Listening to the Heart: Unifying Open Audio Databases for Cardiology Research

Jing Han[1], Erika Bondareva[1], Tomasz Jadczyk[2], Cecilia Mascolo[1]

[1] University of Cambridge, UK
[2] Medical University of Silesia, Poland / International Clinical Research Centre, Czechia

Abstract

While computer-assisted cardiac auscultation is highly promising, its accuracy, effectiveness, and robustness require rigorous evaluation. Particularly, to the best of our knowledge, no prior study has assessed the generalisability of heart sound analysis methods across multiple datasets. In this work we aggregated multiple open source datasets to study the robustness of machine learning-based abnormal heart sound detection algorithms. Specifically, we evaluated a machine learning model on four publicly available heart sound datasets under four different cross-validation settings: within-corpus, cross-corpus, and two multi-corpus settings (data aggregation and decision aggregation). Our findings reveal that the multi-corpus setting with data aggregation outperforms the cross-corpus setting, suggesting that combining varied data sources enhances generalisability. However, despite this improvement, there are still challenges that require further investigation, which we discuss in detail. Overall, the study emphasises the need for clear protocols in data collection, labelling, and sharing to ensure fair comparisons and a deeper understanding of model generalisability.

1. Introduction

Cardiovascular diseases (CVDs) are the leading cause of death globally, accounting for about 32% of all deaths.Traditional cardiac auscultation remains fundamental for the screening and clinical diagnosis of certain CVDs, particularly valvular heart diseases (VHDs). While non-invasive and widely accessible, auscultation has limitations in diagnostic accuracy and detail compared to other diagnostic techniques, such as echocardiography, which offers higher sensitivity, specificity, and a more comprehensive assessment of cardiac health [1].

Recently, artificial intelligence (AI)-enhanced medical devices show promise as community-based screening tools for identifying patients with clinically significant VHDs in the general unselected population [2]. A digital stethoscope combined with AI-based acoustic features extraction can support CVD diagnostic process. While current heart sound analysis (HSA) technology is minimally used in clinical practice, novel approaches in the emerging field of audiomics are showing promise for population-based screening [3]. These studies have shown encouraging results when trained and tested on a single dataset, but there has been limited effort to combine varied data sources. To the best of our knowledge, no prior study has assessed the generalisability of heart sound analysis methods across multiple datasets.

In this work, we address the critical need for generalisability evaluation of a heart sound classification model across different datasets, examining its performance when faced with variations in data collection devices and acoustic environments.

To this end, we evaluate our chosen model across four publicly available heart sound datasets under four distinct settings: within-corpus, cross-corpus, data aggregation, and decision aggregation. Our findings indicate that data aggregation outperforms the cross-corpus setting, suggesting that integrating diverse data sources enhances model generalisability. Despite this improvement, our results also reveal persistent challenges that warrant further investigation. We discuss these challenges in detail, emphasising the need for standardised protocols in data collection, labelling, and sharing. Such protocols are crucial for facilitating fair comparisons and advancing our understanding of model generalisability. Through this study, we aim to encourage collaboration and transparency within the heart sound analysis community, promoting a more robust framework for future research.

2. Methods

2.1. Abnormal Heart Sound Detection

In the PhysioNet 2022 Challenge, the HearTech+ team proposed a novel deep learning method for HSA [4]. It employs a hierarchical multi-scale convolutional neural network (HMS-Net) designed for both murmur and clinical outcome classification. The network establishes long short-term dependencies between multi-scale features, enhancing classification performance. Predictions are based on ensembled segment predictions using a sliding window, and a recording is considered 'abnormal' if more than one-third of its segments are labelled 'abnormal'. Moreover, for patient-level prediction, a patient is predicted as 'abnormal' if they have at least one 'abnormal' recording. For more detailed information about this HearTech+ model, readers are kindly referred to [4].

HearTech+ was chosen as our base model because: (1) the code is publicly available, allowing for fair comparison; (2) the model achieved notable performance, securing 2nd place in heart murmur detection and 9th place in abnormal cardiac function detection tasks among 53 teams, and (3) it does not require segmentation information, facilitating its evaluation across diverse datasets, particularly those lacking such detailed annotations.

