Heart Murmur Detection Using Ensemble of Deep Learning Classifiers for Phonocardiograms Recorded from Multiple Auscultation Locations

Saman Parvaneh, Zaniar Ardalan, Joomyung Song, Kathan Vyas, Cristhian Potes

Edwards Lifesciences, Irvine, USA

Abstract

A digital phonocardiogram (PCG) provides an opportunity for automated screening in resource-constrained environments. As part of the George B. Moody PhysioNet Challenge 2022, our team, Life_Is_Now, developed a computational approach using an ensemble of deep learning classifiers for identifying abnormal cardiac function from PCG. A stratified 5-fold cross-validation was used for model development and evaluation for murmur and clinical outcome identification. The backbone of our trained classifiers is a modified pre-trained deep convolutional neural network on AudioSet-Youtube corpus (YAMNet) and transfer learning. The YAMNet model is modified and finetuned on the publicly available PhysioNet dataset. Our murmur and clinical outcome classifiers received a weighted accuracy score of 0.831 and a Challenge cost score of 14,850 from cross-validation on the public training set. Our murmur scores were 0.678 and outcome score were 10,518 on the hidden validation set. However, we did not receive the official score for the hidden test set as our entry crashed in evaluation on the test set.

1. Introduction

Cardiovascular disease (CVD) is a major cause of mortality worldwide [1]. CVD covers a heterogeneous group of heart and vessel disorders, including heart valve disease (HVD) and congenital heart disease (CHD). Cardiac auscultation using a stethoscope is an accessible and common screening tool that can identify patients with heart murmurs for referral to a specialized doctor. However, accurate interpretation of phonocardiogram (PCG) requires training and long-term practice [2]. Digital PCG provides an opportunity for developing automated screening algorithms for heart sound analysis and diagnosis, especially in resource-constrained environments. The main goal of the 2022 George B. Moody PhysioNet Challenge is to explore the potential of an objective and automated algorithm for pre-screening of abnormal heart function [3, 4]. Deep learning models [5], traditional featured-based algorithms [6], and their combinations [7] are successfully used for PCG analysis. Deep learning advances allow to perform automated high-level feature extraction and classification with minimal signal pre-processing [8]. Our proposed method for the challenge uses ensemble of deep learning models for murmur detection and clinical outcome identification.

2. Materials and Methods

2.1. Data

The CirCor DigiScope dataset [9] provided 5,282 PCG recordings from four main auscultation locations for 1,568 patients. Auscultation locations are pulmonary valve (PV), aortic valve (AV), mitral valve (MV), tricuspid valve (TV), and other (Phc). Each recording is labeled manually by a human expert to identify whether a cardiac murmur is present, absent, or unknown at each auscultation location (location-based murmur label). If murmur is present at least in one location, the patient is annotated with present murmur label; if the annotator was unsure about the presence of a murmur in at least one location, the patient murmur label is annotated as "unknown" (patient-based murmur label). Note that PCGs for all locations are not available for all patients and, there are more than one PCG at a specific location for some patients. Besides murmur labels, clinical outcome labels (normal and abnormal) for each patient were included in the challenge dataset based on more comprehensive screening such as an echocardiogram interpretation [4]. Out of the CirCor dataset, 60% were publicly available as training set (942 patients and 3,163 recordings), 10% were put aside as hidden validation, and 30% as hidden test set. Please refer to [4, 9] for more details about the challenge and the associated data.

To have a more generalizable model, a stratified 5-fold cross-validation were used over both murmur and outcome labels. More specifically, we created a joint outcome-murmur label by combining both labels together and used stratified splitting over new label. Table 1 shows the details about the number of patients and PCG length at every fold. As can be seen, this length varies across patients. A further 3,454 PCGs from the PhysioNet/Computing in Cardiology Challenge 2016 (now called the George B. Moody PhysioNet Challenge) were used to create a complementary model for outcome

 ISSN: 2325-887X DOI: 10.22489/CinC.2022.241

identification [3, 10].

