Cross-Domain Detection of Pulmonary Hypertension in Human and Porcine Heart Sounds

Alex Gaudio[1,2], Noemi Giordano[3], Miguel Coimbra[2], Benedict Kjaergaard[4], Samuel Schmidt[5], Francesco Renna[2]

[1] Carnegie Mellon University, Pittsburgh, United States
[2] INESC TEC, Faculty of Sciences of the University of Porto, Porto, Portugal
[3] Politecnico di Torino, Torino, Italy
[4] Aalborg University Hospital, Aalborg, Denmark
[5] Aalborg University, Aalborg, Denmark

Abstract

Detection of Pulmonary Hypertension (PH) via the automated analysis of cardiac auscultation may offer a non-invasive, accurate, and reliable solution with low resource requirements. We detect PH in human and in porcine datasets and demonstrate domain generalization across the two datasets. Extending our previous work, we train a deep network on a representation of segmented second heart sounds (S2). The human dataset contains digital stethoscope (PCG) recordings of 42 patients. The porcine dataset contains 110 samples of PCG and seismocardiography (SCG) recordings obtained from pigs with chemically induced PH. In both datasets, ground truth reference indicators of PH were obtained via right heart catheterization (RHC). The area under the ROC curve (auROC) and area under the Precision-Recall curve (AP) on human data are 0.92 and 0.97, respectively. On the porcine dataset, leave-one-out cross-validation gives 0.84 auROC and 0.85 AP. Moreover, we demonstrate transferability across domains, where training on the porcine dataset and evaluating on the human dataset gives 0.702 auROC and 0.848 AP. Results show that it is possible to use porcine data for developing human AI models, and that Phonocardiogram (PCG) and Seismocardiogram (SCG) training data can be used to evaluate PCG data.

1. Introduction

Pulmonary Hypertension (PH) is a hemodynamic state describing high blood pressure in the pulmonary system. It affects approximately 1% of the human population [**?**]. Guidelines published by the European Society of Cardiology (ESC) and European Respiratory Society (ERS) prioritize early detection in order to fast-track patients to specialized PH treatment centers [**?**]. PH is challenging to diagnose [**?**]. In this work, we consider the analysis of heart sounds for low-cost and non-invasive detection of PH, and we demonstrate reliability of the approach via within-domain and across-domain generalization experiments.

Existing approaches for PH detection: Existing technologies for screening either lack reliability, are too invasive or expensive, or cannot detect PH in some patients. Right heart catheterization (RHC) is the gold standard for diagnosis and detection of PH, but it is not applicable as a screening tool due to the highly invasive procedure, high cost, and need for a specialized clinical environment [**?**, **?**]. Other technologies are helpful in combination to diagnose PH, and these include electrocardiography (ECG), echocardiography (ECHO), chest radiography, magnetic resonance imaging (MRI), pulmonary function tests, and blood tests. None of these approaches are known, at the time of this writing, to give a decisive detection of PH in all settings. ECHO, in particular, is low-cost method that can be useful in screening settings, but current recommendations [**?**] and a meta-survey of ECHO for PH detection [**?**] state that the technology is too limited to rely upon alone. Estimates needed for PH detection were unreadable in 40% of patients [**?**], unreliable in 50% of patients [**?**] or frequently in limited agreement with measurements obtained via RHC [**?**]. Similarly, ECG can help diagnose PH [**?**] in approximately 40% of analyzed patients [**?**]. A premise of our work is that the computer analysis of digitized heart sounds, as recorded via phonocardiography or seismocardiography, can give superior reliability with non-invasive, low-cost, and user-friendly technology.

Heart sound physiology for PH detection: The two main audible events of a heart sound can be identified as the First Heart Sound (S1) and Second Heart Sound (S2). The observed S1 is predominantly formed by the closing of the atrioventricular valves (the mitral valve and to a lesser

extent the tricuspid valve), and we assume it is not relevant for PH detection. The S2 is formed by the Aortic valve closure (A2) and Pulmonary valve closure (P2). The relative time delay between the A2 and P2 events as observed in the sound signal are clinically significant for PH diagnosis [**?**].

Analysis of cross-domain generalization enhances trustworthiness of our approach: A machine learning algorithm, such as our proposed PH detector, trained on one dataset may or may not preserve its performance when evaluated on a similar dataset. Domain generalization is a type of domain adaptation [**?**] that analyzes empirical performance on out-of-distribution test sets. In our setting, the predictive task of PH detection is unchanged, but the data domain changes. We consider both Human and porcine datasets, and both PCG and SCG heart sound recording technologies.

