Adapting Audio Foundation Models for Heart Sound Analysis

Carla Biermann, Jing Han, Cecilia Mascolo

University of Cambridge, Cambridge, United Kingdom

Abstract

Foundation models - large pretrained neural networks - have shown potential for heart sound classification tasks. However, a key question is still how to best adapt a general audio foundation model to these tasks. This work systematically studies three domain adaptation techniques, freezing the foundation model and training a linear layer on top (linear probing, LP), fine-tuning (FT), and continued pretraining (CP), on two audio foundation models using four public heart sound databases. Our findings demonstrate that LP alone is insufficient for heart sound analysis. While FT improves performance over LP, it yields models that generalise poorly to unseen datasets. To address this, we introduce CP as a novel method for heart sounds. We investigate three CP variants that differ only in their data and evaluate them via subsequent LP or FT. We find that further pretraining on the downstream dataset enhances the learned representations and boosts LP and FT performance the most. Combining all datasets for CP produces a heart-sound-specific yet task-agnostic foundation model, which improves LP and FT performance by up to 13%. These findings underscore the importance of choosing the correct adaptation strategy for heart sound analysis tasks.

1. Introduction

Deep learning approaches for heart sound analysis face challenges due to the characteristics of heart sound data. Firstly, public annotated heart sound data is sparse, with datasets containing up to a few thousand recordings. Secondly, these datasets are heterogeneous, with variations in recording equipment, patient populations, and recording environments. This heterogeneity complicates the combined use of multiple datasets and raises questions about the generalisability of models trained on a single corpus. Han et al. [1] highlighted these generalisation issues and called for methods which can address this.

Foundation models, large neural networks pretrained on vast amounts of data, have demonstrated potential in solving heart sound classification tasks. Since a public heart sound foundation model does not yet exist, likely due to the limited availability of heart sound data, prior works have utilised models pretrained on general audio or on other specialised domains such as respiratory sounds. When adapting these models, fine-tuning (FT) has been shown to be effective [2, 3]. However, fine-tuning often results in highly specialised models that perform well on the training task but fail to generalise to other datasets. Conversely, using frozen foundation models as feature extractors has been shown to be less effective on the noisy CirCor DigiScope dataset [4].

This work aims to address these challenges by providing a systematic benchmark of three domain adaptation methods: linear probing (LP), fine-tuning (FT), and continued pretraining (CP). CP involves training the foundation model from its public checkpoint on unlabelled data using the original pretraining objective. This technique has been used successfully in NLP [5–7] and automatic speech recognition [8,9]. To the best of our knowledge, CP has not been used for heart sound data. Our research is twofold: 1) to establish the performance trade-offs of LP and FT for heart sound data, and 2) to introduce CP as a novel strategy to investigate how multiple heart sound datasets can be leveraged to train a more robust model without requiring label alignment or data standardisation.

To this end, this work considers two foundation models and four heart sound databases. We first benchmark foundation model performance in the linear probing, i.e., freezing the foundation model and training a simple linear head on top, and fine-tuning scenarios. We then demonstrate that while FT improves performance upon LP, it leads to models that generalise poorly across different heart sound datasets and tasks. We propose continued pretraining as a potential solution to this limitation and study three CP scenarios: in-corpus (pretraining on a single dataset), cross-corpus (on all other datasets), and all-corpora (on a combination of all four datasets). The further pretrained models are then evaluated using LP and FT, respectively. Our results show that when applying CP using only the downstream dataset, the foundation model learns better representations, which boost LP performance. Moreover, all-corpora CP results in a single, general heart sound model that generalises well across datasets, achieving enhanced performance in both LP and FT.

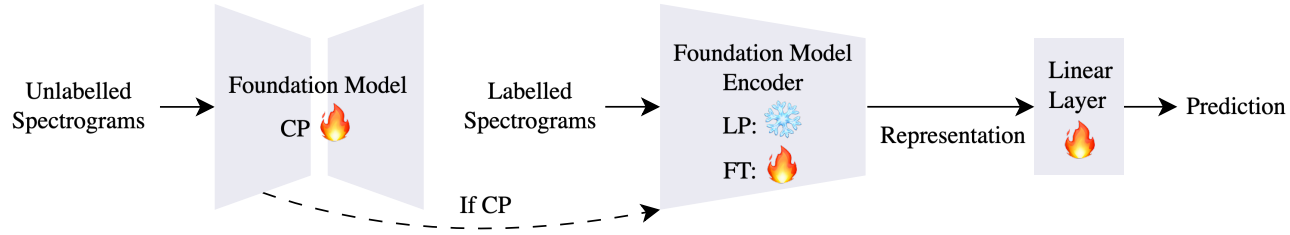

Figure 1. Overview of domain adaptation methods: linear probing (LP), fine-tuning (FT), and continued pretraining (CP).

