Maiby's Algorithm: A Two-Stage Deep Learning Approach for Murmur Detection in Mel Spectrograms for Automatic Auscultation of Congenital Heart Disease

Matheus Araujo[1] , Dewen Zeng[2], Joao Palotti[3], Xirong Xi[2] , Yiyu Shi[2], Lee Pyles[4], Quan Ni[5]
[1]Cleveland Clinic Foundation, Cleveland, United States
[2] University of Notre Dame, South Bend, United States
[3] Qatar Computing Research Institute, Doha, Qatar
[4]West Virginia University, Morgantown, United States
[5]One Heart Health, Minneapolis, United States

Abstract

Congenital heart disease (CHD) is a major cause of death for newborns, especially in low resources countries, due to limited access to heart specialists for timely diagnosis. As part of the George B. Moody PhysioNet Challenge 2022, we propose an automatic algorithm to detect CHD murmurs from digitally recorded heart sounds annotated by specialists. To train and validate our model, we use a dataset with 5282 heart sounds collected from 1568 children in the Paraiba state of Brazil recorded from multiple auscultation locations. Our team, named One_Heart_Health, used a two-stage strategy that combines embeddings from Mel spectrograms generated from audio segments and a final classifier that combine those embeddings to deliver the final classification per individual. On the official hidden test, we reached a weighted accuracy score of 0.729 (ranked 17th out of 40) and a challenge cost score of 13283 (ranked 23th out of 39). In our internal 5-fold cross-validation experiments, our approach reached a sensitivity of 0.76 ± 0.10 and a specificity of 0.85 ± 0.11. We have shown that a deep learning approach for murmur detection has the potential to mimic heart specialists to provide timely identification of CHD.

1. Introduction

Congenital heart disease (CHD) affects about 1% of newborns, causing approximately 260,000 death per year in 2017, with 87% of them from the low- or middle-income countries [1]. While universal newborn screening for critical CHD has been adopted in high-income countries, many low resources parts of the world continue to struggle with timely diagnosis, especially in geographically remote areas with limited access to heart specialists. We have previously shown that telemedicine by heart specialists can accurately detect CHD with an overall accuracy of 91% based on digital heart sounds [2]. A promising next step to scale auscultation in consistency and reachability while maintaining affordability is to develop automatic detection of CHD with a machine learning model. The model should leverage a large dataset of digital heart sounds and be adjudicated by heart specialists or verified by echocardiographic diagnoses. In this context, the 2022 George B Moody PhysioNet Challenge enables the use of machine learning by providing a dataset where we can identify murmurs and clinical outcomes associated with CHD from digital heart sounds collected from a large-scale CHD screening program with digital heart sounds of 1568 children participants and respective expert annotations [3].

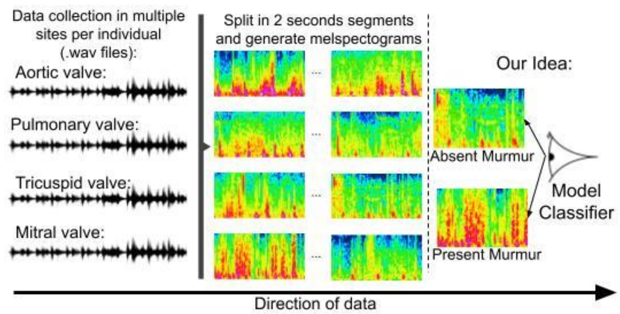

Figure 1. Mel spectrograms from 2-second audio segments with absent and present murmurs. The x-axis in each figure is time, the y-axis is frequency, and reddish colors represent high intensity. We propose a deep learning approach based on visual perception to assist in auscultation for congenital heart disease. From left to right, we show how the information flows from data multi-site heart sound collection to the final classification of an individual.

In this work, we leverage a large dataset of labeled heart sounds to generate an auscultation-like analysis regarding the presence or absence of murmurs. Figure 1 shows our

 ISSN: 2325-887X DOI: 10.22489/CinC.2022.249

core idea of using Mel spectrograms. It is evident to a trained human visual cortex that the top Mel spectrogram has clear-spaced and intense low-frequency events on the bottom region, an acoustic behavior that we expect from normal heartbeats. The Mel spectrogram on the bottom part of Figure 1 has additional events that are regular and longer on mid-level frequency bands. When heard by a specialist, these events can be identified as murmurs. With our two-stage deep learning framework, we aim to mimic human visual cortex classification and aggregated the information for final decision.

