Quantifying Domain Shift in 3D-CNNs for Heart-Sound-Based Pulmonary Hypertension Detection

Noemi Giordano[1], Cátia Isabel Costa[2], Cristina Oliveira[3], André Lobo[2], Ricardo Fontes-Carvalho[3], Francesco Renna[4]

[1]Department of Health Science and Technology, Aalborg University, Aalborg, Denmark
[2]Unidade Local de Saùde de Gaia e Espinho (ULSGE), Porto, Portugal
[3]UnIC@RISE, Porto, Portugal
[4]INESC-TEC, Faculdade de Ciências da University of Porto, Porto, Portugal

Abstract

***Aims**: Heart sounds offer a promising, noninvasive alternative for Pulmonary Hypertension (PH) screening. Our 3D-CNN, using the beat-wise time-frequency distribution of the second heart sound (S2), achieved an AUC of 0.89 on dataset D1. However, AUC dropped to 0.48 on dataset D2, despite similar training metrics. This study quantifies the domain shift causing this discrepancy in generalizability.* **Methods:** *We applied transfer learning from D1 to D2 (with and without fine-tuning). We also analyzed Grad-CAMs and quantified input-activation alignment via correlation and Intersection over Union (IoU) to assess if the model focused on the S2 region.* **Results:** *Zero-shot transfer learning outperformed fine-tuning (AUC 0.66 vs. 0.48). Grad-CAMs revealed that while the model correctly focused on S2 in D1, it relied on non-generalizable shortcuts in D2. Both correlation and IoU were strongly associated with AUC, successfully quantifying these detrimental domain shifts.* **Conclusions:** *Deep learning for PH detection is highly vulnerable to domain shifts. Integrating robust domain adaptation techniques is essential for reliable real-world clinical deployment.*

1. Introduction

Pulmonary Hypertension (PH) is a severe cardiopulmonary condition involving an increased pulmonary artery pressure (PAP). Current prevalence is estimated around 1%, but many cases remain undiagnosed due to a lack of screening tools [1]. According to guidelines, the gold standard tool to diagnose PH is Right Heart Catheterization (RHC) [2]. Nevertheless, given the invasive nature, high cost, and inherent risks associated with RHC, there is a critical need for noninvasive, accessible screening alternatives. Phonocardiograms (PCG) is a promising modality for this purpose, as the acoustic properties of the second heart sound (S2) are tightly linked to pulmonary hemodynamics and intracardiac pressures [3].

Advancements in deep learning have significantly improved the automated analysis of cardiac acoustic signals [4], [5], [6]. In our previous work [7], we developed a 3D Convolutional Neural Network (3D-CNN) that leverages the beat-wise time-frequency distribution (TFD) of S2. By capturing the complex, transient spectro-temporal features of heart sounds, this architecture achieved a high discriminatory performance, yielding an Area Under the Curve (AUC) of 0.89 on a private dataset (D1). However, the translation of such deep learning models into robust clinical tools is frequently hindered by poor generalizability across heterogeneous data sources [8]. This limitation became acutely evident when applying our established 3D-CNN to a second, independent dataset (D2). Despite comparable training metrics, the model's predictive performance degraded dramatically to an AUC of 0.48. Such discrepancies in biomedical machine learning are typically driven by underlying domain shifts, such as variations in sensor hardware, acoustic environments, and patient demographics, that alter the underlying feature distributions [8]. Therefore, this study aims to systematically quantify the domain shift between D1 and D2 to identify the specific factors driving this discrepancy in generalizability, ultimately laying the groundwork for more robust, cross-dataset PH screening methodologies.

2. Materials and Methods

2.1. Datasets

This study utilizes two distinct datasets to evaluate cross-domain generalizability. Dataset D1 was recorded on sedated patients in a controlled environment. Dataset D2 was recorded in a more typical setting, resulting in

higher noise both in the signals and in the labels.

