Searching for Effective Neural Network Architectures for Heart Murmur Detection from Phonocardiogram

Hao Wen[1], Jingsu Kang[2]

[1]LMIB and School of Mathematical Sciences, Beihang University, Beijing, China
[2]Tianjin Medical University, Tianjin, China

Abstract

Aim: The George B. Moody PhysioNet Challenge 2022 raised problems of heart murmur detection and related abnormal cardiac function identification from phonocardiograms (PCGs). This work describes the novel approaches developed by our team, Revenger, to solve these problems.

Methods: PCGs were resampled to 1000 Hz, then filtered with a Butterworth band-pass filter of order 3, cutoff frequencies 25 - 400 Hz, and z-score normalized. We used the multi-task learning (MTL) method via hard parameter sharing to train one neural network (NN) model for all the Challenge tasks. We performed neural architecture searching among a set of network backbones, including multi-branch convolutional neural networks (CNNs), SE-ResNets, TResNets, simplified wav2vec2, etc.

Based on a stratified splitting of the subjects, 20% of the public data was left out as a validation set for model selection. The AdamW optimizer was adopted, along with the `OneCycle` scheduler, to optimize the model weights.

Results: Our murmur detection classifier received a weighted accuracy score of 0.736 (ranked 14th out of 40 teams) and a Challenge cost score of 12944 (ranked 19th out of 39 teams) on the hidden validation set.

Conclusion: We provided a practical solution to the problems of detecting heart murmurs and providing clinical diagnosis suggestions from PCGs.

1. Introduction

Heart murmur, defined as heart sounds produced by the turbulent blood flow through the heart, is a common clinical indicator in pediatric cardiology [1]. Accurate detection of heart murmurs and distinguishment between innocent murmurs and pathological murmurs help early clinical intervention of vital heart diseases such as congenital heart diseases, hence having a significant medical value.

Based on such motivations, the George B. Moody PhysioNet Challenge 2022 [2, 3] raised questions about detecting heart murmurs and identifying abnormal cardiac functions from phonocardiograms (PCGs), which are non-invasive heart sound recordings collected from multiple auscultation locations. In this paper, we present our methods of tackling these problems.

2. Methods

2.1. Preprocess Pipeline

After a careful study of spectral characteristics of heart murmurs from medical literature [4], and with reference to previous work [5], we constructed the PCG signal preprocessing pipeline as follows:

- Resampling to 1000 Hz;
- Butterworth bandpass filtering of order 3 and cutoff frequencies 25 - 400 Hz;
- Z-score normalization to zero mean and unit variance.

2.2. Neural Network Backbones

Inspired by the work of wav2vec2 [6], and under the consideration of exploring and utilizing the powerfulness of pretraining models on larger databases, we adopted a shrunken `wav2vec2` as one of our neural network (NN) backbones. We used the time-domain signals, namely the PCG waveforms, as model input, rather than the derived time-frequency-domain signals, for example, the spectrograms. Since PCGs have significantly lower sampling rates than conventional human voice audio signals, we reduced the dimension (number of channels) of the 'wav2vec2' model's encoder and its depth (number of hidden layers).

Considering that PCGs share a similar physiological origin as electrocardiograms (ECGs), we further adjusted and tested several NN backbones that have proven effective in ECG problems, including MultiBranch CNN, SE-ResNet, TResNetS, TResNetF [7], and ResNet-NC [8] etc. We enlarged the kernel sizes of each convolution in these backbones by a factor of 2 (the ratio of the sampling rates).

The efficacy of most of the NN backbones is validated via experiments as illustrated in Figure 1. The learning process of the wav2vec2 model was interrupted at an early

Backbone	# Params	Input Type
MultiBranch [7]	17.7M	waveforms
SE-ResNet [7]	15.9M	waveforms
ResNet-NC [8]	15.4M	waveforms
TResNetS [7]	41.0M	waveforms
TResNetF [7]	4.0M	waveforms
wav2vec2 [6]	19.8M	waveforms

Table 1: NN backbones tested for the Challenge tasks. wav2vec2 used the `transformers` implementation `Wav2Vec2Model` rather than the `torchaudio` implementation.