1.1. Related Work

Heart murmurs have been a key signal for CHD diagnosis and the use of automatic computation to detect them has been investigated for decades. We can specify two main reasons for automation: mass screening, especially in places with deficient health care systems and the physical limitations, including subjectivity judgment of a high-skilled examiner trained for many years [4] .

The advances in applied artificial intelligence in the last decade contributed to the performance of computer models to detect pathological heart murmurs with performance similar to expert cardiologists [5]. Previously, the 2016 PhysioNet Challenge [6] also promoted teams to find the best algorithm for heart sound classification for general heart diseases; but our work in PhysioNet Challenge 2022, specifically focused on murmur detection of CHD.

2. Methodology

We propose a two-stage strategy. After data collection and audio preprocessing, in Stage 1, we deploy a CNN-based neural network on top of Mel spectrogram segments to classify whether a 2-second audio piece contains or not a murmur. Then, in Stage 2, we use the model in Stage 1 as a feature extractor, and from randomly selected pieces of participants audios, we develop another model that classifies the participant as normal or abnormal.

2.1. Data

The dataset used in this study was provided during the official phase of the 2022 PhysioNet Challenge [7]. It is a subset of the data collected by two independent cardiac screening campaigns organized to screen a large pediatric population in the Northeast region of Brazil [3]. We used heart sound recordings ranging from 8 to 312.5 seconds from 1,568 participants. These were recorded using a Littmann 3200 stethoscope embedded with the DigiScope

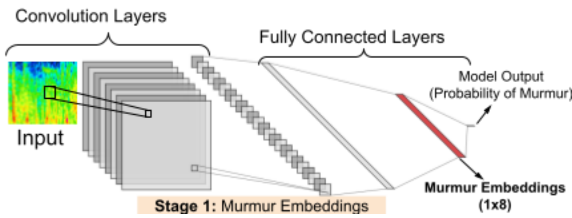

Figure 2: In Stage 1, we train a CNN model that classifies whether or not a Mel spectrogram contains a murmur. We flatten the output of the convolution layers and process it with two fully connected layers before making a prediction. In Stage 2, we extract features for another classifier by using the second last layer with dimension 1x8.

Collector at 4 kHz. Two cardiac physiologists manually annotated the beginning and end of each fundamental heart sound. The recording locations included aortic, pulmonary, tricuspid, and mitral on healthy (normal) and pathological (abnormal) subjects, with various congenital heart diseases, a single individual was associated with multiple audio recordings from one or more locations, and each audio was classified on whether a murmur was present, absent, or unknown. While preprocessing the files, we ignored recordings of locations without a murmur from participants that had the murmur identified in another audio. To increase the sensitivity of our algorithm, we considered all unknown cases as cases of present murmur.

2.2. Audio Preprocessing

We first processed the audio files (.wav) of all participants passing a Butterworth Bandpass filter allowing frequency between 20hz and 530hz; most pathogenic murmurs were found to be between these bands' thresholds. We also normalized the audio samples using the Librosa python package. Next, since each recording has long audio files of different lengths, we split recording into multiple segments of 2 seconds. We chose 2 seconds to guarantee at least one entire heartbeat cycle in each segment. Finally, we generate a Mel spectrogram from each 2-second segment, similar to those shown in Figure 1. Aiming to mimic human-like performance, we chose to work with Mel spectrograms because they are audio representations constructed after applying Mel filters crafted to enhance frequencies more distinguishable by the human auditory system.

