Impact of the input representation on pulmonary hypertension detection from heart sounds through CNNs

Noemi Giordano[1,2,3], Alex Gaudio[4], Samuel Schmidt[1], Francesco Renna[2]

[1] University of Aalborg, Denmark
[2] INESC TEC, Faculdade de Ciências da Universidade do Porto, Portugal
[3] Politecnico di Torino, Italy
[4] Johns Hopkins University, Baltimore, MD, US

Abstract

Pulmonary hypertension (PH) is a hemodynamic condition describing elevated pulmonary artery pressure. To date, right heart catheterism is the gold standard diagnostic test for PH, but it is an invasive and expensive procedure. Deep learning (DL) techniques applied to heart sounds have previously shown promising performances for PH screening. In this work, we analyze the impact of different input representations for PH detection with convolutional neural networks (CNNs). We found that considering each heartbeat as an independent input yielded systematically lower performance than considering the recordings as a whole: preserving the information about the variability over the heartbeats is key. Time-domain feature maps outperformed handcrafted features and combining the time- and frequency-domain proved consistently most effective. Reducing the number of heartbeats to 30 did not affect the performance, and even reducing to 10 beats preserves the diagnostic value. The proposed analysis moves one step further the applicability of DL for PH detection from heart sounds in the clinical practice.

1. Introduction

Pulmonary hypertension (PH) is a hemodynamic condition, involving an increase of the pressure in the pulmonary artery and the right ventricle. The prevalence of PH was estimated around 1% [1], but many cases may be missed due to the coexistence of comorbidities in conjunction with the lack of appropriate screening [1], [2]. A delayed diagnosis was proven to reduce the 5-year survival rate by almost 50% [3]. According to guidelines, the gold standard test for PH diagnosis is Right Heart Catheterism (RHC) [1]. Given its complexity, invasiveness and cost, though, only critical patients undergo RHC, leaving a gap for screening technologies. To date, echocardiography is the main PH screening tool, but it depends on the detection and analysis of a tricuspid regurgitation jet, which was found to have a low negative predictive value for PH [4], [5]. Auscultation may offer a promising complement or alternative to echocardiography as a screening tool for PH. The ease of use, portability and low-cost of heart sound technology may improve the screening options in low-resource scenarios and enable domiciliary screening. Deep learning (DL) techniques applied to heart sounds have previously shown promising performances for PH screening [6], [7], [8], [9]. Nevertheless, the application of DL methods to signals is not straightforward and the input representation may impact the performances of the model. In the literature, the effect of input representation for PH detection from heart sounds based on DL is still widely unexplored.

In this work, we focus on Convolutional Neural Networks, a family of DL models which previously showed promising performances for the task of interest [6], [7]. CNNs are designed to work on images: multiple options may be devised to translate heart sounds into an image to be fed as input to the CNNs. The goal of this work is to explore the effect of input representation on the performances of the model for PH detection.

2. Methods

Our pipeline has three main steps: (a) feature extraction, for translating the heart sound recordings into an appropriate input for the model; (b) DL modelling; and (c) empirical validation. The next paragraphs will present details for each step.

2.1. Feature extraction

CNNs were originally designed to process image data, where the spatial arrangement of pixels encodes meaningful patterns. This spatial consistency allows CNNs to effectively capture local features, such as edges, textures, and shapes, which are critical for image analysis. Spatial consistency can also be leveraged for signal analysis, provided that the signal is mapped into an