Contributions: This work demonstrates the ability to generalize across domains for PH detection. It is shown that: (a) porcine training data can be used to evaluate human data and (b) a training dataset consisting of SCG and PCG data can be used to evaluate PCG data.

2. Methodology

We perform cross-domain and within-domain evaluations of a PH detector on two datasets.

2.1. Datasets:

Porcine: The porcine dataset contains a mixture of SCG and PCG samples from pigs undergoing right heart catheterization. A reversible PH condition was produced in the animals by subjecting them to either nitrogen asphyxiation (to cause hypoxemia) or CO_2 asphyxiation (to cause hypercapnia) [**?**]. A total of 59 experiments performed on 10 animals yielded 118 samples, of which 8 were excluded for quality concerns. The experiments took place at the Aalborg University Hospital, Aalborg, Denmark.

As shown in Figure 1, the heart sounds were recorded by means of two different types of equipment. Most of the recordings collected SCG data using an iWorx™ commercial system equipped two triaxial accelerometers located respectively over the lower border of the sternum and over the fourth intercostal space next to the sternum. The remaining recordings collected PCG data using a custom-made multi-sensor array developed at Politecnico di Torino. The array embeds 48 microphones and 3 electrodes for ECG recording. ECG was collected in all recordings. Details about the device are previously published [**?**]. For the purposes of this study, the SCG/PCG signal with the highest Signal-to-Noise Ratio (SNR) was selected. For each experiment, two one-minute segments were extracted from the recordings: a one-minute segment at the beginning of the recording, before any trigger was applied (baseline condition, labelled as no PH); and a one-minute segment centered on the peak of the PAP (labelled as PH). The porcine dataset includes 118 one-minute recordings, of which 8 recordings were excluded for quality concerns. Dataset statistics are summarized in Table 1. The dataset is unpublished.

Table 1. Porcine Dataset

	PH neg	PH pos
SCG	47	36
PCG	13	14
Total	60	50

Table 2. Human Dataset

	PH neg	PH pos
SCG	0	0
PCG	13	29

Human: The human dataset contains PCG samples from 42 human subjects undergoing right heart catheterization. Table 2 shows that 29 subjects have PH. Inclusion and exclusion criteria are unknown. The data was collected from Centro Hospitalar Universitário do Porto, Portugal, using a custom stethoscope connected to a Rugloop Waves system. Auscultation was performed over the second left intercostal space on patients in the supine position and at rest in a quiet environment. This private dataset was previously introduced in [**?**].

2.2. Model Design:

Pre-processing: The audio signal data for a given subject is a digital recording of the subject's beating heart. In the Porcine dataset, the signals were band-pass filtered between 25 Hz and 450 Hz and then segmented into heartbeats using the R-wave peaks extracted from the ECG reference signal. S2 segments were extracted into 200 ms windows using a double thresholding approach and band-pass filtered once more to 25 Hz and 200 Hz. The sample rate of the data was 1 kHz. Similarly, the Human samples were band-pass filtered between 25 Hz and 400 Hz, resampled to 1 kHz, and then segmented into S2 sounds of 200 ms windows. The Human pre-processing procedure is identical to our previous work [**?**]. In both datasets, the S2 sounds were represented as a 2-dimensional image matrix, such as shown in Figure 2. The images of the Porcine and Human datasets were zero-padded to 157 and 454 rows, respectively, by adding rows of zeros to the bottom of the figure. Padding facilitated using a deep learning library to stack image matrices from multiple samples into a tensor to improve training speed, and padding also ensures the input to the deep network satisfies minimum dimensions required by the current choice of deep network. Zero padding is not necessary to the overall approach. The columns of Human samples were normalized to unit variance after padding to facilitate stable gradient backpropagation, following our previous work.

