, we can see that the highest ACC, macro PRE, macro recall and F1 score on Fold 4, with an ACC of 98.74%, a macro PRE of 98.76%, a macro recall of 98.74%, and F1 score of 98.76%.

For the 10-fold cross-validation, the model has an average ACC of about 99.40% across all folds, with a range of ACC in 97.93–100%. And the model achieved an average macro PRE, an average macro recall, and an F1 score of 99.40%. From the confusion matrix we

can see that the best performing model comes from the Fold 7, with all positive indicators reaching 100%.

Comparing the result of the two cross-validation techniques, one can see that the model's performance is generally similar for both techniques, with high average ACC, macro PRE, macro recall, and F1 scores in both cases. However, there are some slight differences in the range of scores achieved by the model across the fold. For example, the model has a slightly wider range of accuracy scores using 5-fold cross-validation (95–98%) compared to 10-fold cross-validation (98% to 100%), which may indicate that the model's performance is more variable using 5-fold cross-validation. Moreover, for the 10-fold cross-validation, the average ACC, Average macro PRE, average Macro Recall, and average F1 score are close to or even equal to 100% for more than half of the results. One potential reason for the slightly higher average performance of the model using 10-fold cross-validation compared to 5-fold cross-validation could be the additional data points used for training and testing in the 10-fold cross-validation. With 10 folds, the model is trained and tested on a larger number of data points, which provided a more accurate representation of the model's generalizability to new data. This could be particularly important if the dataset is large and diverse, as the additional folds may allow the model to better capture the full range of variation in the data.

However, it is also worth considering the trade-off between the increased evaluation provided by 10-fold cross-validation and the additional computational resources required to run the additional folds. Depending on the size of the dataset and the complexity of the model, 10-fold cross-validation may take significantly longer time to run compared to 5-fold cross-validation, which could be a consideration for some users. In addition, the larger number of folds in the 10-fold cross-validation may make it more difficult to identify patterns or trends in the results, as there are more data points to analyze.

Overall, 10-fold cross-validation may provide relatively better performance due to learning from more data, while 5-fold cross-validation offers a more realistic evaluation closer to the real scenarios. The advantage of 10-fold cross-validation is that it uses more data to train and validate the model, resulting in more accurate estimates of model performance. However, it also requires more computational resources and time to complete the analysis. On the other hand, 5-fold cross-validation offers a more efficient and faster evaluation of model performance, making it more suitable for large datasets or computationally demanding models. However, it may lead to higher variance and overfitting due to a smaller amount of training data. Therefore, it is essential to consider the specific circumstances and objectives when choosing the appropriate cross-validation method.

Table 1. Performance of the 5-fold cross-validation.

Fold	ACC (%)	Macro-PRE (%)	Macro-Recall (%)	F1-Score (%)
1	95.73	96.16	95.73	95.94
2	95.14	95.85	95.14	95.49
3	96.21	96.27	96.21	96.24
4	98.74	98.76	98.76	98.76
5	97.57	97.63	97.56	97.60

4.2.2. Classification Results

Table 3 shows the classification results by using the best model from 5-fold cross-validation. We used PRE, SENS, SPEC, and F1 score to evaluate the performance of the model. The results show that the model has an average PRE of 98.94% and an average F1 score of 98.74%, which indicates a good performance. However, looking at the performance of each category individually, one can find that the model has some limitations in distinguishing between the "MR" class and the "MVP" class, especially in models that are not well trained, as can also be clearly seen from the confusion matrices in Figure 6.