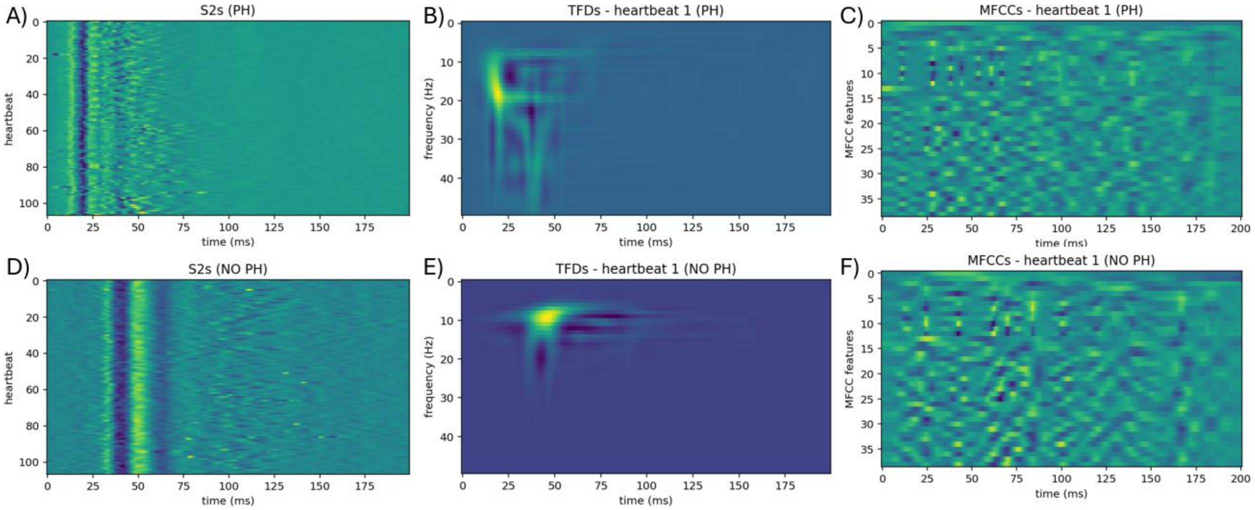

Figure 1. Examples of the three proposed feature maps for a patient with PH (A-C) and without PH (D-F).

appropriate feature map where local features convey relevant information for the downstream task.

In this work, we tested three different types of feature maps, meant to translate the signal-domain information into an image-domain input for the model. All three approaches have previously shown promising performances for similar tasks [8], [10], [11]. All three feature maps are based on 200-millisecond-long second heart sound (S2) segments. The segments were obtained from the original recordings by leveraging a DL-based Hidden Markov Model [12].

Feature map 1: S2s. All the S2 segments from the a given recording were stacked to create a 2D matrix. When visualized as an image, the x axis represents the time, the y axis the heartbeat index. In this sense, the representation is fully in the time-domain. Each image has a size of N by 200, where N is the number of heartbeats.

Feature map 2: TFDs. Each S2 segment was represented in the time-frequency domain. This was performed using the Choi-Williams distribution (CWD), an improved alternative to the Wigner-Ville distribution (WVD) that reduces cross-term interference by applying an exponential kernel in the ambiguity domain [13]. CWD was applied separately to each S2 segment, thus obtaining an image per beat with the x axis representing time, and y axis representing frequency. Each image has a size of 50 by 200, with the frequency bandwidth spanning from DC to 50 Hz. Images representing different heartbeats from the same recordings could be stacked together into a N-by-50-by-200 3D matrix.

Feature map 3: MFCCs. Mel Frequency Cepstral Coefficients (MFCC) were extracted from each S2 segment to capture perceptually relevant spectral features. 13 MFCCs were computed, along with their first-order and second-order temporal derivatives to incorporate dynamic information. In this way, an image per beat was obtained with the x axis representing time and the y axis representing the MFCC features. Each image has a size of 39 by 200. Also in this case, images representing different heartbeats from the same recording could be stacked together into a N-by-100-by-200 3D matrix.

Figure 1 shows an example of the three feature maps for a sample recording.

2.2. Deep Learning modelling

Two different CNN architectures were designed and tested for the PH detection task.

Model 1: 2D-CNN. The 2D-CNN model has a traditional 2D architecture composed of five convolutional blocks. The first 2D convolutional layer uses 32 3x3 filters, followed by batch normalization, 2x2 max-pooling layer and spatial dropout. Subsequent convolutional blocks increase the number of filters to 64, 128, 256, and 512. All convolutional layers use the ReLU activation. The output of the final convolutional layer is flattened and passed through three fully connected layers of sizes 512, 256, and 128, each followed by a 50% dropout layer. The output layer has a sigmoid activation function, returning the estimated probability of the PH class. Each image is meant to be provided to the model as a separate input: S2 segments get the same label as their parent recording and are treated as independent. The advantage is that the size of the dataset increases two orders of magnitude. The total number of trainable parameters is 203,979,457.

Model 2: 3D-CNN. The 3D-CNN model is a 3D architecture designed to capture spatial features across multiple dimensions. It begins with a 3D convolutional layer employing 16 filters of size 3x3x3 with ReLU activation, followed by batch normalization, a 3D 2x2x2 max-pooling layer, and spatial dropout. This is followed by a second 3D convolutional block with 64 filters, again using ReLU activation, followed by batch normalization. The output is flattened and passed through two fully connected layers of sizes 128 and 64, each followed by a dropout layer with a 50% dropout rate. The final output layer leverages a sigmoid activation function to the estimated probability of the PH class. Images from the same recording are stacked and treated as a single input. The advantage is that the information regarding the variability of the sounds over the heartbeats is preserved. The total number of trainable parameters is 74,731,425.

For both 2D-CNN and 3D-CNN, weights were randomly initialized (no pre-training) and optimized using Adam with a learning rate of 1e-5 and binary cross-entropy as loss function.