2.3. Stage 1: Murmur Detection Model

Figure 2 shows the first stage of Maiby's algorithm. We build a convolutional neural network (CNN) model that

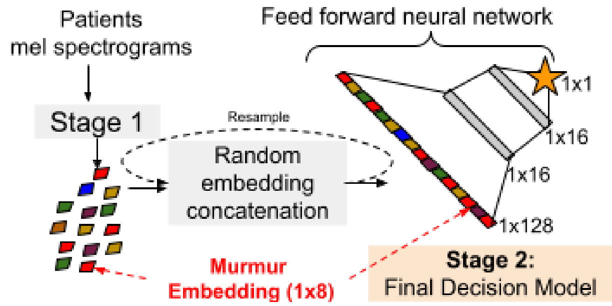

Figure 3: For Stage 2, we train a feed-forward neural network that outputs the probability of a participant presenting or not murmur. Each participant has multiple embeddings generated by each of the 2-second Mel spectrograms. We randomly concatenate 16 of these embeddings using the second last layer of Stage 1 to generate an input tensor of 1x128 for the Stage 2 model. Interestingly, we can resample embeddings to generate new inputs many times for the same participant since the order of concatenation does not matter.

receives as input a Mel spectrogram generated during preprocessing and outputs in its last layer a single value from a sigmoid function that can be interpreted as the probability of the segment having a murmur.

For weight initialization, we leverage the 2016 PhysioNet Challenge dataset to "warm-up" our Stage 1, but without freezing any layers before training with the 2022 PhysioNet data.

Since the Stage 1 model only works on a single Mel spectrogram, we need to devise a strategy to combine the various spectrograms generated for a participant. Our approach for this problem is to take advantage of the second last layer of the network with an embedding size of 1x8 and combine them to form the input of our second model to generate the final prediction.

2.4. Stage 2: Final Decision Model

As shown in Figure 3, from all audio recordings of a patient, we randomly sampled 16 2-second audio clips and input them into our Stage 1 network to generate the embeddings of size 1x8. Next, we randomly concatenated 16 embeddings to generate an array of size 1x128 to serve as input of Stage 2 that will decide whether a participant should or not be further screened. Our Stage 2 model is a simple feed-forward neural network made of fully connected layers with an input size of 1x128, two layers of shape 1x16, and a final layer of shape 1x1 that outputs after a sigmoid activation the probability of a participant having a murmur. Due to the unbalanced nature of the problem, we also set the class weight according to the estimated value from the scikit-learn Python package.

We do not expect any time dependency between the murmur embeddings since they are randomly selected. Also, an interesting property, we can repeat the random concatenation multiple times for the same patient while training leveraging all the embeddings from the annotated data from a participant. In our experiments, we generate 4 arrays of 1x128 for each patient during training.

For both Stage 1 and Stage 2 models, we set to 1000 the maximum epochs to train the models. We used the area under the precision-recall curve (AUC-PR) as the metric for early stopping with the patience parameter set to 40 epochs (i.e., the number of epochs without improvement until halt training). We trained this network using an Adam optimizer with a binary cross-entropy loss function. TensorFlow 2.8.2 was used to implement the models. We named our approach Maiby's algorithm and our code is available at: https://github.com/maraujo/physionet22/.

For the cost of the clinical outcome identification task in the challenge, we used a vanilla feed-forward neural network with 2 layers, each with 8 neurons and a 25% dropout rate. The input was a standardized vector with participants' demographic information (pregnancy status, weight, height, body mass index, gender, and age) concatenated with the murmur probability computed by the Stage 2 model.

3. Results

The weighted accuracy is the official challenge score for murmur detection; it is computed by using weights 5, 3, and 1 for the accuracies of the present, unknown and absent classes, respectively. This metric places more importance or weight on participants with murmurs. First, we analysed our model in the hidden validation set (10% of the dataset) from the Physionet challenge. For comparison, we used a baseline model implemented in Python with a random forest classifier using participants' demographic data and the mean, variance, and kurtosis of each recording. Our weighted accuracy performance reached 0.699 against 0.394 for the baseline.

In the official hidden test set (30% of the dataset), our model reached for the murmur detection task a weighted accuracy performance of 0.729 (17th out of 40), a score 6.5% lower than the best-ranked team (0.780). For the cost of clinical outcome, we reached 13283 (ranked 23th out of 39).