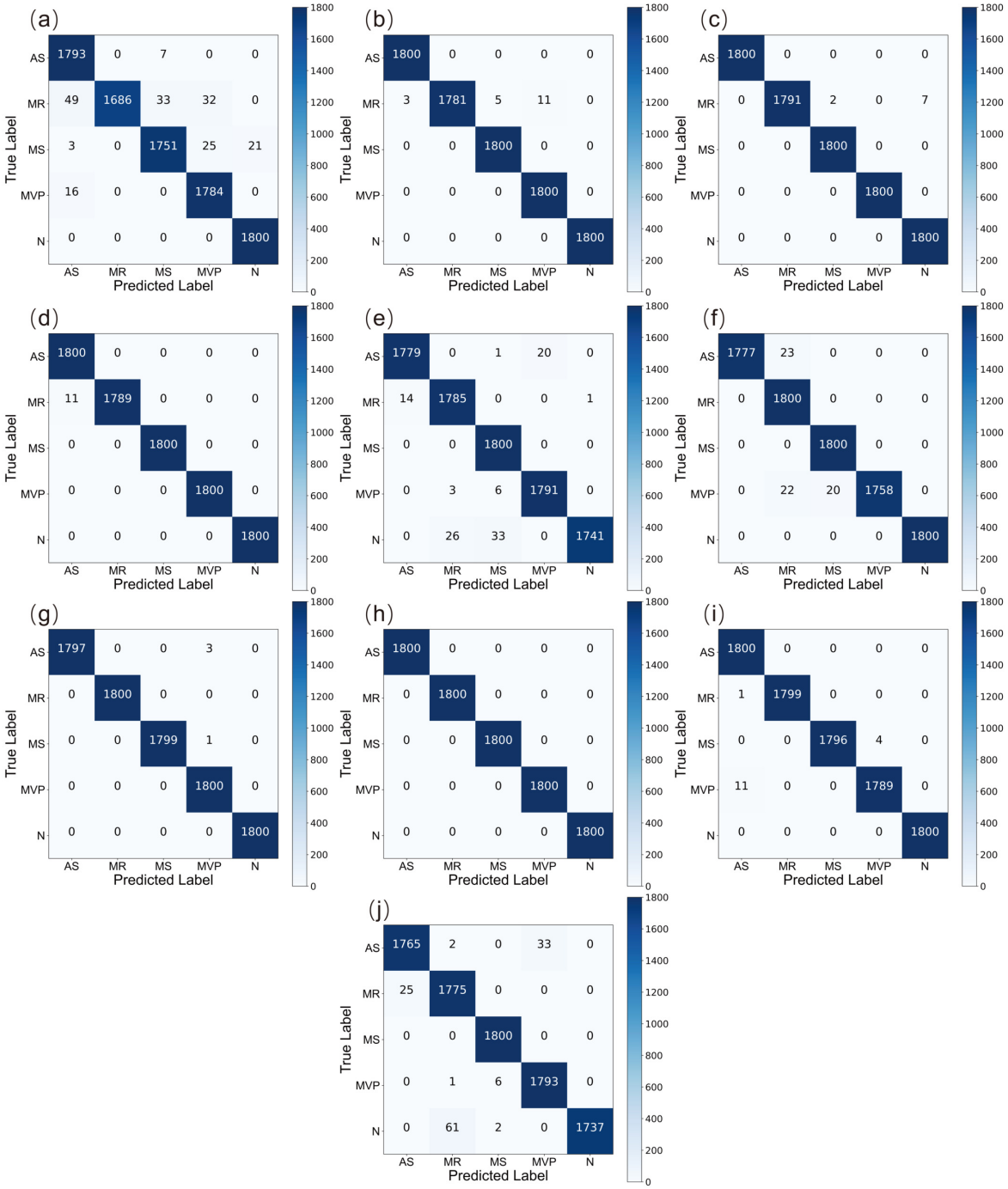


Figure 7. Confusion matrices for the 10-fold cross-validation of the proposed model.

Table 4 shows the classification results using the optimal model from 10-fold cross-validation. This model achieves perfect scores for all indicators and has a promising classification effect. This further demonstrates that the 10-fold cross-validation results in a better HS multi-classification model than 5-fold cross-validation for the model we proposed.

Table 2. Performance of the 10-fold cross-validation.

Fold	ACC (%)	Macro PRE (%)	Macro-Recall (%)	F1-Score (%)
1	97.93	97.97	97.93	97.95
2	99.79	99.79	99.79	99.79
3	99.90	99.90	99.90	99.90
4	99.88	99.88	99.88	99.88
5	99.84	98.86	99.84	98.85
6	99.28	99.29	99.28	99.28
7	99.96	99.96	99.96	99.95
8	100	100	100	100
9	99.82	99.82	99.82	99.82
10	98.56	98.56	98.56	98.56

Table 3. Classification results using the best model from 5-fold cross-validation.

