Murmur Detection and Clinical Outcome Classification Using a VGG-like Network and Combined Time-Frequency Representations of PCG Signals

Zhongrui Bai[1], Baiju Yan[1], Xiangxiang Chen[2], Yirong Wu[1,2], Peng Wang[2]

[1]School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China
[2]Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China

Abstract

For the George B. Moody PhysioNet Challenge 2022, our team, PhysioDreamfly, developed a deep neural network approach for detecting murmurs and identifying abnormal clinical outcomes from phonocardiograms (PCGs). In our approach, a VGG-like CNN model is used as the classifier. Images consisting of Log-Mel spectrograms and wavelet scalogram that transformed from unsegmented PCGs are used as model inputs. We combined the murmur and outcome labels to address the two tasks as one multi-label task, and introduced a weighted focal loss function to optimize the model. Our murmur detection classifier received a weighted accuracy score of 0.752 (ranked 11th out of 40 teams) and Challenge cost score of 12831(ranked 18th out of 39 teams) on the hidden test set.

1. Introduction

The phonocardiogram (PCG) signal conveys information regarding the mechanical property of the heart. Automated PCG analysis is significant for improving diagnostic efficiency and reducing overall healthcare costs, especially in poor countries that lack trained professionals with auscultation skills and basic healthcare facilities[1].

Traditional PCG signal analysis methods generally first segments the signal, then extracts features from the segmented PCGs, and finally trains a classifier to classify these features[2]. In such methods, the effectiveness of heart sound segmentation predominantly affects the accuracy of classification results. With the development of deep learning and the improvement of hardware environment, the processing of heart sound signals using deep learning methods has become the focus of research[3].

The objective of Challenge 2022 is to identify the presence, absence, or unclear cases of murmurs (task 1) and the normal vs. abnormal clinical outcomes(task 2) using heart sound recordings from multiple auscultation locations on the body and routine demographic information[4]. In this paper, as part of the George B. Moody PhysioNet Challenge 2022, we transformed PCG signals into two-dimensional time-frequency representations and developed a VGG-like CNN model for detecting murmurs and identifying abnormal clinical outcomes. More details are described in the following sections.

2. Methods

For classifier training, we used all the recordings from the public training set, which contains 3163 records from 942 patients[5]. Details about the Challenge 2022 dataset can be found in[1].

2.1. TF-Domain Representation

Two different methods were applied to transform PCGs into time–frequency (TF)-domain representations. And these two representations are used together as the input of the CNN model that described in Section 2.2.

2.1.1. Wavelet Scalogram

The first TF-domain representation is the wavelet scalograms obtained using the continuous wavelet transform. The scalograms obtained from the wavelet transform tends to perform better than Short-time Fourier transform(STFT) spectrums because it has a better time-frequency resolution.

Before the transformation, the original PCGs were downsampled to 1KHz, and subsequently passed through a Butterworth bandpass filter with a pass band of 10-400Hz. We chose the cgaus3 wavelet basis after trying several continuous wavelet basis functions.

The wavelet scaling parameter varies from 1 to 16; the complex wavelet coefficients for the different scales were then returned. And we calculated the absolute values of the coefficients to obtain a two-dimensional array of real numbers. We then downsampled the array in the horizontal axis

by a factor of 8 to reduce the size of the model input. An image of length $N/8$ and height 16 was obtained after the continuous wavelet transform for a heart sound recording of length N.

2.1.2. Log-Mel Spectrogram

The second TF-domain representation is the Log-Mel spectrogram. Mel spectrogram is obtained by transforming the STFT spectrum into the Mel scale, which is more suitable for human auditory responses to different frequencies. Log-Mel spectrogram is the logarithmic transformation of the Mel spectrogram, and it shows a powerful capability in sound identification tasks.

The PCGs were also downsampled to 1KHz before transformation. The PCGs were not filtered here to retain components that might be useful for identifying low quality signals. For Log-Mel spectrum calculation, the window length of STFT was set to 16 and step length of STFT was set to 8, that is, each window has 50% overlapping signal. The number of Mel banks was 16, thus another image of length $N/8$ and height was generated for a heart sound recording of length N.