The original proposed structure incorporated patient information such as age, gender, height, and weight to distinguish patients with abnormal clinical outcomes. However, since this patient information is not consistently available across all datasets we evaluated, we removed these patient feature embeddings from the original structure to facilitate a fair comparison across all evaluated datasets.

2.2. Evaluating Method Generalisability

To evaluate the generalisability of the HearTech+ model for clinical outcome prediction, we employ the following four evaluation strategies:

Within-corpus Cross-validation (CV): We perform 5-fold CV on each database. When patient ID is available, the folds are made individual-independent, ensuring that no samples from the same individual appear in more than one fold. This evaluation aims to report the performance of the selected model within each heart sound dataset.

Cross-corpus Evaluation: It involves evaluating a trained model on entirely different datasets. Here, we used four datasets, meaning that each of the four trained models was tested independently on the remaining three datasets.

Data Aggregation Evaluation: Rather than training on a single dataset and testing on the others, this strategy expands the training corpus by combining all available datasets, excluding the one designated as the test set. The model is then evaluated on the remaining test corpus.

Decision Aggregation Evaluation: Similar to data aggregation evaluation, this strategy leverages multiple datasets. Classifiers are trained on each single dataset. During testing, their decisions are combined via majority voting for the final evaluation on the unseen test corpus.

2.3. Datasets

For our evaluation, we explored four publicly available databases, which are described in detail below and summarised in Table 1.

The CirCor DigiScope Dataset[1] This dataset includes heart sound recordings collected during two mass screening campaigns conducted in Brazil [5], and it was later used in the 2022 PhysioNet Challenge [9]. The database comprises 5,282 heart sound recordings from 1,568 patients, with participants' ages ranging from 3 days to 30 years. The recordings were captured using a Littmann 3200 stethoscope from four standard auscultation points at a sampling frequency of 4 kHz. The dataset also includes demographic information, murmur-related labels, outcome-related labels, annotations of murmur characteristics, and heart cycle segmentation information.

2016 PhysioNet Challenge Dateset[2] This database was created for the PhysioNet Challenge 2016 [6, 10]. It consists of nine different heart sound databases compiled from various research groups. The dataset includes recordings from 1,297 subjects, both healthy individuals and patients with a range of conditions such as VHD and coronary artery disease. Recordings were collected across diverse clinical and non-clinical settings using various equipment. All recordings were resampled to a frequency of 2 kHz.

The PASCAL Challenge Database [3] This dataset was introduced as part of the PASCAL Classifying Heart Sounds Challenge in 2011 [7]. The dataset consists of two sets: Set A and Set B. Set A contains 176 samples collected from an unspecified population using a smartphone app, while Set B includes 656 recordings obtained with a digital stethoscope system in a clinical unit in Recife, Brazil. All recordings were made at a sampling rate of 4 kHz, in both clinical and non-clinical settings. The annotations differ between the two sets: Set A was categorised into four classes: normal, murmur, extra heart sound, and artifact; while Set B was labelled into three classes: normal, murmur, and extra systole.

The ZCHSound Dataset [4] This dataset is an open-source collection of heart sound recordings, primarily focused on paediatric heart sounds, with participants' ages ranging from 2 days to 14 years [8]. It includes data from 1,259 participants and is divided into two main subsets: a high-quality heart sound dataset containing recordings from 941 participants, and a low-quality set comprising recordings from 318 newborns within the first five days of birth. The recordings are sampled at 8 kHz and categorised into five classes based on diagnosed cardiac conditions: normal, atrial septal defect (ASD), patent ductus

[1] https://physionet.org/content/circor-heart-sound/
[2] https://physionet.org/content/challenge-2016/1.0.0/files
[3] https://istethoscope.peterjbentley.com/heartchallenge
[4] http://zchsound.ncrcch.org.cn/dataset

Table 1. Summary of Four Public Heart Sound Datasets.