Label	PRE (%)	SENS (%)	SPEC (%)	F1-Score (%)
AS	98.23	98.72	99.56	98.47
MR	99.16	98.50	99.79	98.83
MS	98.77	98.33	99.69	98.54
MVP	98.83	98.56	99.71	98.69
N	98.73	99.61	99.68	99.17

Table 4. Classification results using the best model from 10-fold cross-validation.

Label	PRE (%)	SENS (%)	SPEC (%)	F1-Score (%)
AS	100	100	100	100
MR	100	100	100	100
MS	100	100	100	100
MVP	100	100	100	100
N	100	100	100	100

Figures 8 and 9 show the receiver operating characteristic (ROC) plots for the optimal model for the 5-fold cross-validation as well as the 10-fold cross-validation. From Figure 8, it is clear that the curves for "MVP" and "N" class are relatively linear, while the curves for "AS", "MR", and "MS" class are more curved. This suggests that the model is able to distinguish between the positive and the negative examples for "MVP" and "N" class with relatively high accuracy, but may have more difficulties in distinguishing the positive and from the negative examples for "AS", "MR", and "MS" class. From Figure 9, it is clear that every curve is relatively linear. This suggests that the model has the ability to distinguish between the positive and the negative examples for each class.

Compare the AUC values for each class. For the best model in the 5-fold cross-validation, the AUC values for the different classes range from 0.9996 to 0.9999, with the "N" class having the highest AUC value. This suggests that the model is able to classify the examples in these classes with a high accuracy. For the optimal model in the 10-fold cross-validation, the AUC values for each class are 1. This suggests that after 10-fold cross-validation, the model performs better than the model that has only treated with a five-fold cross-validation.

4.2.3. Comparison with the Published Results

Tables 5 and 6 summarize the results compared with other state-of-the-art works. As discussed in Section 2, many researchers previously committed to computer-aided diagnosis or prognosis of heart diseases with excellent results. Several researchers developed studies on the multi-classification (AS, MR, MS, MVP, and N) HS issue for further medical diagnosis [17,19,20]. Compared with Khan's work [14], our model performs similarly in the 5-fold cross-validation as Khan's, our ACC is 1.57% higher, Macro recall is 3.6% higher

in the 10-fold cross-validation. However, it is worth noting that Khan's work also focuses on the multi-task HS classification, i.e., five, four, three as well as binary classifications.

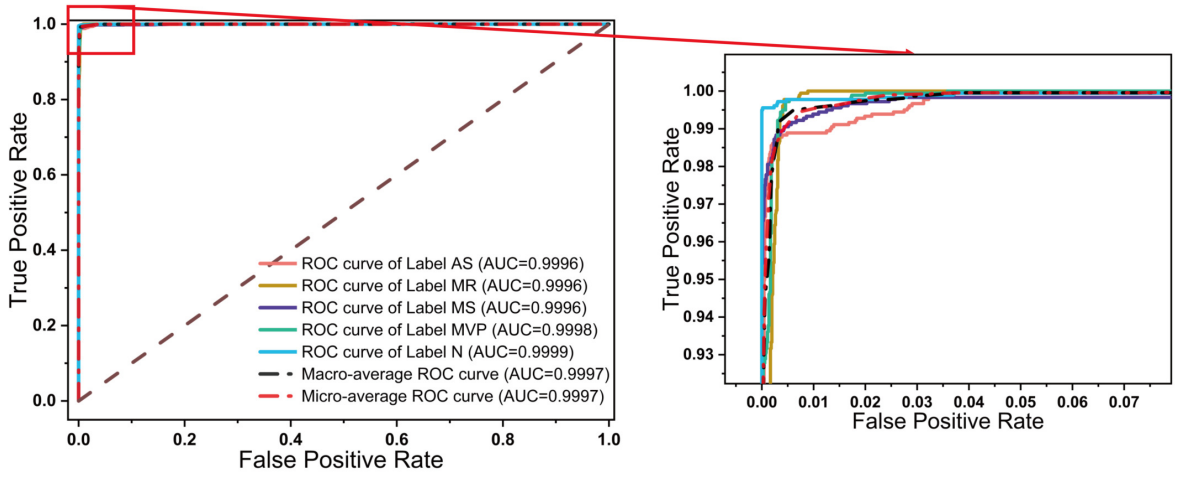


Figure 8. ROC plot for the optimal model for 5-fold cross-validation.