2. Methods

Heart Sound Datasets This study uses four publicly available heart sound datasets.

The CirCor DigiScope Dataset [10], used in the 2022 PhysioNet Challenge, contains 5,272 PCG recordings from 1,568 primarily pediatric patients collected in Brazil. Each patient contributed up to four recordings, each from a different valve location. The data was recorded using electronic stethoscopes at 4 kHz. Recordings are annotated with a "present", "absent", or "unknown" murmur label, murmur characteristics, a clinical outcome label, and socio-demographic information about the patient. We use both the murmur and clinical outcome tasks in this paper.

The 2016 PhysioNet Challenge Dataset [11] comprises 3,240 recordings from 1,297 subjects. This highly heterogeneous database was curated from eight datasets collected by seven independent research groups from seven countries over a decade in clinical and non-clinical environments. For standardisation, all recordings were resampled to 2 kHz and relabelled "normal" or "abnormal".

The PASCAL Challenge Database [12] is made up of two datasets. Dataset A comprises 176 phonocardiograms collected via the iStethoscope Pro iPhone app. Dataset B was recorded in hospitals using digital stethoscopes and includes 656 heart sound recordings.

The ZCHSound Dataset [13] contains 941 high-quality recordings from pediatric patients aged 2 days to 14 years and 318 low-quality, noisy recordings from neonates. Participants include healthy individuals and those with congenital heart disease. Each patient contributed one recording obtained using a smart stethoscope with a sampling frequency of 8 kHz.

We use three downstream tasks: CirCor Murmur, CirCor Outcome, and PhysioNet 2016. The PASCAL and ZCHSound databases are merely used for CP.

Audio Foundation Models We chose two public foundation models that support fine-tuning and continued pretraining. The models were selected due to their strong performance on other low-frequency physiological data [14]. Audio-MAE [15] is a generative model pretrained on approximately 5,800 hours of audio recordings. OPERA-CT is a contrastively pretrained Transformer and one of three OPEn Respiratory Acoustic (OPERA) foundation models [14]. It is pretrained on more than 400 hours of respiratory audio data comprising breathing, coughing, and lung sounds, and achieves SoTA performance in respiratory classification tasks.

Preliminary LP and FT experiments were also conducted for the HeAR [16] and CLAP [17] foundation models. HeAR is a Transformer-based model pretrained on 174,000 hours of health-related audio data, while Contrastive Language-Audio Pretraining (CLAP) learns audio representations from text-audio pairs. HeAR performed comparably to Audio-MAE and OPERA-CT in LP but showed smaller gains after FT. CLAP achieved the highest performance across all downstream tasks in both LP and FT. However, it was excluded from CP experiments, as its text-audio pretraining does not align with our setup.

Linear Probing and Fine-tuning We first benchmark the two audio foundation models when applying LP and FT. Figure 1 illustrates the pipeline. We place a linear layer (dimensions $768\times$ no. of classes) on top of the foundation model to judge the quality of the learnt representations by the foundation models. LP freezes the encoder weights, whereas FT updates them. To test the generalisability of fine-tuned models, we fine-tune OPERA-CT on each task and adapt it to every other task using LP.

Continued Pretraining In CP, a model is further trained from its pretrained checkpoint on new unlabelled data. We consider three CP scenarios (see Figure 2).

- **In-Corpus CP**: A model is pretrained on the unlabelled train and validation data of a single dataset before being adapted to that same dataset's downstream task via LP/FT.
- **Cross-Corpus CP**: A model is pretrained on all but one of the available heart sound datasets in a leave-one-out manner and then adapted to the left-out dataset.
- **All-Corpora CP**: A single model is pretrained on a combination of all four heart sound datasets, including the train and validation set of the downstream dataset, and then adapted to each individual downstream task.

Experiment Setup This work adheres to the official train/val/test splits of the 2022 PhysioNet Challenge. For

Table 1. **Macro AUROC performance of domain adaptation strategies on heart sound test sets.** The mean and standard deviation over 5 seeds are reported. Green cells indicate improvement after CP from baseline LP or FT. † denotes statistically significant differences.