		Murmur			Outcome		Data Length Min-Max (Median)
		Absent	Present	Unknown	Normal	Abnormal	
Fold	0	139 (73.54%)	36 (19.05%)	14 (7.41%)	97 (51.32%)	92 (48.68%)	6.384-58.752 (21.184)
	1	140 (74.07%)	36 (19.05%)	13 (6.88%)	98 (51.85%)	91 (48.15%)	5.952-44.256 (21.728)
	2	140 (74.47%)	35 (18.62%)	13 (6.91%)	97 (51.6%)	91 (48.4%)	5.152-55.856 (21.488)
	3	138 (73.4%)	36 (19.15%)	14 (7.45%)	97 (51.6%)	91 (48.4%)	6.187-64.512 (21.472)
	4	138 (73.40%)	36 (19.15%)	14 (7.45%)	97 (51.6%)	91 (48.4%)	5.248-52.688 (21.648)
Total		695 (73.78%)	179 (19.0%)	68 (7.22%)	486 (51.59%)	456 (48.41%)	5.152-64.512

Table 1. The details of every fold's data using stratified 5-fold cross-validation. The number of patients for each label as well as minimum, maximum, and median length of the PCGs in seconds are included.

2.2. Methods

A pre-trained deep convolutional neural network (CNN) model, YAMNet, was utilized for transfer learning in this study [11]. This model takes audio waveform as input and makes independent predictions for 521 audio events. Audio events include heart and respiratory sounds (e.g., breathing, wheeze, heart murmur, and cough) as well as relevant noise sources (e.g., background noises, baby/infant cry, and human sounds) from Google AudioSet Ontology [12]. YAMNet uses the MobileNet v1 [13] architecture with a 0.98 seconds audio segment sampled at 16 kHz. Raw audios convert to Mel spectrogram with a window length of 25 milliseconds, an overlap of 15 milliseconds, and 64 frequency bands covering the range 125-7500 Hz. Therefore, input to YAMNet will be 96×64 Mel spectrogram images.

2.2.1 Pre-processing

For murmur detection, all available recording locations except Phc are combined with the associated location-based murmur label for training. Furthermore, the majority class (i.e., absent) was randomly undersampled by factor of four as it was identified in our experiments. For outcome identification, only PCGs recorded on AV location were used for training. To pre-process data for YAMNet, PCGs are resampled to 16 kHz and segmented with a window length of 0.98 seconds with 50% overlap.

2.2.2. Murmur Detection

For transfer learning of YAMNet in murmur detection, the last three layers (fully connected, softmax, and classification layers) were removed and replaced for 3-class classification (e.g., absent, presence, unknown). Of note, present, unknown, and absent class weights were set to 5, 3, and 1, respectively. The modified YAMNet was trained with Adam stochastic optimizer (mini-batch size of 32, shuffling of the training set every epoch, the initial learning rate of 0.0001, max epochs of 20, early stopping if the loss on the validation set is larger than or equal to the previously smallest loss for four times). During training phase five models were trained using training and validation folds listed in Table 2. In our proposed approach, every PCG recording is split into a window length of 0.98 seconds; therefore, there are multiple windows per recording. However, since the length of the PCG recordings is not the same for every patient and recording location, there is a different number of windows per recording. Each window is considered as a training data in the training phase, and the model produces multiple murmur labels (one label per window) per recording in the detection phase. Maximum voting among windows is used to produce only one output per recording. With only one output per recording, there are multiple auscultation recordings and five models, and another maximum voting is performed to find one output for every patient.

Model	Training Folds	Validation Fold	Test Fold
1	0,1,2	3	4
2	1,2,3	4	0
3	2,3,4	0	1
4	3,4,0	1	2
5	4,0,1	2	3

Table 2. Training folds as well as validation and test folds for developing five models for murmur detection.

2.2.3. Outcome Identification

For outcome identification, an ensemble of following three models is formed by maximum voting:

- For transfer learning of YAMNet in outcome identification, the last three layers were modified for 2-class classification (Normal and Abnormal). Then, modified YAMNet was trained using data from folds 0, 1, and 2. Folds 3 and 4 were used as validation and test folds, respectively. Adam stochastic optimizer with a mini-batch size of 16 and initial learning rate of 1e-7 was used for training for one epoch. Of note, the weights of all layers except the last three layers were frozen during training. Prevalence of normal and abnormal classes was used to set class weights in the last classification layer.
- A modified YAMNet for the 2-class classification explained above was trained with 80% and 20% of available PCGs from PhysioNet/Computing in Cardiology Challenge 2016 as a training and validation set. Class weights in the last classification layer were set based on the prevalence of normal and abnormal classes in the training data. Adam optimizer with the mini-batch size of 64, training data shuffling at every epoch, max epochs of 10, and an initial learning rate of 0.0001 was used for model training. Freezing the weights of the first 40 layers produced the best the results.