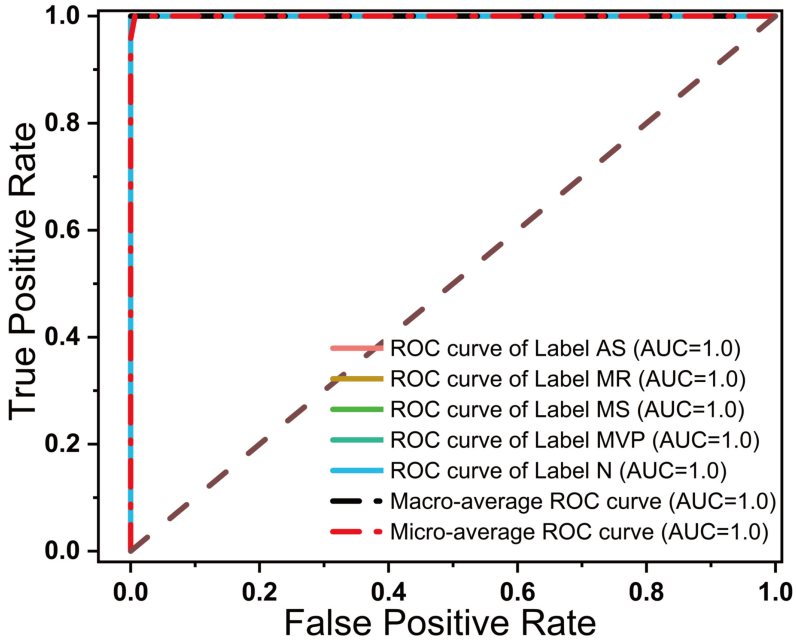


Figure 9. ROC plot for the optimal model for 10-fold cross-validation.

The proposed model can classify the HS signals with a maximum accuracy of 98.74% in a five-fold cross-validation and 100% in a ten-fold cross-validation, as a result of the improved computing power of the computer hardware and the optimization of the computer software. Moreover, for this experiment, we only used the original and augmented datasets for preprocessing and inference and calculated the mean and standard deviation of the required time. The transformer architecture has significant advantages. The preprocessing and inferencing time of the proposed model is only 1.62 ± 0.05 s. Compared with the present works [19], we have achieved not only higher ACC, Macro PRE, and Macro Recall, but also faster processing speeds.

The median value for each evaluation metric has also been calculated in addition to the optimal model. As can be seen from the Tables 5 and 6, the median value provides a more representative measure of the model performance. Specifically, for the 5-fold cross-validation, the ACC, Macro Recall, and Macro PRE are 96.21%, 96.21% and 96.27%, respectively. For the 10-fold cross-validation, these three indicators are 99.80%, 99.81%,

and 99.81%, respectively. These results indicate that the models are highly consistent in performance across different cross-validation folds. It should be noted that the proposed model may be computationally expensive. Our model was run on a Quad RTX 3090 Nvidia graphics card for almost 10 hours. But this is also related to the amount of our data. After data enhancement, our data set has been expanded to a relatively large magnitude.

Table 5. Results of proposed method compared with present works on ACC, Macro Recall, and Macro PRE.

Method Used	Dataset	ACC (%)	Macro Recall (%)	Macro PRE (%)
DNN [20]	5-class dataset [20]	97.00	94.50	-
Deep WaveNet [17]	5-class dataset [20]	97.00	92.50	-
FBSE-EWT [14]	5-class dataset [20]	98.53	99.08	-
Proposed Model *	5-class dataset [20]	98.74	98.76	98.76
Proposed Model **	5-class dataset [20]	100	100	100
Proposed Model ***	5-class dataset [20]	96.21	96.21	96.27
Proposed Model ****	5-class dataset [20]	99.80	99.81	99.81

* The optimal model for 5-fold cross-validation ** The optimal model for 10-fold cross-validation; *** The median model for 5-fold cross-validation **** The median model for 10-fold cross-validation.

