Multi-Task Prediction of Murmur and Outcome from Heart Sound Recordings

Yale Chang[1], Luoluo Liu[1], Corneliu Antonescu[2,3]

[1] Philips Research North America, Cambridge, MA, United States
[2] Banner Health, Phoenix, AZ, United States
[3] University of Arizona College Medicine, Phoenix, AZ, United States

Abstract

As part of the George B. Moody PhysioNet Challenge 2022, our team, prna, developed a novel approach to automatically detect the presence of heart murmurs and predict abornomal clinical outcomes from the heart sound recordings. To train the model, each heart sound recording is first divided into multiple 3-second segments. Then each segment is transformed into a time-domain embedding vector through a convolutional neural network (CNN). In parallel, the Mel-frequency cepstrum (MFCC) representation of the segment is transformed into a frequency-domain embedding vector using CNN. These embedding vectors and the demographic variables are concatenated and then used as input to two separate networks built to predict the presence of heart murmurs and clinical outcomes respectively. The network parameters are optimized to jointly predict both targets using multi-task learning. Our murmur detection classifier received a weighted accuracy score of 0.694 (ranked 20th out of 40 teams) and Challenge cost score of 11,403 (ranked 2nd out of 39 teams) on the hidden test set.

1. Introduction

The objective of the George B. Moody PhysioNet Challenge 2022 is to detect the presence of heart murmurs and predict abnormal clinical outcomes from heart sound recordings collected from pediatric subjects[1]. Since heart sound recordings can be obtained from cardiac auscultations in a non-invasive manner, a machine learning model trained to automatically detect the presence of heart murmurs and predict abnormal clinical outcomes from heart sound recordings can facilitate the identification of children with congenital or acquired heart diseases.

Traditional heart sound classification algorithms [2] consist of 1) preprocessing (noise removal) ; 2) segmentation (S1, systolic, S2, diastolic states); and 3) classification using machine learning algorithms (support vector machine, k-nearest neighbors, multiple layer perceptron, etc). Recently, deep learning approaches [3] are shown to achieve improved performance compared to traditional approaches. Therefore, we chose to build deep learning models for this problem. Furthermore, considering the strong association between the presence of murmur and the abnormal outcome, we applied multi-task learning [4] to jointly predict the murmur presence and abnormal outcome. We compared performance with models built to predict those targets separately. Besides presenting the results of the proposed model, we also ran ablation study to demonstrate the effects of adding MFCC features and multi-task prediction.

2. Methods

The public training data consists of 3163 heart sound recordings collected from 942 pediatric subjects [5]. Most of the patients have multiple recordings from multiple auscultation locations (for example Mitral, Aortic, Pulmonary or Triecuspid Valves). The murmur label (*Present, Unknown, Absent*) is assigned to each recording and a subject is labeled as murmur present if any recording of the subject contains murmur. The clinical outcome label (*Abnormal, Normal*) is assigned to each subject. The trained model is expected to output predicted probabilities and labels for both targets (presence of the murmur and abnormal outcome) at the subject-level.

2.1. Preprocessing

First, each recording was downsampled from 4000 Hz to 1000 Hz. To remove the noise in the recording, we applied an order-2 Butterworth filter [6] with frequency bandpass of 25 Hz to 400 Hz. Following z-normalization applied to make each recording have zero mean and unit standard deviation, each recording was dividied into multiple consecutive non-overlapping 3-second segments and each segment was labeled using the murmur label of the recording and outcome label of the subject. Here we assumed both the murmur signal and outcome signal will be present at each segment of the recording [7].

Besides using the time series of each segment as model input, we also computed its Mel-frequency cepstrum

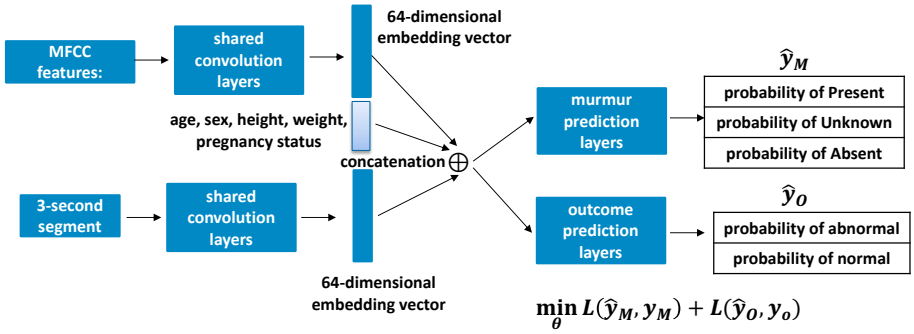

Figure 1. The network architecture consists of 1) the representation learning layers transforming the inputs into embedding vectors; and 2) the target prediction layers transforming the concatenattion of the embedding vectors and demographic variables into the predicted probabilities of both murmur and outcome.

(MFCC) [8] representation and use it as model input, which can provide complementary information besides the temporal representation. We summarize the detailed settings of the MFCC transformation in Table 1.