- The top performing pre-trained model in the Physionet/Computing in Cardiology Challenge 2016 by Potes et al. [7] was slightly changed as a third model for outcome identification. In this model, outputs of an AdaBoost-abstain classifier (AdaBoost) and a CNN classifier (CNN) were combined using the following rule to produce Normal and Abnormal, where the corresponding threshold (thr1 and thr2) in the original algorithm were tuned to maximize the 2016 challenge score:

```
if (AdaBoost > thr1) OR (CNN>thr2) then
        Abnormal PCG
Else                                    (1)
        Normal PCG
end if
```

We used the above thresholds to maximize the 2022 Challenge cost score in available data in folds 0, 1, and 2. In summary, the same pre-trained models with optimized thresholds based on 2022 challenge data (thr1= 0.28 and thr2= 0.47) is used to form the third model. Ensemble of three models on all available PCG recordings recorded in different locations for a specific patient using maximum voting will be used for outcome identification. All pre-processing and model development were performed using MATLAB R2022a.

2.3. Model Evaluation

There are two classifications tasks for murmur detection and outcome identification in this challenge. For the murmur classification, the loss function was a weighted average of accuracy where the weight of the present class is five, the weight of the unknown class is three, and the weight of the absent class is one. For the outcome classification, a Challenge cost score is provided by the PhysioNet challenge organizers. Please refer to [4] for details of evaluation metrics. The average and standard deviation of each model in 5-fold cross-validation are used to select the best models in our experiments for evaluation on hidden validation and test sets in the official phase of the challenge.

3. Results

Challenge scores for murmur detection and outcome identification in training and validation are reported in Tables 3 and 4. Our entry was not scored on the test set due to a crash in the evaluation code. Furthermore, average of F-measure and accuracy for present, unknown, and absent classes across 5-fold are shown in Figure 1.

4. Discussion and Conclusion

Several approaches, including transfer learning with YAMNet and VGGish [11], were explored to find the best model. VGGish is a CNN architecture that outputs a 128-dimensional feature vector for each PCG as an input.

Furthermore, YAMNet and VGGish were used as feature extractors across four auscultation recordings and fed to different classification configurations such as parallel Long Short-Term Memory (LSTM). In parallel LSTM, one model is trained for each PCG location and then all outputs of parallel channels are merged to produce one representation for every patient. Then the merged output is trained using three dense layers and a final classification layer. We evaluated multiple optimizers, batch size, number of units at each channel, and regularization methods for all our experiments. Our best performing model in 5-fold cross-validation was the transfer learning using modified YAMNet. One challenge with parallel LSTM was missing auscultation location for some patients. To address this challenge, the missing PCG locations were replaced by one of the available auscultation locations or with zero vectors. Another challenge was differences in the length of PCG recordings across patients where we needed to crop long recordings with the length of the shortest recording, which led to information loss.

Through our exploration, YAMNet performed better than VGGish in all explored configurations. We believe a smaller number of trainable parameters in YAMNet compared to VGGish makes it a better option for settings with limited training data. Also, the smaller number of trainable parameters in YAMNet makes training faster than VGGish. In contrast to parallel LSTM, where all PCG locations for a specific patient are fed as one training data, different auscultation recordings were considered independent training data in our best-performing model for murmur detection. Therefore, more data was available to train a model with higher performance. This observation is consistent with training independent modified YAMNet models for each recording location. In our exploration, we found outcome identification using deep learning more challenging than murmur detection. A modified YAMNet trained on the Physionet/Computing in Cardiology Challenge 2016 showed a proper convergence in the loss function and promising Challenge cost score in training and validation on 2016 challenge data. However, such convergence was not present when we trained the same model on available data for 2022 challenge. A better understanding of the process for labeling clinical outcomes in the 2022 challenge dataset might help to better explain this observation. After listening to PCGs in the training data, we noticed noisy segments, including ambient noise (e.g., baby cries and human voice) and friction/abrasion between the recording device and the skin/chest. Filtering unwanted noise can potentially enhance murmur detection and outcome identification. Therefore, we plan to explore the impact of denoising of recordings on model performance. Also, data augmentation for the proposed classification tasks considering a limited number of

training data needs to be explored.