Table 6. Results of proposed method compared with present works on ACC and inference time.

Method Used	Dataset	ACC (%)	Time Required (s)
PCG-TEP, INCA & DT [19]	5-class dataset [20]	95.10	1.18
PCG-TEP, INCA & LD [19]	5-class dataset [20]	98.30	7.67
PCG-TEP, INCA & BT [19]	5-class dataset [20]	98.60	19.96
PCG-TEP, INCA & SVM [19]	5-class dataset [20]	99.90	7.59
PCG-TEP, INCA & kNN [19]	5-class dataset [20]	100.00	12.07
Proposed model *	5-class dataset [20]	98.74	1.62
Proposed model **	5-class dataset [20]	100	1.62
Proposed model ***	5-class dataset [20]	96.21	1.62
Proposed model ****	5-class dataset [20]	99.80	1.62

* The optimal model for 5-fold cross-validation ** The optimal model for 10-fold cross-validation; *** The median model for 5-fold cross-validation **** The median model for 10-fold cross-validation.

5. Discussion

From the results obtained above, the merits of our model can be discussed as follows.

Firstly, the proposed framework improves the accuracy of automatic multi-classification of HS signals and the computing speed of prediction. On one hand, one of the main advantages of using self-attention Transformer for HS classification is the improved performance. Self-attention mechanisms allow the model to capture long-range dependencies in the input data, which can be particularly useful for analyzing the sequential data like HSs. On the other hand, another advantage of using self-attention Transformer for HS classification is the efficiency when it processes long sequences of data. Traditional CNN models can struggle with long sequences, as the number of operations required to process them increases linearly with the length of the input data. In contrast, self-attention Transformer can process long sequences more efficiently, as the number of operations required is constant regardless of the length of the input. These can lead to a better classification accuracy compared to other methods. Our method can not only distinguish between the normal and the abnormal HSs but also refine the classification of abnormal HSs into AS, MR, MS, and MVP. When the medical resources, experience, and hardware are insufficient, doctors can use the specific abnormal results to make a quick diagnosis. In the meantime, the typical results can be used to determine whether the patient has recovered to reduce the waste of medical resources.

Secondly, the proposed approach achieves outstanding and promising performance compared to the current work. A diverse approach to audio data enhancement is used to improve the generality and robustness of the model. Furthermore, the spectrograms of

the HS signals contain a large amount of detailed information that can be extracted as the specific features of different types of HSs. In this way, self-attention mechanisms allow the model to attend to different parts of the input data at different times, which can help learn more robust and transferable features. This can lead to better performance on unseen data, making the model more useful. Based on these two points, our model has shown promising results.

It can be seen that the proposed method outperforms existing works in terms of both the classification accuracy and the computing speed. The proposed model is proved to be effective in HS multi-classification. Despite of the achievements of our model, there are still some limitations worth discussing.

One potential limitation of our model is its applicability to different measuring systems. Our model was trained on Yaseen's dataset [31], which is randomly collected from books or websites without the data containing extreme noise. It is possible that the results may be different if the model is applied to HSs recorded using a different type of measuring system. For example, if our model was trained on HSs recorded using an electronic stethoscope, it may not perform as well as on HSs recorded using a chest-mounted sensor. To address this limitation, it will be important for future research to evaluate the performance of our model on different types of measuring systems. This could involve collecting additional data using a range of measuring systems and retraining the model on this expanded dataset. In addition, it will be important to consider any practical considerations related to the use of different systems. For example, chest-mounted sensors may be more costly or less portable than stethoscopes, which could impact the feasibility of using these systems in different settings. Besides, in terms of the potential impact of collecting HSs from different parts of the body, our model is currently based on the best results achieved using open-source datasets. These datasets themselves are sourced from medical official websites or textbooks. However, further optimization is required to account for the variability introduced by HSs collected in different medical determination of locations. For example, the mitral valve is usually ausculated at the left intercostal space, medially from the midclavicular line. In contrast, the agortic valve is usually auscultated in the 2nd parasternal line on the right side of the chest. This is an ongoing area of research for us.