2.3. Validation

Model performance was assessed using a bootstrapped 5-fold cross-validation strategy. In each iteration, three folds were used for training, one for internal validation, and one for testing on unseen data. Stratified sampling was applied to ensure that each fold preserves the original proportion of patients with and without PH. Folds were constructed at the patient level to ensure that all heartbeats from the same patient were assigned to the same fold. This process was repeated 12 times using different random partitions to enhance robustness. For each repetition, the area under the receiver operating curve (AUC) was computed, and overall performance was summarized as the micro-averaged AUC across all folds, with its standard deviation over the 12 repetitions.

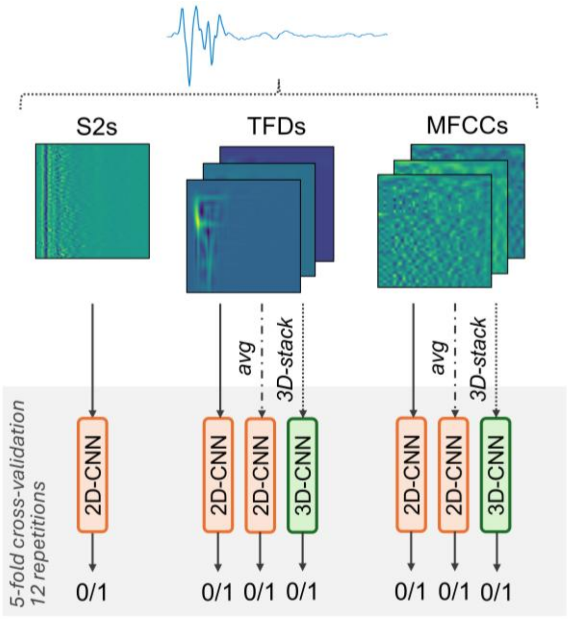

Figure 2. Graphical representation of the experiments conducted.

2.4. Dataset

We performed our experiments on a private dataset including 42 subjects (29 with, 13 without PH), acquired at Centro Hospitalar Universitario do Porto, Portugal [8]. Ground truth pulmonary artery pressure was assessed by RHC: patients with a mean PAP higher than 25 mmHg or a systolic PAP higher than 30 mmHg were labelled as PH, as per current guidelines [1]. Heart sounds were recorded using a custom stethoscope head connected to a Rugloop Waves system, with a sampling frequency of 8 kHz and a 16-bit dynamics. Recordings lasted 5 minutes.

3. Results

We designed three possible setups: using the 2D-CNN with each heartbeat as an independent input; using the 2D-CNN with the average heartbeat; using the 3D model with stacked images. All three setups were applied to TFDs and MFCCs. The first was applied to S2s as only one image describing the full recording is available. Figure 2 proposes a graphical representation and results are presented in Table I.

Table I. Area under the ROC curves (average ± standard deviation over 12 repetitions)

Input	3D (2D for S2s)	2D avg beat	2D single beats
S2s	**0.73 ± 0.08**	-	-
TFDs	**0.88 ± 0.04**	0.85 ± 0.04	0.64 ± 0.04
MFCCs	0.63 ± 0.04	**0.79 ± 0.06**	0.60 ± 0.05

For each setup, we also tested the effect of reducing the number of available heartbeats. This is relevant to define the boundaries of the method for real-life applications. Five-minute recordings, 30 heartbeats, and 10 heartbeats were tested. Results are shown in Figure 3.

4. Discussion and Conclusions

Three main aspects of the input representation were analyzed in this work: a) the nature of the feature map, b) the use of single heartbeats as independent inputs vs the use of the entire recording, c) the duration of the recordings, i.e., number of heartbeats.

We explored three of the most common feature maps previously showing promising results for CNN-based heart sounds classifiers. The three selected options convey different information of the sound: its morphology in the time-domain (S2s), its time-frequency behavior (TFDs), its perceptually relevant features (MFCCs). With MFCCs, leveraging the average heartbeat resulted in significantly better results than preserving the information of the single heartbeats, either as independent

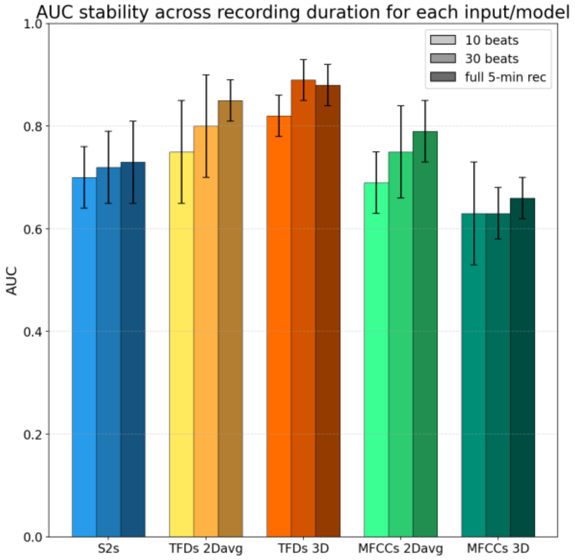

Figure 3. Variation of the average AUC after the reduction of the duration to 30 beats and to 10 beats.