Dataset Name	Subjects No.	Samples No.	Mean Duration (s)	Duration Range (s)	Sampling Rate (Hz)	Labelling Strategy
PhysioNet 2022 [5]	1,568	5,282	20.90	4.75-80.37	4,000	Murmur/No murmur/Unknown or Normal/Abnormal
PhysioNet 2016 [6]	1,297	3,240	22.35	5.31-122.00	2,000	Normal/Abnormal/Unsure
PASCAL Challenge [7]	–	832	6.24	0.76-24.45	4,000	Set A: Normal/Murmur/Extra Heart Sound/Artifact; Set B: Normal/ Murmur/Extrasystole
ZCHSound [8]	1,259	1,259	20.11	6.46-60.12	8,000	Normal/ASD/PDA/PFO/VSD

arteriosus (PDA), patent foramen ovale (PFO), and ventricular septal defect (VSD).

2.4. Prediction and Evaluation

While the original labels of the four selected datasets differ, they were mapped to ‘normal’ or ‘abnormal’. Specifically, samples labelled as ‘unsure’ in PhysioNet 2016 were excluded. In PASCAL, samples with extra heart sound and extra systole labels were relabelled as ‘abnormal’ and artifact samples were removed. In ZCHSound, all four cardiac conditions were relabelled as ‘abnormal’. All recordings were further downsampled to 2 kHz to remove the discrepancy across datasets.

We evaluate the performance in terms of sensitivity, specificity, and cost, per patient, following the same patient aggregation strategy of [4]. This excludes PhysioNet 2016, where patient information is not available. The cost measure was initially introduced in 2022 PhysioNet Challenge, to rank clinical outcome classifiers. This measure accounts for the costs associated with algorithmic prescreening, expert screening, treatment, and diagnostic errors that can lead to delayed or missed treatments [9].

3. Results and Discussion

The experimental results are presented in Table 2. For both within-corpus and cross-corpus evaluations, averaged performance is reported either across five folds or across three training datasets. For the within-corpus evaluation on PASCAL, we created five folds: one from Set A and four independent folds from Set B with patient IDs inferred from file names. In the cases of data aggregation and decision aggregation evaluations, performance metrics are provided separately for each test dataset. Note that the cost metrics can only be compared across different settings within the same datasets, as these measures are intrinsically linked to the size of the testing set.

As shown in Table 2, in the within-corpus evaluations, the models achieved satisfactory results on patient-independent splits of PhysioNet 2016 and ZCHSound. On PhysioNet 2022, the performance was comparable to that reported in [4]. However, this performance diminished significantly during cross-corpus evaluations, demonstrating challenges in model generalisation when applied to unseen datasets. This underscores the difficulty of maintaining accuracy across varied populations and recording conditions.

The implementation of data aggregation showed improvements in several cases. It suggests that data aggregation can mitigate some limitations posed by individual datasets. However, the effectiveness varied, highlighting that while data aggregation can be beneficial, it may not uniformly improve performance across all datasets.

The current decision aggregation method did not address the data source mismatch issue effectively. This indicates that decision aggregation alone may not adequately account for the variability in heart sound recordings from different contexts. Instead of averaging decisions equally from all classifiers, it may be beneficial to consider the confidence levels of individual classifiers. This approach could potentially enhance performance and is worth exploring in future research.

Overall, the generalisability across dataset evaluations reveals persistent challenges in the task of heart sound abnormality classification. Further research is needed to develop more robust models capable of accurately classifying heart sound abnormalities across diverse populations and recording conditions. To address these challenges, it is crucial to consider how future heart sound databases can be compiled and shared more effectively, ensuring they provide the necessary depth and quality for advancing research in this field.

4. HSA Database Compilation Insights

The results of this study indicate significant generalisability issues present in existing ML-based models for HSA. The publicly available heart sound datasets also have limitations that hinder their utility for comprehensive research. Notably, the labelling strategies across these datasets lack standardisation, making direct comparisons and model training more challenging. For example, only the PhysioNet 2022 dataset includes labels for both murmur detection and clinical outcome classification, while

Table 2. Performance in terms of Sensitivity ($se.$), Specificity($sp.$), and Cost Metrics ($cost$) over four cross-validation evaluations on four HSA datasets.