Another potential limitation of our model is its intended recipient. Our model was designed for use by medical professionals, and it is possible that the results may be different if the model is used by laypeople, although our model has been able to specifically and accurately output the type of the underlying cardiac problem (AS, MR, MS, MVP, Normal). For example, medical professionals may have more training and experience in interpreting HSs, which could impact the performance of the model when used by other people. To address this limitation, it will be important for future research to evaluate the performance of our model when used by different groups of recipients. This could involve collecting additional data from a range of recipients and retraining the model on this expanded dataset. It may also be necessary to fine-tune the model or develop new models specifically for use by different groups of recipients. In addition, it will be important to consider any practical considerations related to the use of the model by different groups. For example, medical professionals may require additional training or education to use the model effectively, while laypeople may need more guidance or support to understand the results of the model. Furthermore, it should be noted that a heart disease may present with the presence of multiple HSs concurrently, and our current model demonstrates superior classification performance for individual HSs. However, the primary objective of our study is to facilitate disease diagnosis through HS classification. Even if we cannot accurately distinguish the accompanying HSs, our model can still provide valuable assistance. Hence, we recognize the importance of identifying potential concomitant HSs, and this will be a focus of our ongoing research efforts.

Last but not least, it must be considered the feasibility of implementing our model in different settings when deploying it. Our model relies on self-attention transformer technology, which can be computationally intensive and may require specialized hardware

and software to run effectively. This could impact the feasibility of deploying the model in certain settings, such as resource-limited environments or settings with limited access to specialized hardware. To address this limitation, it will be important for future research to consider the feasibility of deploying our model in different settings. This could involve optimizing the model for use on different hardware platforms or developing more efficient implementations of the model. It may also be necessary to consider alternative approaches, such as using cloud-based services or edge computing, to facilitate the deployment of the model in different settings like wearable devices or Internet of Things (IoT) devices. In addition, it will be important to consider any practical considerations related to the use of the model. This could include the cost and availability of the necessary hardware and software, the training or education required for different groups of users, and any potential ethical considerations.

Overall, in future researches, it will be important to address the applicability of our model to different measuring systems and the intended recipient of our model. We will also consider the feasibility of deploying the model in various situations. By addressing these limitations, we can work to improve the performance and usability of our model for a range of users and measuring systems.

6. Conclusions

For HVDs with high morbidity and mortality, HS signals obtained through auscultation can feedback important information about the patient's heart condition. Our study used these signals to train the proposed model to classify various heart valve diseases, including AS, MR, MS, MVP, and normal types. We validate our model using 5-fold cross-validation as well as 10-fold cross-validation and train it using the datasets after augmentation. In the 5-fold cross-validation, our model achieved the highest ACC of 98.74% and a mean AUC of 0.99. The highest PRE achieved in the proposed model is 99.16% with the label of MR, while the highest ACC is 99.67% with the label of N. In 10-fold cross-validation, our model achieved highest ACC, SENS, SPEC, PRE, and F1 score all at 100%. This study is the first to classify HS signals using a transformer model of HS spectrogram and five classes, achieving extremely high diagnostic reliability. It is expected that, the proposed model, can assist medical experts in the diagnosis of HVDs.

Author Contributions: Conceptualization, D.Y. and S.-G.L.; methodology, D.Y. and Y.L.; software, D.Y.; validation, D.Y. and J.W.; formal analysis, Y.L.; investigation, D.Y. and Y.L.; resources, S.-G.L.; data curation, Y.L. and T.T.; writing—original draft preparation, D.Y. and S.-G.L.; writing—review and editing, D.Y. and X.L.; visualization, X.Z. and Y.Y.; supervision, S.-G.L. and B.L.; project administration, S.-G.L.; funding acquisition, S.-G.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation of China (Grant No. 51372042, 51872053); Guangdong Provincial Natural Science Foundation (2015A030308004); the NSFC-Guangdong Joint Fund (Grant No. U1501246); the Dongguan City Frontier Research Project (2019622101006); and the Advanced Energy Science and Technology Guangdong Provincial Laboratory Foshan Branch-Foshan Xianhu Laboratory Open Fund—Key Project (Grant No. XHT2020-011).