2.1.3. Combining

Although the average duration of each record in the dataset is 23 seconds, the first 18 seconds of each heart sound recording was used to generate the time-frequency map due to limited video memory size of GPU. For recordings less than 18 seconds, the vacant part of the TF image was filled with zeros. For each recording, we put the two TF images together in the vertical axis direction. Then the TF images of five locations were concatenated in horizontal axis direction, from front to back, PV, TV, AV, MV and Phc. For each patient, a grayscale image of size 32×11250 was generated and used as the input features of the classifier. Finally the TF-domain images of all patients were normalized to a range between [0, 1] using the mean and variance, and this mean and variance were retained as normalization parameters for the test samples. This way, the differences in the values of time-frequency images between patients were preserved.

The Python PyWavelets library and Librosa library were used to generate Wavelet scalograms and Log-Mel spectrograms.

2.2. Model Description

Visual Geometry Group(VGG) models[6] have a clean structure and good performance in image classification. We built the classifier based on the structure of VGG11, as shown in Figure 1. We also used some 3×3 convolution kernels and layer-by-layer increasing convolution filters. However, unlike the original VGG11 model, we built four similar convolution blocks and use batch normalization and dropout in each block. We also reduced the number and size of fully connected layers to reduce the model size.

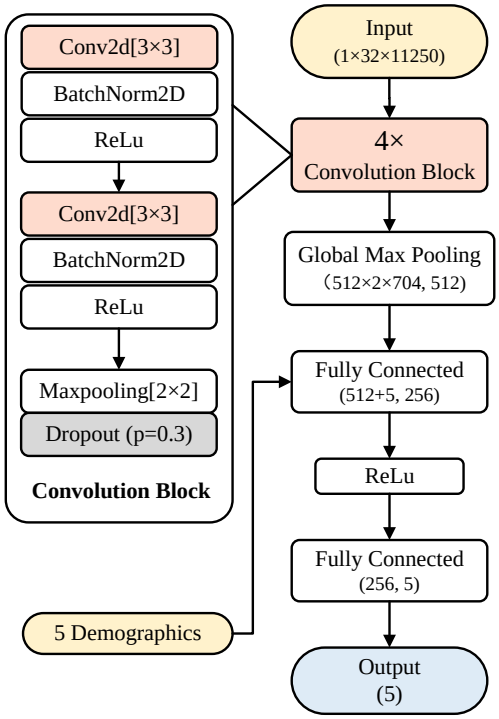

Figure 1. Architecture of the VGG-like model. The numbers in round brackets indicate the sizes of feature maps(channel×height×length, if multiplication signs exist). The number in square brackets inside the convolution module represents the convolution kernel size or pooling size.

The combined time-frequency images are fed into four similar consecutive convolution blocks. Each convolution block contains two duplicate Convolution2D-BatchNorm2D-ReLu structures, as well as a 2×2 size maximum pooling layer to reduce the feature map size, and a dropout layer to avoid overfitting.

Due to the presence of the maximum pooling layer, the height and width of the feature map are reduced to half of the original size after each convolution block. The channels of the feature map become 64 after the first convolution block, and are doubled block by block thereafter, up to 512 channels after the fourth convolution block. A global maximum pooling layer is then used to produce an output array of length 512. Next are two fully-connected layers. In the first fully-connected layer, five normalized demographic features are stitched into the input array. As with

the normalization of the time-frequency images, the means and variances of the demographic features are retained as normalization parameters for the test samples. The second fully connected layer reduces the output to 5, which is the length of the concatenated labels of the two tasks.

2.3. Training Details for Two Tasks

We concatenated the labels of the two tasks as the final labels of the model. The first three values of the labels represent Presence, Unknown and Absence, respectively, and the last two represent Abnormal and Normal clinical outcomes, respectively. This merges two single-label classification tasks into one multi-label classification task. We accomplished both tasks in one training session, and reduced the degree of overfitting in training.