Parameter	Value
sample rate	1000 Hz
analysis window length	0.025 second
steps between successive windows	0.1 second
number of cepstrum	10
# filters in the filterbank	26
FFT size	512
lowest band ege of Mel filters	0 Hz
highest band edge of Mel filters	500 Hz

Table 1. Parameter settings of the MFCC transformation

	Parameter	Value
Input	dim(segment) dim(MFCC) dim(demographic)	3000 2990 5
shared convolution layers	# layers # kernels kernel size	2 20 15
embedding vector	dimension	64
prediction layers	# layers # kernels kernel size	2 20 3

Table 2. Parameter settings of the network architecture

2.2. Network Architecture

The 3-second segment and MFCC features were used as inputs to the network architecture shown in Figure 1. The 3-second segment and MFCC features were transformed into two embedding vectors using two separate CNN networks. The embedding vectors were then concatenated with the demographic features and used as input to two separate CNN networks (murmur prediction layers and outcome prediction layers) built to predict the murmur presence and abnormal outcomes. We summarize the network parameters in Table 2.

2.3. Model Training

We split the 942 patients' data in the public training set into five folds, where the first three folds were used to train the model, the fourth fold and the fifth fold were used for *in-house validation* and *in-house test* respectively. After dividing each recording into multiple 3-second segments, there are 13,169 samples used for model training, 4,822 samples used for in-house validation and 4,474 samples used for in-house testing.

Note that the *in-house validation set* and *in-house test set* are subsets from the *public training set* and therefore different from the *hidden validation set* (used in the official phase of the Challenge) and the *hidden test set* (used to get the final rankings of all teams).

We will report model performances on the *hidden test set* to compare with other teams in a fair manner. To avoid making too many submissions in the official phase, we used the *in-house test set* for ablation study to identify the best model architecture. Therefore, we will also report the performance of different models on the *in-house test set* for the ablation study on how MFCC features and multi-task learning affect the model performance.

The network parameters θ were optimized to minimize the summation of the cross entropy loss of the murmur prediction ($\mathcal{L}(\hat{\mathcal{Y}}_\mathcal{M}, \mathcal{Y}_\mathcal{M})$) and outcome prediction ($\mathcal{L}(\hat{\mathcal{Y}}_\mathcal{O}, \mathcal{Y}_\mathcal{O})$) on the training data. Both the network architecture and the loss function were implemented in PyTorch

[9].

$$\min_{\theta} \mathcal{L}(\hat{\mathcal{Y}}_{\mathcal{M}}, \mathcal{Y}_{\mathcal{M}}) + \mathcal{L}(\hat{\mathcal{Y}}_{\mathcal{O}}, \mathcal{Y}_{\mathcal{O}}) \quad (1)$$

We chose the Adam optimizer [10] with initial learning rate of 0.001 and weight decay of 0.001. The learning rate was decreased to its one tenth if the validation metric was lower than all validation metric values of the previous three epochs. The optimization will terminate if the learning rate fell below 1e-6 or the number of epochs reached 50.

The average area under the curve (AUC) value of murmur prediction and outcome prediction computed on the validation set was used to 1) select the best network parameters given a fixed architecture; and 2) optimize the network architecture, including network depth, the number of kernels and kernel size.

The in-house validation set was also used to select 1) the optimal threshold of murmur prediction to maximize the weighted accuracy; and 2) the optimal threshold of outcome prediction to minimize the cost metric defined by the Challenge organizer [1].

2.4. Model Prediction

When applying the trained model to the test subject, we apply the same preprocessing steps to obtain multiple 3-second segments and its MFCC features from multiple recordings. The predicted probabilities for segments from the same recording are averaged to derive the recording-level prediction. Given multiple recording-level predictions for the same subject, the subject is predicted to have murmur (or abnormal outcome) if any of the recording is predicted to have murmur (or patient is labeled to have an abnormal outcome).

3. Results

3.1. Evaluation on Official Test Set

After training the model to jointly predict murmur and outcome using both 3-second heart sound recording segments and the MFCC features as input, the murmur weighted accuracy (higher is better) and outcome cost (lower is better) on the public training, hidden validation, and hidden test set are shown in Table 3 and Table 4 respectively. Compared to other teams, the ranking of outcome prediction (2nd place) is much higher than murmur prediction (20th place).

3.2. Ablation Study Using Public Training Set

We ran ablation study to show the effects of 1) adding MFCC features to the inputs; and 2) multi-task prediction of murmur and outcome, on the model performance.

Training	Validation	Test	Ranking
0.749	0.696	0.694	20/40

Table 3. Weighted accuracy scores for our final selected entry (team prna) for the murmur detection task, including the ranking of our team on the hidden test set. We used one in-house train/vadliation/test split on the public training set, repeated scoring on the hidden validation set, and one-time scoring on the hidden test set.

Training	Validation	Test	Ranking
12,302	9,688	11,403	2/39

Table 4. Cost scores for our final selected entry (team prna) for the clinical outcome identification task, including the ranking of our team on the hidden test set. We used one in-house train/vadliation/test split on the public training set, repeated scoring on the hidden validation set, and one-time scoring on the hidden test set.