Data Availability Statement: Not applicable.

Acknowledgments: We would like to extend our sincere gratitude to Yaseen, Gui-Young Son, and Soonil Kwon for providing the open-source data that was instrumental in this research.

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

References

1. Li, P.; Ge, J.; Li, H. Lysine acetyltransferases and lysine deacetylases as targets for cardiovascular disease. *Nat. Rev. Cardiol.* **2020**, *17*, 96–115. [CrossRef] [PubMed]
2. Roth, G.A.; Mensah, G.A.; Johnson, C.O.; Addolorato, G.; Ammirati, E.; Baddour, L.M.; Barengo, N.C.; Beaton, A.Z.; Benjamin, E.J.; Benziger, C.P.; et al. Global Burden of Cardiovascular Diseases and Risk Factors, 1990–2019. *J. Am. Coll. Cardiol.* **2020**, *76*, 2982–3021. [CrossRef] [PubMed]
3. Tsao, C.W.; Aday, A.W.; Almarzooq, Z.I.; Alonso, A.; Beaton, A.Z.; Bittencourt, M.S.; Boehme, A.K.; Buxton, A.E.; Carson, A.P.; Commodore-Mensah, Y.; et al. Heart disease and stroke statistics—2022 update: A report from the American Heart Association. *Circulation* **2022**, *145*, e153–e639. [CrossRef] [PubMed]
4. McCullough, P.A.; Olobatoke, A.; Vanhecke, T.E. Galectin-3: A novel blood test for the evaluation and management of patients with heart failure. *Rev. Cardiovasc. Med.* **2011**, *12*, 200–210. [CrossRef]
5. Taylor, A.M.; Thorne, S.A.; Rubens, M.B.; Jhooti, P.; Keegan, J.; Gatehouse, P.D.; Wiesmann, F.; Grothues, F.; Somerville, J.; Pennell, D.J. Coronary artery imaging in grown up congenital heart disease: Complementary role of magnetic resonance and X-ray coronary angiography. *Circulation* **2000**, *101*, 1670–1678. [CrossRef] [PubMed]
6. Lankveld, T.A.R.; Zeemering, S.; Crijns, H.J.G.M.; Schotten, U. The ECG as a tool to determine atrial fibrillation complexity. *Heart* **2014**, *100*, 1077–1084. [CrossRef] [PubMed]
7. Chambers, J.B.; Garbi, M.; Nieman, K.; Myerson, S.; Pierard, L.A.; Habib, G.; Zamorano, J.L.; Edvardsen, T.; Lancellotti, P. Appropriateness criteria for the use of cardiovascular imaging in heart valve disease in adults: A European Association of Cardiovascular Imaging report of literature review and current practice. *Eur. Heart J.-Imaging* **2017**, *18*, 489–498. [CrossRef]
8. Jia, L.; Song, D.; Tao, L.; Lu, Y. Heart Sounds Classification with a Fuzzy Neural Network Method with Structure Learning. In *Advances in Neural Networks—ISNN 2012: 9th International Symposium on Neural Networks, Shenyang, China, 11–14 July 2012*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 130–140. [CrossRef]
9. Bentley, P.; Nordehn, G.; Coimbra, M.; Mannor, S.; Getz, R. Classifying Heart Sounds Challenge. Available online: http://www.peterjbentley.com/heartchallenge/ (accessed on 8 March 2023). [CrossRef]
10. Zhang, W.; Han, J.; Deng, S. Heart sound classification based on scaled spectrogram and partial least squares regression. *Biomed. Signal Process. Control* **2017**, *32*, 20–28. [CrossRef]
11. Zhang, W.; Han, J.; Deng, S. Heart sound classification based on scaled spectrogram and tensor decomposition. *Expert Syst. Appl.* **2017**, *84*, 220–231. [CrossRef]
12. Hamidi, M.; Ghassemian, H.; Imani, M. Classification of heart sound signal using curve fitting and fractal dimension. *Biomed. Signal Process. Control* **2018**, *39*, 351–359. [CrossRef]
13. Zeinali, Y.; Niaki, S.T.A. Heart sound classification using signal processing and machine learning algorithms. *Mach. Learn. Appl.* **2022**, *7*, 100206. [CrossRef]