Model Design: For our cross-domain generalization ex-

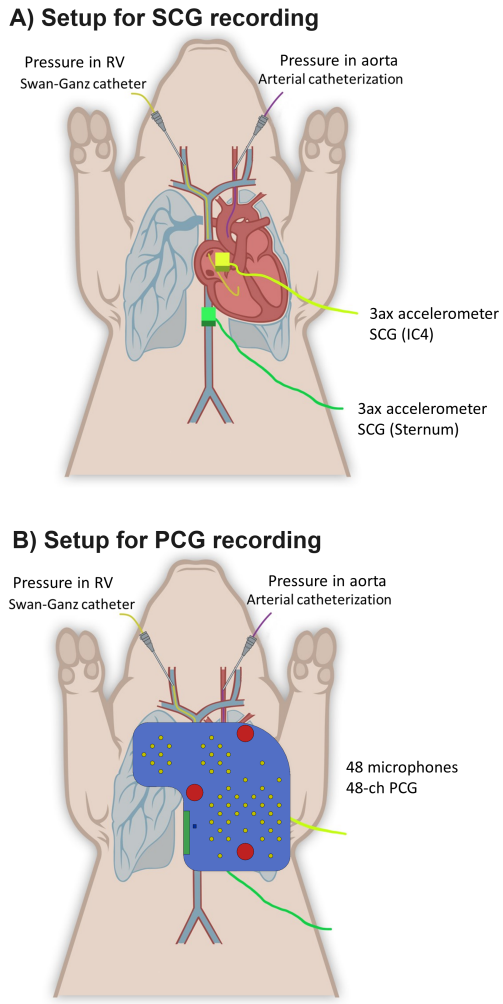

Figure 1. **Porcine Dataset Experimental Setup.** Representation of the animal setup using accelerometers (panel A) or the multi-sensor array (panel B).

periments, we adapt the DenseNet121 [**?**] model with random (rather than pre-trained) initialization because it was previously used in [**?**]. The AdamW optimizer used learning rate 0.00001 and weight decay 0.000001; all other hyperparameters were default values from the PyTorch library [**?**]. The loss function was unweighted binary cross entropy loss. To facilitate backpropagation of meaningful gradients with the small datasets, batch gradient descent was employed rather than stochastic or minibatch gradient descent. Models on both datasets were trained for 600 epochs, and then evaluated at that epoch.

Evaluation Methodology: Two evaluation schemes are subsequently described: within- and across-domain.

Within-domain evaluation considers a model's empirical PH detection performance on a single dataset by making use of bootstrapped cross validation (CV). By bootstrap, we mean that CV was applied independently N times, where each independent run had a different random seed to initialize the model parameters and split the training and validation sets. The Human dataset leveraged k-fold CV. The Porcine dataset employed leave-one-group-out CV with k groups, where a group corresponds to all recordings from one of the ten pigs. Both datasets used $N = 12$ and $k = 10$, yielding kN independently trained models and validation sets for a given experiment. The model was trained on the training data, evaluated on the validation data, and performance statistics from each of the kN validation sets were computed. Our previous work with Human used $N = 1$ [**?**], and we now use $N = 12$ to have improved confidence that the model does not overfit based on the way the training and validation sets were split. Note that while the random seed has no effect on dataset split for the leave-one-group-out CV, it does affect the deep network's random initialization and subsequent parameter updates (via backpropagation).

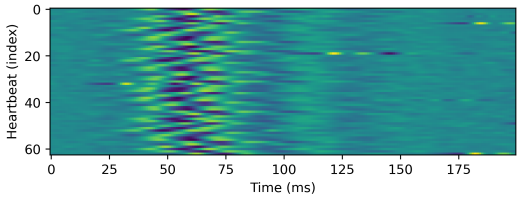

Figure 2. **Pre-processing illustration:** Example S2 matrix, prior to zero padding the bottom rows.

Across-domain evaluation trains a model on one dataset and evaluates on the other in order to evaluate how well the dataset generalizes to new domains. In this paper, the Porcine dataset serves as the training set, and the Human dataset is the test set. Cross validation is not applicable ($k = 1$). We trained twelve independent models, so that $kN = 12$, where each model was initialized randomly by a different random seed.

Metrics: Each experiment consists of kN models and validation sets. Each model was evaluated on its corresponding validation set with micro and macro averaged scores. The scores considered are area under the receive operating characteristic curve (auROC) and average precision (AP). The auROC ranges from [0,1], where 0.5 is a completely random classifier. The AP summarizes the area under the precision recall curve, ranges from [0,1], and describes a completely random classifier with a value of $\frac{n_1}{n}$ where the fraction relates the number of PH positive samples (n_1) to the total number of samples (n). The AP and auROC, analyzed together rather than individually, give a better sense of performance [**?**]. The metrics were aggregated via macro and micro averaging over the kN validation sets. Macro average computes the average of the AP or auROC computed on each of the kN validation sets,

Table 3. Within-domain and Across-domain Results

	Macro auROC	AP	Micro auROC	AP
Human (PCG)	0.92	0.97	0.92	0.96
Porcine (PCG and SCG)	0.84	0.85	0.86	0.84
Porcine → Human	0.70	0.84	0.70	0.83

while micro average computes the score by assembling all predictions from all kN validation sets into a single vector and then evaluating the score function once.