Since the distribution of the data is unbalanced, based on[7], we design a weighted focal loss function, which is defined as follows.

$$Loss = \frac{1}{\mathrm{c}} \sum_{i=1}^{\mathrm{c}} \mathrm{w}_i * FL_i \tag{1}$$

$$FL_i = \begin{cases} -(1-y_i')^{\gamma} \log y_i', & y_i = 1 \\ -y_{\mathrm{i}}^{'\gamma} \log (1-y_i'), & y_i = 0 \end{cases} \tag{2}$$

Where c is the number of label categories after one-hot encoding, $\mathrm{c} = 5$ in this case. And y_i' is the model output of the data belonging to category i after sigmoid transformation; y is the true label (0 or 1) for data belonging to category i; The parameter γ is set to 2. $\mathrm{w_i}$ is the weight coefficient of each category. According to our experience, too large a gap between the weights tends to make the output result worse. Therefore the weight vector is set to $\mathbf{w} = [1.2, 1, 1, 1.2, 1]^{\mathrm{T}}$.

Model parameters were optimized using the ADAM optimizer. L2 normalization were used, with a weight decay parameter of 0.01. The initial learning rate was set to 0.0003 and decayed by a factor of 0.7 every 10 epochs. We set the training batch size to 16, and after propagating every 2 batches in the forward direction, the accumulated gradients are then used to update the weight parameters. This allows for a larger equivalent batch size with limited video memory.

As shown in Figure 2, when the 48th epoch of training is completed, the model was saved and used as a classifier for the murmur detection task. Then the weight vector of the loss function was adjusted to $\mathbf{w} = [1.2, 1, 1, 2, 1.5]^{\mathrm{T}}$. Another ten epochs are trained, and the final model is saved for the outcome classification task.

All these hyper-parameters are empirically optimized.

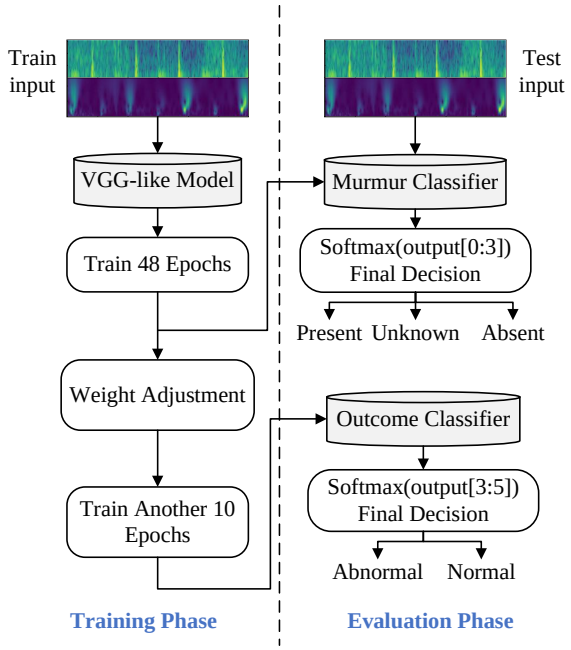

Figure 2. Training phase and evaluation phase of the classifier model. Note that the normalization process and demographic information of the features are omitted for simplicity.

2.4. Model Evaluation

The training phase and evaluation phase of the model is shown in Figure 2. The test PCGs are processed using the same method described in section 2.1 and the TF images are normalized using the mean and variance of the training set.

The data for each patient in test dataset were fed into the trained murmur model and the outcome model, and softmax operations were performed on the first three values and the last two values of the model output respectively. For the murmur task, it was judged as Absent only if its probability was greater than 0.6, otherwise it was judged as the category with higher probability among Unknown and Absent. For the outcome task, it was judged as Normal only if its probability was greater than 0.52. Both thresholds were empirically determined by changing the thresholds between 0 and 1 with a step of 0.01, using the weighted score and cost as criteria.

3. Result

For both tasks, we evaluated our classifier model on the public training set using repeated stratified 5-fold cross-validation.