	PhysioNet 2022			**PhysioNet 2016**			**PASCAL**			**ZCHsound**		
	$se.$	$sp.$	$cost$	$se.$	$sp.$	$cost$	$se.$	$sp.$	$cost$	$se.$	$sp.$	$cost$
within-corpus	.695	.580	11529	.829	.967	3523	.574	.783	16076	.862	.841	7384
cross-corpus	.795	.290	14788	.641	.307	7014	.667	.286	15961	.505	.720	13641
data agg.	.616	.617	12764	.577	.379	6738	.592	.333	16714	.512	.719	13446
decision agg.	.976	.037	14082	.635	.279	7296	.763	.242	14519	.509	.792	13454

PASCAL is the only dataset that labels the extra heart beats. To aid training models that generalise better, we believe that, where feasible, establishing a more standardised labelling approach is essential.

Beyond labelling, several other key aspects should be considered, including:

• *Optimal Audio Quality*: higher sampling rates are recommended, and longer recording durations (over 20 seconds) are essential to capture sufficient heart cycles with good quality. Recording in a quiet environment is also recommended, to reduce the amount of environmental noise captured, which would negatively affect the HSA.

• *Rich Recording Documentation*: it is beneficial to record clear information about auscultation locations on the chest, and specify the quality and type of recording devices used.

• *Comprehensive Patient Information*: it is essential to include standard demographic details such as age, gender, and medical history to provide valuable context for the data, along with subject IDs to facilitate individual-independent validation.

Incorporating these factors will significantly enhance the quality, informativeness, and clinical relevance of heart sound datasets. Such improvements are crucial for developing more robust and generalisable HSA models, ultimately advancing research in this field.

5. Conclusion

In this study, we aggregated four public heart sound datasets for the first time to investigate the generalisation capability of models across varied datasets. We evaluated the performance of HMS-Net on a binary task of abnormal heart sound identification using four different setups: within-corpus, cross-corpus, data aggregation, and decision aggregation. While the models demonstrated acceptable performance in within-corpus evaluations, their ability to generalise across datasets presents a significant challenge. These findings highlight the necessity for ongoing research aimed at enhancing the robustness and applicability of heart sound classification models, emphasising the importance of developing methods that can effectively address the variability inherent in diverse datasets. Furthermore, we provide recommendations for future heart sound dataset compilation, aiming to improve dataset quality, standardisation, and clinical relevance, which are crucial for advancing research in this field.

Acknowledgments

This work was supported by ERC Project 833296 (EAR) and Statutory funds of the Medical University of Silesia in Poland (no. (PCN-1-005/N/0/K and PCN-1-139/N/2/K).

References

[1] Jariwala N, et al. Clinically undetectable heart sounds in hospitalized patients undergoing echocardiography. JAMA Internal Medicine 2022;182(1):86–87.

[2] Sengupta PP, et al. The future of valvular heart disease assessment and therapy. The Lancet 2024;.

[3] Ghanayim T, et al. Artificial intelligence-based stethoscope for the diagnosis of aortic stenosis. The American Journal of Medicine 2022;135(9):1124–1133.

[4] Xu Y, et al. Hierarchical multi-scale convolutional network for murmurs detection on pcg signals. In Proc. Computing in Cardiology (CinC), volume 498. 2022; 1–4.

[5] 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.

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

[7] Bentley P, et al. The PASCAL Classifying Heart Sounds Challenge 2011 (CHSC2011) Results. http://www.peterjbentley.com/heartchallenge/index.html.

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

[9] Reyna MA, et al. Heart murmur detection from phonocardiogram recordings: The george b. moody physionet challenge 2022. PLOS Digital Health 2023;2(9):e0000324.

[10] Clifford GD, et al. Classification of normal/abnormal heart sound recordings: The physionet/computing in cardiology challenge 2016. In Proc. Computing in Cardiology Conference (CinC). 2016; 609–612.

Address for correspondence:

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