D1. The primary dataset is a private cohort comprising 42 subjects (29 diagnosed with PH, 13 without PH) acquired at Centro Hospitalar Universitário do Porto, Portugal. Ground truth labels were established via RHC, following standard clinical guidelines (mean pulmonary arterial pressure > 25 mmHg or systolic pressure > 30 mmHg). Heart sounds were captured over 5-minute intervals using a custom stethoscope head connected to a Rugloop Waves system at a sampling frequency of 8 kHz. Recordings were performed on sedated patients.

D2. The second dataset consists of recordings from 93 patients admitted to the Cardiology Department of the Unidade Local de Saùde de Gaia e Espinho (ULSGE), Portugal. 30-second recordings were performed using a Rijuven CardioSleeve digital stethoscope located on the pulmonary auscultation area, with a sampling frequency of 3 kHz. Ground truth for D2 was determined either from RHC when available, (20 patients, mean pulmonary arterial pressure > 20 mmHg), or from echocardiography (flow velocity across the tricuspid valve > 2.8 m/s). This results in 27 PH patients and 66 no-PH patients.

2.2. PH detection using 3D-TFD-CNN

The PH detection pipeline is object of our previous work [7]. Following our previously validated pipeline, all signals were downsampled to 1 kHz and S2 segments of 200 ms were extracted from the phonocardiograms using a Hidden Markov Model-based segmentator [9]. Each S2 segment was transformed into the time-frequency domain via the Choi-Williams distribution (CWD). This yielded an image per beat capturing frequencies up to 50 Hz. These beat-wise representations were stacked into 3D matrices (time-by-frequency-by-heartbeat) to preserve information about the inter-beat temporal variability.

The classification model is a 3D-CNN designed to capture spectro-temporal and beat-to-beat features. The network consists of two 3D convolutional blocks (16 and 64 filters, respectively, with 3D 3-by-3-by-3 kernels, ReLU activation, batch normalization, and max-pooling), followed by two fully connected layers (128 and 64 units) with 50% dropout, and a final sigmoid output layer.

2.3. Transfer learning strategy

First, the PH detection model was applied to each dataset independently using a leave-one-out (LOO) validation scheme. After assessing the performance drop from D1 to D2, a transfer learning protocol was implemented to determine whether robust, domain-invariant features learned from D1 could successfully transfer to D2. First, the 3D-CNN pre-trained on D1 was evaluated directly on D2 (zero-shot transfer). Subsequently, fine-tuning was applied. During fine-tuning, the convolutional feature extractors were frozen while the fully connected layers were re-initialized and trained. Again, a LOO strategy was employed due to the limited size of the datasets. In total, four experiments were conducted:

1) Train on D1, test on D1 with LOO
2) Train on D2, test on D2 with LOO
3) Train on D1, zero-shot transfer to D2
4) Train on D1, fine-tune on D2 and test with LOO

2.4. Interpretability via 3D Grad-CAM

To understand the morphological focus of the network and how it degrades under domain shift, Gradient-weighted Class Activation Mapping (Grad-CAM [10]) was employed. Grad-CAM utilizes the gradients of the PH class flowing into the final 3D convolutional layer to produce a coarse localization map, highlighting the regions of the time-frequency-beat volume most critical for the model's prediction. For each patient, the 3D Grad-CAM heatmaps were generated to visualize the model's focal points. Projections along the frequency vs time, heartbeat vs time, and heartbeat vs frequency planes were extracted to generate patient-level explainability reports.

2.5. Input-activation alignment analysis

To systematically quantify the generalization gap, we analyzed the spatial alignment between the original S2 TFD input energies and the resulting Grad-CAM activations. Alignment was assessed using two metrics:

Pearson Correlation. Used to measure the linear relationship between the continuous input pixel intensities and the corresponding Grad-CAM activation weights.

Intersection over Union (IoU). Used to evaluate the spatial overlap of the most relevant features. The input TFD volume was binarized at 10% of its maximum intensity, to isolate S2 acoustic energy from background noise. Conversely, the Grad-CAM volume was binarized at 50% of its peak activation to isolate the model's main decision focus. The IoU was then calculated as the area of overlap divided by the area of union between the thresholded input and the thresholded activation map.