14. Khan, S.I.; Qaisar, S.M.; Pachori, R.B. Automated classification of valvular heart diseases using FBSE-EWT and PSR based geometrical features. *Biomed. Signal Process. Control* **2022**, *73*, 103445. [CrossRef]
15. Deng, M.; Meng, T.; Cao, J.; Wang, S.; Zhang, J.; Fan, H. Heart sound classification based on improved MFCC features and convolutional recurrent neural networks. *Neural Netw.* **2020**, *130*, 22–32. [CrossRef] [PubMed]
16. Krishnan, P.T.; Balasubramanian, P.; Umapathy, S. Automated heart sound classification system from unsegmented phonocardiogram (PCG) using deep neural network. *Phys. Eng. Sci. Med.* **2020**, *43*, 505–515. [CrossRef]
17. 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]
18. Raza, A.; Mehmood, A.; Ullah, S.; Ahmad, M.; Choi, G.S.; On, B.W. Heartbeat Sound Signal Classification Using Deep Learning. *Sensors* **2019**, *19*, 4819. [CrossRef]
19. Tuncer, T.; Dogan, S.; Tan, R.S.; Acharya, U.R. Application of Petersen graph pattern technique for automated detection of heart valve diseases with PCG signals. *Inf. Sci.* **2021**, *565*, 91–104. [CrossRef]
20. Yaseen; Son, G.Y.; Kwon, S. Classification of Heart Sound Signal Using Multiple Features. *Appl. Sci.* **2018**, *8*, 2344. [CrossRef]
21. Dhar, P.; Dutta, S.; Mukherjee, V. Cross-wavelet assisted convolution neural network (AlexNet) approach for phonocardiogram signals classification. *Biomed. Signal Process. Control* **2021**, *63*, 102142. [CrossRef]
22. Sherstinsky, A. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. *Phys. D Nonlinear Phenom.* **2020**, *404*, 132306. [CrossRef]
23. Levin, E. A recurrent neural network: Limitations and training. *Neural Netw.* **1990**, *3*, 641–650. [CrossRef]
24. Yu, Y.; Si, X.; Hu, C.; Zhang, J. A review of recurrent neural networks: LSTM cells and network architectures. *Neural Comput.* **2019**, *31*, 1235–1270. [CrossRef] [PubMed]
25. Liu, C.; Springer, D.; Li, Q.; Moody, B.; Juan, R.A.; Chorro, F.J.; Roig, J.M.; Silva, I.; Johnson, A.E.; Syed, Z. An open access database for the evaluation of heart sound algorithms. *Physiol. Meas.* **2016**, *37*, 2181–2213. [CrossRef] [PubMed]
26. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Heigold, G.; Gelly, S.; Uszkoreit, J. An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 26–30 April 2020; pp. 285–288. [CrossRef]

27. Chen, K.; Du, X.; Zhu, B.; Ma, Z.; Berg-Kirkpatrick, T.; Dubnov, S. HTS-AT: A hierarchical token-semantic audio transformer for sound classification and detection. In Proceedings of the ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 646–650. [CrossRef]
28. Gong, Y.; Chung, Y.A.; Glass, J. AST: Audio Spectrogram Transformer. In Proceedings of the Interspeech 2021, Brno, Czech Republic, 30 August–3 September 2021; pp. 571–575. [CrossRef]
29. Gemmeke, J.F.; Ellis, D.P.W.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R.C.; Plakal, M.; Ritter, M. Audio Set: An ontology and human-labeled dataset for audio events. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 776–780. [CrossRef]
30. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the ICLR, San Diego, CA, USA, 7–9 May 2015; pp. 1–13. [CrossRef]
31. Cheng, J.; Dong, L.; Lapata, M. Long Short-Term Memory-Networks for Machine Reading. In Proceedings of the EMNLP, Association for Computational Linguistics, Austin, TX, USA, 1–5 November 2016; pp. 551–561. [CrossRef]