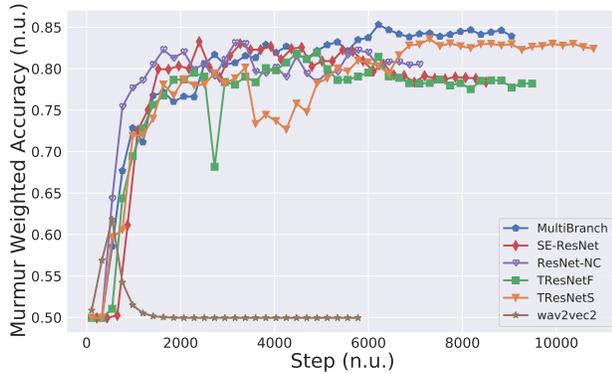stage. The cause for this abnormal phenomenon is left for further studies.

Figure 1: Curves of weighted accuracies of murmur detection on the cross-validation set 2.4 using 7 different NN backbones. The model heads, optimizers, loss functions, as well as other training setups, were kept the same.

2.3. Multi-Task Learning

The 2 Challenge tasks [3] are per-patient classification tasks. It should be noted that the Challenge database [9] provides per-recording annotations for the murmur detection task and heart sound segmentation annotations as well. We applied the multi-task learning (MTL) paradigm [10] on each recording via hard parameter sharing. More precisely, we use one NN model for all the tasks. Each task has its specific model head, typically a stack of linear layers concatenated to the shared backbone as discussed in Section 2.2. Our MTL paradigm is illustrated in Figure 2.

As depicted in Figure 3, experiments showed that models (with the same backbone) using an additional segmentation head (denoted as "MTL3") usually outperformed models with only two classification heads (denoted as "MTL2") for the Challenge tasks.

Our NN models produce per-recording predictions for the Challenge tasks. To obtain per-patient predictions, we
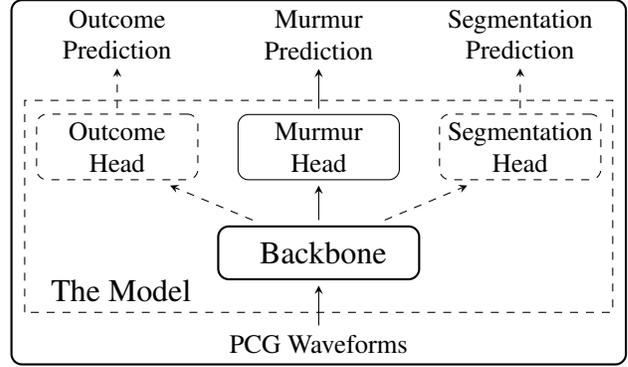

Figure 2: The paradigm of multi-task learning (MTL) used in our team's approach. The dashed lines indicate optional model heads. The "Outcome Head" and the "Murmur Head" use pooled features from the "Backbone", while the "Segmentation Head" uses the unpooled features. The heads correspond to different tasks and share the same backbone.

used the simple greedy rule described in Algorithm 1.

Algorithm 1: The algorithm to obtain per-patient predictions

---

if *at least on recording positive* **then**
 | **Positive** for the patient;
else if *all recording negative* **then**
 | **Negative** for the patient;
else // for murmur detection only
 | **Unknown** for the patient;

2.4. Training Setups

For algorithm development, we divided the publicly available part of the Challenge database into the training set and the cross-validation set with a ratio of 8:2. This split was stratified on the attributes "Age", "Sex", "Pregnancy status" and the prediction targets "Murmur", "Outcome".

The batch size was set at 32 for model training, with the maximum number of epochs set at 60. Model parameters were optimized using the AMSGrad variant of the AdamW optimizer [11] along with the `OneCycle` scheduler [12]. We froze the backbone from a specific epoch (usually 30), only updating the parameters of the task heads.

To alleviate overfitting on the training set, an early stopping callback was added. To further improve model transferability, we applied several types of augmentations to the batched training data stochastically:
- adding coloured noises;
- polarity inversion (flipping).