3. Results

Table 3 shows strong PH detection performance. The Human dataset has auROC and AP above 0.92, and Porcine is above 0.84. The Porcine ground truth may be different due to the way hypertension was labeled. The within-domain performance serves informally as a baseline for the across-domain performance. Note that each within-domain model was trained on $\frac{9}{10}$ of the dataset, and each across-domain model was trained on all ($\frac{10}{10}$) of the training dataset. Training on Porcine data and then evaluating that model on human data demonstrates that the datasets are quite similar, even though they differ in animal type, data collection methodology, and distribution of PH positive versus negative samples.

4. Conclusion

The results of this research support the analysis of digitized heart sounds as a solution for early detection of PH. Our results suggest that developing models on porcine data may offer a safe and suitable alternative to working with human data. Our future work will evaluate out-of-distribution generalization in additional datasets. Our results also suggest that the analysis of SCG sounds may be a viable pathway for PH detection, though further study needs to be done with SCG. Our future work will also explore ways to validate our approach to PH detection in screening settings with upright or standing human patients not undergoing right heart catheterization. The value of our approach is that sound analysis methods proposed in this work are non-invasive and very accurate. It is hoped that our approach may be able to replace the need for highly invasive methods in the detection and diagnosis of pulmonary hypertension.

Acknowledgments

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project UIDB/50014/2020. An international patent application pertains to this work: PCT/IB2023/058675.

References

[1] Hoeper MM, Ghofrani HA, Grünig E, Klose H, Olschewski H, Rosenkranz S. Pulmonary hypertension. Deutsches Arzteblatt International 2017;114(5):73.

[2] Humbert M, Kovacs G, Hoeper MM, Badagliacca R, Berger RM, Brida M, Carlsen J, Coats AJ, Escribano-Subias P, Ferrari P, et al. 2022 ESC/ERS guidelines for the diagnosis and treatment of pulmonary hypertension. European Heart Journal 2023;.

[3] Bazan IS, Fares WH. Pulmonary hypertension: diagnostic and therapeutic challenges. Therapeutics and Clinical Risk Management 2015;1221–1233.

[4] Janda S, Shahidi N, Gin K, Swiston J. Diagnostic accuracy of echocardiography for pulmonary hypertension: a systematic review and meta-analysis. Heart 2011;97(8):612–622.

[5] Kovacs G, Avian A, Foris V, Tscherner M, Kqiku X, Douschan P, Bachmaier G, Olschewski A, Matucci-Cerinic M, Olschewski H. Use of ECG and other simple noninvasive tools to assess pulmonary hypertension. Plos One 2016;11(12):e0168706.

[6] Perloff JK. Auscultatory and phonocardiographic manifestations of pulmonary hypertension. Progress in Cardiovascular Diseases 1967;9(4):303–340.

[7] Guan H, Liu M. Domain adaptation for medical image analysis: a survey. IEEE Transactions on Biomedical Engineering 2021;69(3):1173–1185.

[8] Schmidt SE, Dolmer MH, Hansen J, Struijk JJ, Søgaard P, Kjærgaard B. Porcine model for validation of noninvasive estimation of pulmonary artery pressure. In 2022 Computing in Cardiology (CinC), volume 498. IEEE, 2022; 1–4.

[9] Giordano N, Rosati S, Balestra G, Knaflitz M. A wearable multi-sensor array enables the recording of heart sounds in homecare. Sensors 2023;23(13):6241.

[10] Gaudio A, Coimbra M, Campilho A, Smailagic A, Schmidt SE, Renna F. Explainable deep learning for non-invasive detection of pulmonary artery hypertension from heart sounds. In 2022 Computing in Cardiology (CinC), volume 498. IEEE, 2022; 1–4.

[11] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017; 4700–4708.

[12] Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, Lin Z, Desmaison A, Antiga L, Lerer A. Automatic differentiation in PyTorch 2017;.

[13] Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd International Conference on Machine Learning. 2006; 233–240.

Address for correspondence:

Alex Gaudio
Johns Hopkins University, Baltimore, MD, 21218
agaudio2@jh.edu