	PhysioNet 2016		CirCor Murmur		CirCor Outcome	
Strategy	**OPERA**	**Audio-MAE**	**OPERA**	**Audio-MAE**	**OPERA**	**Audio-MAE**
LP	0.831 ± 0.000	0.831 ± 0.001	0.673 ± 0.001	0.671 ± 0.007	0.571 ± 0.005	0.593 ± 0.001
CP (in) + LP	0.844 ± 0.001†	0.922 ± 0.001†	0.688 ± 0.000†	0.756 ± 0.001†	0.573 ± 0.006	0.637 ± 0.001†
CP (cross) + LP	0.798 ± 0.001†	0.853 ± 0.002†	0.651 ± 0.001†	0.739 ± 0.008†	0.561 ± 0.002†	0.626 ± 0.003†
CP (all) + LP	0.763 ± 0.001†	0.866 ± 0.002†	0.672 ± 0.004	0.755 ± 0.001†	0.577 ± 0.006	0.639 ± 0.003†
FT	0.927 ± 0.014	0.933 ± 0.014	0.757 ± 0.007	0.744 ± 0.072	0.606 ± 0.010	0.606 ± 0.026
CP (in) + FT	0.905 ± 0.022	0.951 ± 0.010†	0.732 ± 0.012†	0.841 ± 0.007†	0.607 ± 0.008	0.625 ± 0.024
CP (cross) + FT	0.916 ± 0.012	0.949 ± 0.004	0.727 ± 0.014†	0.837 ± 0.014†	0.605 ± 0.019	0.644 ± 0.005†
CP (all) + FT	0.901 ± 0.012†	0.948 ± 0.008	0.722 ± 0.018†	0.839 ± 0.005†	0.596 ± 0.012	0.642 ± 0.003†

Table 2. Mean macro AUROC over five seeds of LP applied to baseline and fine-tuned OPERA models on three heart sound tasks.

		Test Set		
		PhysioNet 2016	**CirCor Murmur**	**CirCor Outcome**
OPERA FT on	OPERA Baseline	0.831	0.673	0.571
	PhysioNet 2016	0.927	0.639	0.563
	CirCor Murmur	0.622	0.757	0.562
	CirCor Outcome	0.630	0.628	0.606

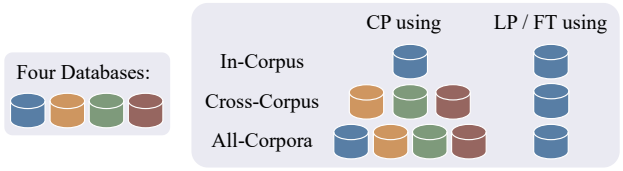

Figure 2. Continued Pretraining (CP) variants using four datasets.

the 2016 Challenge data, we used only the training set, which we divided into train, validation, and test sets. No pre-processing or data augmentation techniques are applied to isolate and evaluate the capabilities of audio foundation models. To address the high data imbalance, a weighted cross-entropy loss is employed in LP and FT, with class weights set inversely proportional to their frequencies. The Adam optimizer is used, and the learning rate is reduced at each epoch to stabilise training. The primary evaluation metric used in this study is the macro AUROC, computed as the unweighted average of the AUROC per class.

3. Results

The baseline LP and FT results can be seen in Table 1. Unsurprisingly, FT improves performance upon LP up to 12%. However, this effect is accompanied by a loss of generalisability. Table 2 shows that models fine-tuned on one task, such as CirCor Murmur, perform worse when a linear layer is applied to them on a different task, like PhysioNet 2016, compared to the baseline OPERA-CT model. This is true even for tasks originating from the same dataset (e.g., CirCor Murmur and Outcome). This suggests that models fine-tuned on different tasks, even within the same dataset, learn to emphasise different features of the spectrograms.

Table 1 also shows the results of the CP experiments. The generative Audio-MAE model benefits from all three CP scenarios. The effect of CP on OPERA-CT is more nuanced and depends on the task and scenario. While in-corpus CP improves its LP performance, this gain often disappears after FT, and in some cases, it even worsens performance. Cross-corpus and all-corpora CP are less effective for OPERA-CT, resulting in enhanced performance in only one case. This suggests that the effectiveness of CP is model-specific and may depend on the pretraining objective, data, architecture, and other factors.

Furthermore, for LP adaptation, including the downstream dataset in the CP phase (in-corpus and all-corpora CP) yields better representations and superior performance compared to excluding it. When fine-tuning, these differences disappear, and the trend sometimes is even reversed, with cross-corpus CP models occasionally outperforming all-corpora CP models.

4. Discussion

While in-corpus CP improves the learnt representations most consistently, it requires further pretraining a separate model for each dataset. A more scalable alternative is all-corpora CP, where only one central model is further pretrained on all data. As demonstrated by the Audio-

MAE results, this technique yields a heart sound-specific yet task-agnostic model with enhanced performance compared to the model checkpoint baselines.