Page 2
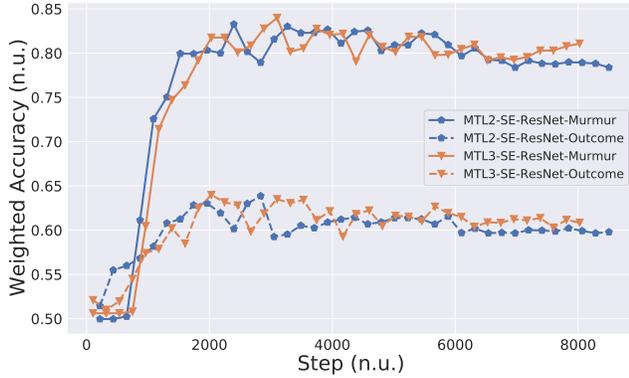
(a) Experiments using SE-ResNet as the backbone.
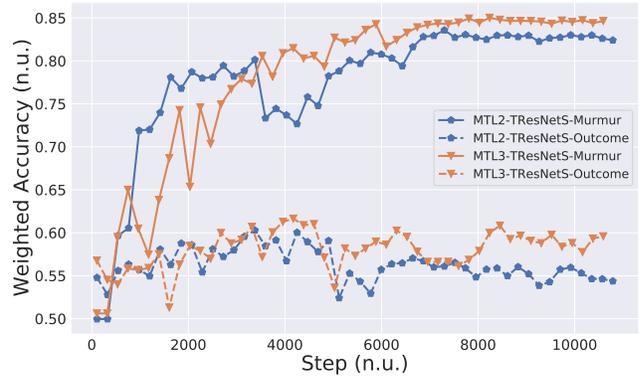
(b) Experiments using TResNetS as the backbone

Figure 3: Experiments of the MTL method with 2 heads (for murmur classification and outcome classification) and with 3 heads (an additional head for heart sound segmentation) using 2 typical backbones.

We experimented with two types of loss functions: the asymmetric loss, denoted "Loss-A"; the weighted binary cross entropy (BCE), denoted "Loss-B". The weights were obtained from the weight matrix of the Challenge scoring functions [3]. The superiority of Loss-B was observed, as was illustrated in Figure 4.
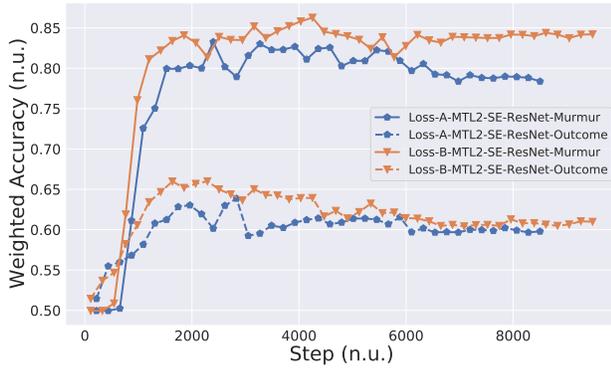

Figure 4: Experiments of comparison of the 2 loss functions (Loss-A for the asymmetric loss, Loss-B for the weighted BCE loss. The model with 2 classification heads and with SE-ResNet as the backbone was used.

2.5. Demographic Features

For the public data of the Challenge, some demographic features are strongly correlated with the prediction target "Outcome", as can be inferred from Figure 5. Experiments and official phase submissions showed that an auxiliary random forest classifier using these features and the murmur predictions improved the outcome scores (reduced the outcome cost). However, we did not use such auxiliary models in our final submission, since the distribution

of these features might be completely different in the hidden data. Moreover, no supportive medical literature was found to support this point.
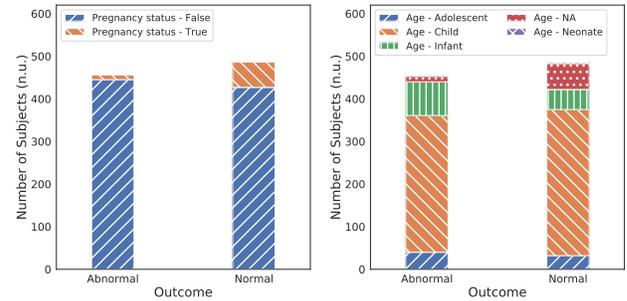
(a) Dist. against pregnancy status.

(b) Dist. against age.

Figure 5: Distributions (Dist.) of the "Outcome" against 2 typical categorical demographic variables.