inputs or together. Nevertheless, we found that the MFCCs were consistently suboptimal for the task of interest. The fact that perceptual features perform poorly is consistent with the complexity of detecting PH with traditional auscultation. For time-domain feature maps, considering each heartbeat as an independent input yielded systematically lower performance than using the recordings as a whole: preserving the information about the variability over the heartbeats is key. Previous literature shows that PH may provoke changes into time-related cardiac biomarkers [5]: this seems consistent with our findings. Combining time and frequency proved most effective showing that the key information for PH detection resides in the intersection between the two domains. This may open to further research concerning time- and frequency-acoustic biomarkers of PH.

We showed that reducing the number of heartbeats to 30 did not significantly affect the performances of the tested input/model combinations. We can conclude that a 30-second recording is sufficient for clinical applicability. Even reducing to 10 beats preserves the diagnostic value of the test, with an AUC of the best input/model combination (TFDs with 3D-CNN) higher than 80%. This is relevant in real-life situations where collecting long recordings is often complicated.

We believe that the proposed analysis clarifies the importance of input representation in DL and moves one step further the applicability of DL for PH detection from heart sounds in the clinical practice.

Acknowledgments

The work was funded by INESC TEC International Visiting Researcher Program and by the European Union with MSCA Postdoctoral Fellowship (GA 101207722).

References

[1] M. Humbert *et al.*, “2022 ESC/ERS Guidelines for the diagnosis and treatment of pulmonary hypertension,” 2022, *Oxford University Press*. doi: 10.1093/eurheartj/ehac237.

[2] M. M. Hoeper, H. A. Ghofrani, E. Grünig, H. Klose, H. Olschewski, and S. Rosenkranz, “Pulmonary hypertension,” *Dtsch Arztebl Int*, vol. 114, no. 5, pp. 73–84, Feb. 2017, doi: 10.3238/arztebl.2017.0073.

[3] E. M. T. Lau, M. Humbert, and D. S. Celermajer, “Early detection of pulmonary arterial hypertension,” Mar. 25, 2015, *Nature Publishing Group*. doi: 10.1038/nrcardio.2014.191.

[4] A. Frost *et al.*, “Diagnosis of pulmonary hypertension,” in *European Respiratory Journal*, European Respiratory Society, Jan. 2019. doi: 10.1183/13993003.01904-2018.

[5] 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, no. 1, pp. 76–80, 2002, doi: 10.1136/heart.88.1.76.

[6] 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,” *Biomed Signal Process Control*, vol. 81, Mar. 2023, doi: 10.1016/j.bspc.2022.104451.

[7] L. Guo *et al.*, “Development and Evaluation of a Deep Learning–Based Pulmonary Hypertension Screening Algorithm Using a Digital Stethoscope,” *Journal of the American Heart Association* , vol. 14, no. 3, Feb. 2025, doi: 10.1161/JAHA.124.036882.

[8] A. Gaudio, N. Giordano, M. Elhilali, S. Schmidt, and F. Renna, “Pulmonary Hypertension Detection from Heart Sound Analysis,” *IEEE Trans Biomed Eng*, 2025, doi: 10.1109/TBME.2025.3555549.

[9] M. Wang, B. Guo, Y. Hu, Z. Zhao, C. Liu, and H. Tang, “Transfer Learning Models for Detecting Six Categories of Phonocardiogram Recordings,” *J Cardiovasc Dev Dis*, vol. 9, no. 3, Mar. 2022, doi: 10.3390/jcdd9030086.

[10] T. Kaddoura *et al.*, “Acoustic diagnosis of pulmonary hypertension: Automated speech- recognition-inspired classification algorithm outperforms physicians,” *Sci Rep*, vol. 6, Sep. 2016, doi: 10.1038/srep33182.

[11] V. G. Andreev *et al.*, “Time–Frequency Analysis of The Second Heart Sound to Assess Pulmonary Artery Pressure,” *Acoust Phys*, vol. 66, no. 5, pp. 542–547, Sep. 2020, doi: 10.1134/S1063771020050012.

[12] 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, doi: 10.1109/JBHI.2023.3312597.

[13] H.-I. Choi and W. J. Williams, “Improved Time-Frequency Representation of Multicomponent Signals Using Exponential Kernels,” *IEEE Trans Acoust*, vol. 37, no. 6, pp. 862–871, 1989.

Address for correspondence:

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