First, we evaluate the model with the same architecture (PCG and MFCC as input to jointly predict murmur and outcome) on the in-house test set, which is a subset of public training set, and show the results in Table 5.

Input	PCG AND MFCC
Target	murmur AND outcome
in-house test set: murmur weighted accuracy	0.749
in-house test set: outcome cost	12302.800

Table 5. The murmur weighted accuracy (higher is better) and outcome cost (lower is better) on the in-house test set when using both 3-second heart sound recording segments and MFCC features to jointly predict murmur and outcome

3.2.0.a **Effect of MFCC Features** After removing the MFCC features from the model input, the model performance is shown in Table 6. The performance of both murmur prediction and outcome prediction decreases, indicating MFCC features do provide complementary information to the heart sound recording time series in predicting both targets.

3.2.0.b **Effects of Multi-Task Prediction of Murmur and Outcome** In addition to jointly predicting both murmur and outcome, we also trained two independent models to predict murmur and outcome separately. The model performance is shown in Table 7. Compared to the multi-task training, the performance of the independent murmur prediction model becomes worse (weighted accuracy decreases from 0.749 to 0.712). In contrast, the performance of the independent outcome prediction model becomes

Input	PCG **Only**
Target	murmur AND outcome
in-house test set: murmur weighted accuracy	0.723 (↓)
in-house test set: outcome cost	12500.724 (↓)

Table 6. Ablation study on the effects of MFCC features: the murmur weighted accuracy (higher is better) and outcome cost (lower is better) on the in-house test set when only using 3-second heart sound recording segments (without MFCC features) to jointly predict murmur and outcome

better (cost decreases from 12302.800 to 12065.997). This seems to indicate the multi-task prediction improves the performance of murmur prediction at the expense of outcome prediction.

Input	PCG AND MFCC
Target	murmur **OR** outcome
in-house test set: murmur weighted accuracy	0.712 (↓)
in-house test set: outcome cost	12065.997 (↑)

Table 7. Ablation study on the effects of multi-task prediction of murmur and outcome: the murmur weighted accuracy (higher is better) and outcome cost (lower is better) on the in-house test set when using both 3-second heart sound recording segments and MFCC features to predict murmur and outcome separately

4. Conclusions

In this work, we develop a multi-task learning model to jointly predict the murmur and outcome from heart sound recordings. The shared convolution layers enable the extraction of relevant embedding vectors from the heart sound recordings represented as raw time series in the time domain and MFCC features in the frequency domain. Ablation study of removing MFCC features demonstrates MFCC features can provide complementary information to the time series data in predicting both murmur and outcome. The multi-task prediction of murmur and outcome is shown to lead to better performance of murmur prediction at the expense of outcome prediction.

The model can be further improved by adding more traditional hand-enginerred features to the inputs of prediction layers. The potential benefits include improved model performance and robustness to data distribution shift. Model interpretation is also needed to highlight how the model make predictions. This is important in understanding the failure mood of the model.

References

[1] Reyna MA, Kiarashi Y, Elola A, Oliveira J, Renna F, Gu A, Perez-Alday EA, Sadr N, Sharma A, Mattos S, Coimbra MT, Sameni R, Rad AB, Clifford GD. Heart murmur detection from phonocardiogram recordings: The George B. Moody PhysioNet Challenge 2022. medRxiv 2022;URL https://doi.org/10.1101/2022.08.11.22278688.

[2] Liu C, Springer D, Li Q, Moody B, Juan RA, Chorro FJ, Castells F, Roig JM, Silva I, Johnson AE, et al. An open access database for the evaluation of heart sound algorithms. Physiological Measurement 2016;37(12):2181.

[3] Chen W, Sun Q, Chen X, Xie G, Wu H, Xu C. Deep learning methods for heart sounds classification: a systematic review. Entropy 2021;23(6):667.

[4] Caruana R. Multitask learning. Machine Learning 1997; 28(1):41–75.

[5] Oliveira J, Renna F, Costa PD, Nogueira M, Oliveira C, Ferreira C, Jorge A, Mattos S, Hatem T, Tavares T, Elola A, Rad AB, Sameni R, Clifford GD, Coimbra MT. The CirCor DigiScope dataset: from murmur detection to murmur classification. IEEE Journal of Biomedical and Health Informatics 2021;26(6):2524–2535.

[6] Butterworth S, et al. On the theory of filter amplifiers. Wireless Engineer 1930;7(6):536–541.

[7] Xiao B, Xu Y, Bi X, Li W, Ma Z, Zhang J, Ma X. Follow the sound of childrens heart: a deep-learning-based computer-aided pediatric chds diagnosis system. IEEE Internet of Things Journal 2019;7(3):1994–2004.

[8] Mermelstein P. Distance measures for speech recognition, psychological and instrumental. Pattern Recognition and Artificial Intelligence 1976;116:374–388.

[9] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, et al. Pytorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems 2019;32.

[10] Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv Preprint arXiv14126980 2014;.

Address for correspondence:

Yale Chang
222 Jacobs St, Cambridge, MA 02141
yale.chang@philips.com