3. Results

The Challenge scores (weighted accuracy for murmur detection and cost for clinical outcome identification) with an extra metric of weighted accuracy for clinical outcome identification on the train, cross-validation 2.4, the hidden validation, and the hidden test sets are gathered in Table 2. Scores on the former two sets are provided with mean and standard deviation over most of our offline experiments searching for the best NN architecture 2.2, 2.3, and loss functions 2.4.

4. Discussion and Conclusions

Our MTL paradigm proved effective for the problems of heart murmur detection and clinical outcome identification from PCGs in this study. The rankings of our team on

	Murmur	Outcome	
	wt. acc.	cost	wt. acc.
Training	0.757	12244	0.770
Cross val.	0.86 ± 0.01	11341 ± 336	0.79 ± 0.05
Hidden val.	0.678	10647	0.671
Hidden test	**0.736**	**12944**	0.761
Ranking	**14 / 40**	**19 / 39**	10 / 39

Table 2: Primary scores on the whole public training set, cross-validation set 2.4 left out from the training data, hidden validation, and the hidden test set. "wt. acc." is the abbreviation for weighted accuracy. Scores on the train and validation sets are of the form $mean \pm std.dev.$ calculated over all our offline experiments.

the hidden validation and on the whole public training set were 20 / 40, 26 / 40 for murmur weighted accuracy, and 21 / 39, 32 / 39 for outcome cost respectively. These were all significantly lower than our rankings on the hidden test set as listed in Table 2. This phenomenon is not surprising, since the MLT paradigm has already shown to have the ability to improve generalizability via leveraging latent domain-specific knowledge inherited in the training data of related tasks [10]. As for the problems that the Challenge raised, the additional segmentation head makes the shared representation (the common backbone) learn more general features and thus improves the performances for the original two classification tasks (heads).

The convolutional neural backbones also proved their effectiveness as has already shown in Figure 1. Indeed, this figure exhibits only a small part of the architectures we had experimented with. However, there is still room for improvement, as compared to the top teams on the Challenge leaderboard.

One regret of this study is that the potential of using derived time-frequency-domain signals is not explored. Previous studies on various physiological signals have shown the powerfulness of neural networks combining the derived time-frequency-domain signals with the original time-domain signals.

Another weakness of this work is that we failed to use the wav2vec2 model for tackling the Challenge problems. One possible reason is that transformer-based models need to be trained on larger datasets, and perform worse on smaller datasets than CNNs. Using larger datasets to perform self-supervised pretraining for PCGs would be a direction for our future work.

Acknowledgments

References

[1] Masic I, Begic Z, Naser N, Begic E. Pediatric Cardiac Anamnesis: Prevention of Additional Diagnostic Tests. International Journal of Preventive Medicine 2018;9(1):5.

[2] 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.

[3] 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 8 2022;URL https://doi.org/10.1101/2022.08.11.22278688.

[4] Noponen AL, Lukkarinen S, Angerla A, Sepponen R. Phono-Spectrographic Analysis of Heart Murmur in Children. BMC Pediatrics 6 2007;7(1).

[5] Schmidt SE, Holst-Hansen C, Graff C, Toft E, Struijk JJ. Segmentation of Heart Sound Recordings by a Duration-Dependent Hidden Markov Model. Physiological Measurement 3 2010;31(4):513–529.

[6] Baevski A, Zhou Y, Mohamed A, Auli M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. Advances in Neural Information Processing Systems 2020;33:12449–12460.

[7] Kang J, Wen H. A Study on Several Critical Problems on Arrhythmia Detection using Varying-Dimensional Electrocardiography. Physiological Measurement 6 2022; 43(6):064007.

[8] Ribeiro AH, Ribeiro MH, Paixão GM, Oliveira DM, Gomes PR, Canazart JA, et al. Automatic Diagnosis of the 12-lead ECG using a Deep Neural Network. Nature Communications 2020;11(1):1–9.

[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 6 2022;26(6):2524–2535.

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

[11] Reddi SJ, Kale S, Kumar S. On the Convergence of Adam and Beyond. In 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net, 2018; .

[12] Smith LN, Topin N. Super-Convergence: Very Fast Training of Neural Networks using Large Learning Rates. In Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications, volume 11006. International Society for Optics and Photonics, 2019; 1100612.

Address for correspondence:

Jingsu Kang
No. 22, Qixiangtai Road, Heping District, Tianjin, China
kangjingsu@tmu.edu.cn,kjs890223@gmail.com