By comparing these alignment metrics between D1 and D2 in our four experiments, we objectively quantified whether the performance drop arises from the model shifting its focus away from the S2 acoustic signatures.

3. Results

The 3D-CNN demonstrated strong discriminatory ability (AUC = 0.89) when trained and evaluated on D1, with a huge performance drop when trained and evaluated on D2 (AUC = 0.48). When the D1-trained model was applied directly to D2 without any weight updates, it

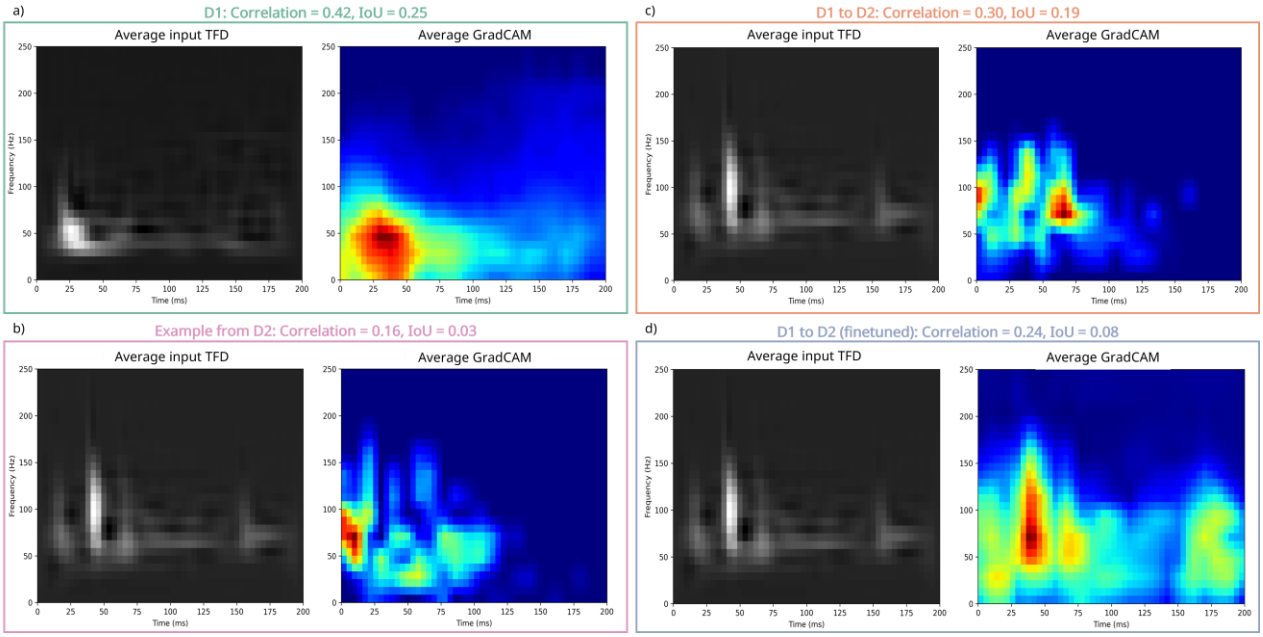

Figure 1. Example of Grad-CAMs on two example subjects, one from each dataset. Images report the average over the heartbeats, for both the input TFD volume and the 3D Grad-CAM when the model is: a) trained and evaluated on D1, b) trained and evaluated on D2, c) trained on D1 and evaluated on D2 (one-shot transfer learning), and d) trained on D1 and evaluated on S2 (with finetuning). Panels b, c and d refer to the same subject belonging to D2.

achieved an AUC of 0.66. Attempting to adapt the model to the new domain via fine-tuning resulted in a drop in performance again, yielding an AUC of 0.50, comparable to a model trained from scratch on D2. The superiority of direct transfer over fine-tuning highlights a severe domain adaptation challenge, suggesting that fine-tuning aggressively overfits back to domain-specific noise rather than the clinically relevant S2 features learnt from D1.