This article used an ensemble of deep learning models for murmur detection and clinical outcome identification using PCG. The promising results of the internal cross-validation on the public training and hidden validation indicate the potential of deep learning models for automated PPG analysis for murmur screening in resource-constrained environments. However, screening algorithm for clinical outcome identification needs to be improved.

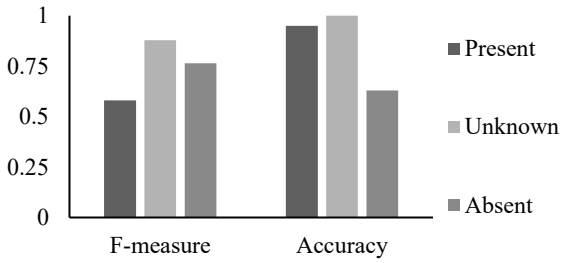

Figure 1. Average of F-measure and accuracy for present, unknown, and absent classes across 5-folds.

Training	Validation	Test	Ranking
0.831±0.022	0.678	NS	NS

Table 3. Weighted accuracy metric scores (official Challenge score) for our final selected entry (team Life_Is_Now) for the murmur detection task, including the ranking of our team on the hidden validation set. We used 5-fold cross validation on the public training set and repeated scoring on the hidden validation set. NS for one-time scoring on the hidden test set and ranking indicates our entry was not officially scored because of a crash in our evaluation code.

Training	Validation	Test	Ranking
14,850±16.73	10,518	NS	NS

Table 4. Cost metric scores (official Challenge score) for our final selected entry (team Life_Is_Now) for the clinical outcome identification task, including the ranking of our team on the hidden validation set. We used 5-fold cross validation on the public training set and repeated scoring on the hidden validation set. NS for one-time scoring on the hidden test set and ranking indicates our entry was not officially scored because of a crash in our evaluation code.

References

[1] Roth GA, Johnson C, Abajobir A, Abd-Allah F, Abera SF, Abyu G *et al.* Global, Regional, And National Burden of Cardiovascular Diseases For 10 Causes, 1990 to 2015. Journal of the American College of Cardiology 2017; 70 (1):1-25.

[2] Jiang Z, and Choi S A Cardiac Sound Characteristic Waveform Method For In-home Heart Disorder Monitoring With Electric Stethoscope. Expert Systems with Applications 2006; 31 (2):286-298.

[3] 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.

[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; https://doi.org/10.1101/2022.08.11.22278688.

[5] Chen W, Sun Q, Chen X, Xie G, Wu H, and Xu C Deep Learning Methods For Heart Sounds Classification: A Systematic Review. Entropy 2021; 23 (6):667.

[6] Zabihi M, Rad AB, Kiranyaz S, Gabbouj M, and Katsaggelos AK, Heart Sound Anomaly And Quality Detection Using Ensemble of Neural Ntworks Without Segmentation, in 2016 Computing In Cardiology Conference (CinC), 2016, pp. 613-616.

[7] Potes C, Parvaneh S, Rahman A, and Conroy B, Ensemble of Feature-based And Deep Learning-based Classifiers For Detection Of Abnormal Heart Sounds, in 2016 Computing In Cardiology Conference (CinC), 2016, pp. 621-624.

[8] Parvaneh S, Rubin J, Babaeizadeh S, and Xu-Wilson M Cardiac Arrhythmia Detection Using Deep Learning: A Review. Journal of Electrocardiology 2019; 57 S70-S74.

[9] 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 2021; 26 (6):2524-2535.

[10] Liu C, Springer D, Li Q, Moody B, Juan RA, Chorro FJ *et al.* An Open Access Database For The Evaluation Of Heart Sound Algorithms. Physiological Measurement 2016; 37 (12):2181.

[11] Hershey S, Chaudhuri S, Ellis DP, Gemmeke JF, Jansen A, Moore RC *et al.*, CNN Architectures For Large-scale Audio Classification, in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 131-135.

[12] Gemmeke JF, Ellis DP, Freedman D, Jansen A, Lawrence W, Moore RC *et al.*, Audio set: An Ontology And Human-labeled Dataset For Audio Events, in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 776-780.

[13] Sinha D, and El-Sharkawy M, Thin Mobilenet: An Enhanced Mobilenet Architecture, in 2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 2019, pp. 0280-0285.

Address for correspondence:
Saman Parvaneh
1 Edwards Way, Irvine, CA, USA
parvaneh@ieee.org