Table 1 shows an additional analysis of Maiby's algorithm performance based on a 5-fold cross-validation experiment of the training data (283 participants used for the test). For the area under the receiver operating characteristic curve (AUC-ROC), the mean and confidence

AUC-ROC	Weighted Accuracy	True Neg.	False Pos.	False Neg.	True Pos.	Sens.	Spec.	Fold
0.70	0.72	193	37	13	40	0.76	0.84	1
0.70	0.70	155	67	12	49	0.80	0.70	2
0.78	0.79	201	18	9	55	0.86	0.92	3
0.73	0.67	209	21	18	35	0.66	0.91	4
0.71	0.69	192	31	17	43	0.72	0.86	5

Table 1. Our 5-fold-cross-validation results using the training

interval was 0.72 ± 0.04; for sensitivity, we had 0.76 ± 0.10 and specificity 0.85 ± 0.11. We observed that our performance for weighted accuracy in the hidden test set is within the confidence interval from our experiment (0.71 ± 0.05).

4. Discussion and Conclusion

The most complex challenge of Maiby's algorithm approach was hyperparameter tuning. We have a large number of parameters, each one with ample space to be evaluated. This problem ranges from the split proportion of the dataset, the size of the murmur embeddings (1x8), the number of embeddings as input for Stage 2 (16 embeddings), number of embeddings resamples per participant (4). In addition, we have to find the best deep learning architecture, which includes specific parameters such as the number of layers, the number of neurons in each layer, the type of activation function, use or not dropout layers, and normalize or not input. Due to limited computing resources and the long training time (3h49min), we had a minimal number of randomized parameter search.

The stochastic concatenation of embeddings in Stage 2 is another weakness. Consider the case of sampling a patient with more than 16 embeddings, meaning that some will be left unseen. The final classification will be impacted if the embeddings containing murmur are unseen.

5. Conclusion

The George B. Moody PhysioNet Challenge 2022 provides a large dataset that motivates the creation solutions for the automatic auscultation of congenital heart disease. We proposed a two-stage algorithm using a convolutional neural network and another feed-forward neural network that, when combined, outputs a participant's probability of having a present heart murmur in their collected heart sounds. Our approach was only 6.5% worse than the best team in terms of weighted accuracy performance for the murmur classification task. We acknowledge that the current algorithm is not as accurate as experts in detecting congenital heart disease. Still, we demonstrate that automatic auscultation of congenital heart disease is possible and has incredible potential in low resources areas to improve CHD diagnosis.

Acknowledgments

This work was made with support from One Heart Health, a non-profit organization focused on developing medical technologies in low resources parts of the world (onehearthealth.org). We dedicated this work to Maiby Maria Araujo, who lost her life to heart disease complications.

References

1. Zimmerman MS, Smith Agc, Sable CA, Echko MM, Wilner LB, Olsen He, et al. Global, Regional, and National Burden of Congenital Heart Disease, 1990–2017: A Systematic Analysis for the Global Burden of Disease Study 2017. The Lancet Child & Adolescent Health. 2020;4(3):185-200.
2. Pyles L, Hemmati P, Pan J, YU X, LIU K, Wang J, et al. Initial Field Test of a Cloud-Based Cardiac Auscultation System to Determine Murmur Etiology in Rural China. Pediatric Cardiology. 2017;38(4):656-62.
3. 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):2524-35.
4. Delgado-Trejos E, Quiceno-Manrique AF, Godino-Llorente Ji, Blanco-Velasco M, Castellanos-Dominguez G. Digital Auscultation Analysis for Heart Murmur Detection. Annals of Biomedical Engineering. 2009;37(2):337-53.
5. Chorba JS, Shapiro Am, Le L, Maidens J, Prince J, Pham S, et al. Deep Learning Algorithm for Automated Cardiac Murmur Detection via a Digital Stethoscope Platform. Journal of the American Heart Association. 2021;10(9).
6. Clifford GD, LIU C, Moody B, Millet J, Schmidt S, LI Q, et al. Recent Advances in Heart Sound Analysis. Physiological Measurement. 2017;38(8):E10-E25.
7. Reyna, M. A., Kiarashi, Y., Elola, A., Oliveira, J., Renna, F., GU, A., Perez-Alday, E. A., Sadr, N., Sharma, A., Mattos, S., Coimbra, M. T., Sameni, R., Rad, A. B., Clifford, G. D. (2022). Heart Murmur Detection From Phonocardiogram Recordings: The George B. Moody Physionet Challenge 2022.Medrxiv, Doi: 10.1101/2022.08.11.22278688

Address for correspondence:
Matheus Araujo
9500 Euclid Avenue, Sleep Disorder Center, Cleveland, Ohio, 44195
limadim@ccf.org