Grad-CAM analysis provided more insights into the observed phenomenon. As it can be observed in Figure 1, reporting one sample subject from each dataset, in D1, the 3D-CNN correctly localized its attention on the spectro-temporal signature of the S2 region. However, when evaluated on D2, the activation maps were more scattered and unfocused, revealing that the model relied on non-generalizable shortcuts. It can be further observed that in the one-shot transfer learning experiment from D1 to D2 the focus of the model is more consistent with the input energy, but it tends towards scattering again in finetuning.

This visual assessment was corroborated by the quantitative input-activation alignment metrics. As illustrated in Figure 2, both correlation and IoU showed a positive association with predictive performance, with the highest alignment observed in the baseline D1 experiment. Conversely, the lowest metrics were observed for the D2 internal-training experiment, directly mirroring the poor AUC scores. This indicates that the domain shift causes a spatial decoupling between the acoustic energy of the S2 input and the features the neural network uses for classification. No significant difference was found between the two classes, as reported in Table I.

4. Discussion and conclusions

The deployment of deep learning models into clinical practice is often limited by their inability to generalize across heterogeneous datasets. In this study, we analyzed and quantified the vulnerability of deep learning-based PH detection to domain shifts. While our 3D-CNN model demonstrated robust predictive capabilities on a clean, good-quality dataset, its performance severely degraded when exposed to a second, independent dataset recorded in a more typical screening scenario, with higher noise in

Table 1. Average metrics of input-activation alignment differentiating the two classes (mean ± standard deviation).

Exp.	Correlation			IoU		
	All	PH = 0	PH = 1	All	PH = 0	PH = 1
D1	.41 ± .06	.43 ± .08	.40 ± .05	.26 ± .07	.26 ± .10	.27 ± .05
D2	.17 ± .13	.16 ± .13	.20 ± .13	.04 ± .06	.04 ± .06	.06 ± .06
D1toD2	.34 ± .08	.34 ± .08	.34 ± .09	.22 ± .07	.22 ± .07	.24 ± .07
D1toD2 (with ft)	.34 ± .17	.34 ± .17	.34 ± .17	.14 ± .12	.14 ± .12	.13 ± .12

both signals and labels. Standard fine-tuning strategies only partially recovered this, suggesting a fundamental mismatch in the feature distributions in the two domains. It should be noted that the domain shift was not notable by applying descriptive statistics to the two input sets.

By comparing the model activation, obtained through 3D Grad-CAMs, and the original input, where higher energy is located in correspondence of the physiological S2, we provided a quantitaive method to analyse this performance drop and the domain shift. The quantitative analysis revealed that the model's failure on D2 was directly assoicated with a spatial decoupling between input and model's activation: instead of focusing on the clinically relevant time-frequency acoustic signatures of S2, the model learned to exploit non-generalizable, domain-specific shortcuts in the new dataset. The strong positive association between the alignment metrics and the AUC scores confirms that correlation and IoU are highly effective proxy metrics for quantifying detrimental domain shifts in biomedical signal processing.

These findings highlight a critical vulnearability in DL-based heart sound analysis which should drive future research towards external validation. Our results provide quantitative motivation for moving beyond standard transfer learning and considering the integration of explicit domain adaptation techniques to induce DL models to focus on physiological biomarkers rather than environmental or hardware-induced noise, even when applied to noisy datasets. We believe that what presented helps research into acoustic-based PH screening towards safe and reliable real-world clinical deployment.