A consideration for adopting CP is its computational expense, which is often more intensive than FT, especially when combining multiple datasets for CP. While CP+LP can perform on par with FT, they have distinct use cases. For many distinct tasks on data from the same domain (e.g. heart sounds), further pretraining one central model, like all-corpora CP, and adapting it via LP might save computational power as foundation model weights are only updated once during pretraining. This can be a great advantage in resource-constrained settings, such as on edge devices.

A major potential of CP is its ability to learn from unlabelled data. While this work uses labelled data and omits the labels for unsupervised pretraining, a valuable future direction could involve using aggregated unlabelled in-domain data for CP. This could make more data accessible for pretraining and boost LP and FT performance in experiments where additional unlabelled data is available.

Despite the performance gains, the adapted foundation models in this study do not achieve state-of-the-art performance in heart sound tasks. This may be attributed, in part, to the use of a simple linear head for adaptation, as more expressive heads, like a Transformer, have been shown to significantly improve performance in related literature [4].

5. Conclusion

This paper evaluated the domain adaptation strategies linear probing, fine-tuning, and continued pretraining of audio foundation models for heart sound analysis. By further pretraining the foundation models on a combination of heterogeneous heart sound datasets, we demonstrate the feasibility of creating a robust and generalisable heart sound foundation model. The insights gained in this paper can inform other specialised, data-sparse and heterogeneous domains that seek to utilise foundation models.

Acknowledgments

This work was supported by ERC Project 833296 (EAR).

References

[1] Han J, et al. Listening to the heart: Unifying open audio databases for cardiology research. In Proc. Computing in Cardiology (CinC), volume 51. 2024; 1–4.

[2] Niizumi D, et al. Exploring pre-trained general-purpose audio representations for heart murmur detection. In 2024 46th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC). IEEE, 2024; 1–4.

[3] Panah DS, et al. Exploring wav2vec 2.0 model for heart murmur detection. In 2023 31st Eur. Signal Process. Conf. (EUSIPCO). IEEE, 2023; 1010–1014.

[4] Niizumi D, et al. Assessing the utility of audio foundation models for heart and respiratory sound analysis. arXiv preprint arXiv250418004 2025;.

[5] Gururangan S, et al. Don't stop pretraining: Adapt language models to domains and tasks. In Proc. 58th Annu. Meeting Assoc. Comput. Linguistics. Online: Association for Computational Linguistics, July 2020; 8342–8360.

[6] Ji S, et al. Domain-specific continued pretraining of language models for capturing long context in mental health. arXiv preprint arXiv230410447 2023;.

[7] Lee J, et al. Biobert: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36(4):1234–1240.

[8] Fu B, et al. wav2vec-s: Adapting pre-trained speech models for streaming. In Findings of the Association for Computational Linguistics ACL 2024. 2024; 11465–11480.

[9] Getman Y, et al. What happens in continued pre-training? analysis of self-supervised speech models with continued pre-training for colloquial finnish asr. In Interspeech. International Society for Computers and Their Applications (ISCA), 2024; 5043–5047.

[10] Oliveira J, et al. The circor digiscope dataset: From murmur detection to murmur classification. IEEE Journal of Biomedical and Health Informatics 2021;26(6):2524–2535.

[11] Liu C, et al. An open access database for the evaluation of heart sound algorithms. Physiological Measurement 2016; 37(12):2181.

[12] Bentley P, et al. The pascal classifying heart sounds challenge 2011 (chsc2011) results. http://www.peterjbentley.com/heartchallenge/index.html.

[13] Jia W, et al. Zchsound: Open-source zju paediatric heart sound database with congenital heart disease. IEEE Transactions on Biomedical Engineering 2024;71(8):2278–2286.

[14] Zhang Y, et al. Towards open respiratory acoustic foundation models: Pretraining and benchmarking. Advances in Neural Information Processing Systems 2024;37:27024–27055.

[15] Huang PY, et al. Masked autoencoders that listen. Advances in Neural Information Processing Systems 2022;35:28708–28720.

[16] Baur S, et al. Hear – health acoustic representations, 2024. URL `https://arxiv.org/abs/2403.02522`.

[17] Elizalde B, et al. Clap learning audio concepts from natural language supervision. In 2023 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP). IEEE, 2023; 1–5.

Address for correspondence:

Jing Han
15 JJ Thomson Avenue, Cambridge, UK, CB3 0FD
jh2298@cam.ac.uk