Training	Validation	Test	Ranking
0.776 ± 0.026	0.737	0.752	11/40

Table 1. Weighted accuracy metric scores (official Challenge score) for our final selected entry (team PhysioDreamfly) for the murmur detection task, including the ranking of our team on the hidden test set. We used stratified 5-fold cross validation on the public training set, one-time scoring on the hidden validation set and hidden test set.

Training	Validation	Test	Ranking
10870 ± 250	9577	12831	18/39

Table 2. Cost metric scores (official Challenge score) for our final selected entry (team PhysioDreamfly) for the clinical outcome identification task, including the ranking of our team on the hidden test set. We used stratified 5-fold cross validation on the public training set, one-time scoring on the hidden validation set and hidden test set.

The results of the murmur task are shown in Table 1. And the results of the outcome task are shown in Table 2.

We also ran five-fold cross-validation experiments on the public training set using only one TF-domain representation and using a combination of two type of TF-domain representations as inputs. The results are displayed in Table 3. When only one TF-domain representation is used, its height (number of wavelet scales or Mel Banks) is set to 32 to ensure the consistency of the model's input dimensions.

	WS	LMS	Combined
Weighted Accuracy	0.734	0.760	0.776
Cost	11298	11103	10870

Table 3. Weighted accuracy metric scores and Cost metric scores for different input TF-domain representation. The results are all obtained using five-fold cross-validations on the public training set. Explanation of abbreviations: WS: Wavelet Scalogram; LMS: Log-Mel Spectrogram.

4. Discussion and Conclusions

We used two types of 2D representations of PCGs and a VGG-like network with multi-labels for murmurs and outcomes classification, and the results show that this is a competitive solution with a simple architecture.

A limitation of this work is that only the first 18 seconds of each recording were utilized. And the labels of the recordings for each location, which may helpful to improve the classification accuracy, were not fully utilized. Moreover, as a multi-label model, we simply concatenated the labels and did not take full advantage of the correlation between the two classes of labels. We also tried to combine the labels of the two tasks using the random k-labelsets method, but obtained poor results. According to our conjecture, it is because the amount of data in each category is too small after the labels are combined.

Acknowledgments

This work is supported by the National Key Research and Development Project 2020YFC1512304.

References

[1] Oliveira J, Renna F, Costa PD, Nogueira M, Oliveira C, Ferreira C, et al. The Circor DigiScope Dataset: From Murmur Detection to Murmur Classification. IEEE Journal of Biomedical and Health Informatics 2022;26(6):2524–2535. doi: 10.1109/JBHI.2021.3137048.

[2] Clifford GD, Liu C, Moody B, Springer D, Silva I, Li Q, et al. Classification of normal/abnormal heart sound recordings: The Physionet/Computing in Cardiology Challenge 2016. In 2016 Computing in cardiology conference (CinC). IEEE, 2016; 609–612.

[3] Karhade J, Dash S, Ghosh SK, Dash DK, Tripathy RK. Time-Frequency-Domain Deep Learning Framework for the Automated Detection of Heart Valve Disorders Using PCG Signals. IEEE Transactions on Instrumentation and Measurement 2022;71:1–11. doi: 10.1109/TIM.2022.3163156.

[4] Reyna MA, Kiarashi Y, Elola A, Oliveira J, Renna F, Gu A, et al. Heart Murmur Detection from Phonocardiogram Recordings: The George B. Moody PhysioNet Challenge 2022. medRxiv 2022;doi: 10.1101/2022.08.11.22278688.

[5] Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. Physiobank, Physiotoolkit, and Physionet: Components of a new research resource for complex physiologic signals. Circulation 2000;101(23):e215–e220. doi: 10.1161/01.CIR.101.23.e215.

[6] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv 2014;doi: 10.48550/arXiv.1409.1556.

[7] Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision. 2017; 2980–2988.

Address for correspondence:

Peng Wang
State Key Laboratory of Transducer Technology, Aerospace Information Research Institute, Chinese Academy of Sciences, No. 9, North Zhongguancun, Haidian District, Beijing, China.
wangpeng01@aircas.ac.cn