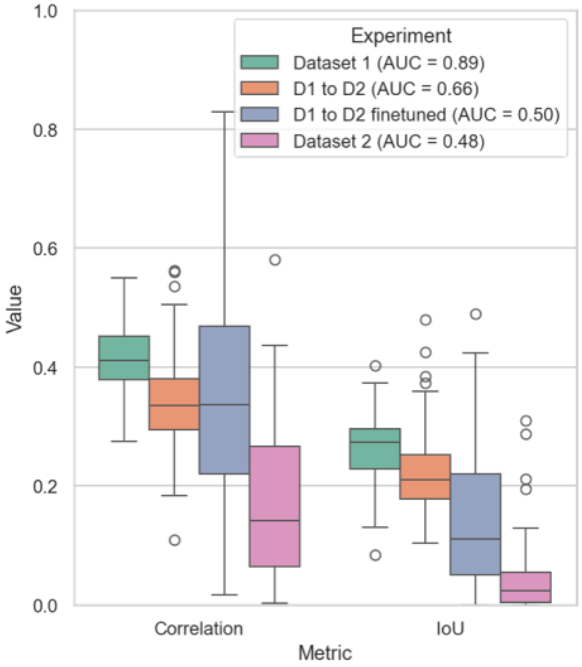

Figure 2. Classification metrics and input-activation alignment metrics for the four tested setups.

Acknowledgments

The work was co-funded by the European Union via MSCA Postdoctoral Fellowship (GA 101207722) and by the European Regional Development Fund (ERDF) through the Innovation and Digital Transition Programme (COMPETE 2030) under Portugal 2030 and by National Funds through the FCT - Fundação para a Ciência e a Tecnologia, I.P. (Portuguese Foundation for Science and Technology) within project PULSE, with reference 17172 (COMPETE2030-FEDER- 00864900).

References

[1] E. M. T. Lau, M. Humbert, and D. S. Celermajer, "Early detection of pulmonary arterial hypertension," *Nat Rev Cardiol*, vol. 12, no. 3, pp. 143–155, Mar. 2015.

[2] M. Humbert *et al.*, "2022 ESC/ERS Guidelines for the diagnosis and treatment of pulmonary hypertension," *Eur Heart J*, vol. 43, no. 38, pp. 3618–3731, Oct. 2022.

[3] J. Xu, L.-G. Durand, and P. Pibarot, "A new, simple, and accurate method for non-invasive estimation of pulmonary arterial pressure," *Heart*, vol. 88, pp. 76–80, Jul. 2002.

[4] B. Ge, H. Yang, P. Ma, T. Guo, J. Pan, and W. Wang, "Detection of pulmonary arterial hypertension associated with congenital heart disease based on time–frequency domain and deep learning features," *Biomedical Signal Processing and Control*, vol. 81, p. 104451, Mar. 2023.

[5] L. Guo *et al.*, "Development and Evaluation of a Deep Learning–Based Pulmonary Hypertension Screening Algorithm Using a Digital Stethoscope," *JAHA*, vol. 14, no. 3, p. e036882, Feb. 2025.

[6] A. Gaudio, N. Giordano, M. Elhilali, S. Schmidt, and F. Renna, "Pulmonary Hypertension Detection From Heart Sound Analysis," *IEEE Trans. Biomed. Eng.*, vol. 72, no. 10, pp. 2902–2914, Oct. 2025.

[7] N. Giordano, A. Gaudio, S. Emil Schmidt, and F. Renna, "Impact of the Input Representation on Pulmonary Hypertension Detection from Heart Sounds through CNNs," in 2025 Computing in Cardiology, Sep. 2025.

[8] B. Wan, B. Caffo, and S. S. Vedula, "A Unified Framework on Generalizability of Clinical Prediction Models," *Front. Artif. Intell.*, vol. 5, p. 872720, Apr. 2022.

[9] M.L. Martins, M.T. Coimbra, and F. Renna, "Markov-based neural networks for heart sound segmentation: Using domain knowledge in a principled way," *IEEE J Biomed Health Inform*, vol. 27, no. 11, pp. 5357–5368, Nov. 2023.

[10] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice: IEEE, Oct. 2017, pp. 618–626.

Address for correspondence:

Noemi Giordano
Department of Health Science and Technology, Aalborg University, Selma Lagerløfs Vej 249, 9260 Gistrup, Denmark.
nogi@hst.aau.dk