Explainable Deep Learning for Non-Invasive Detection of Pulmonary Artery Hypertension from Heart Sounds

Alex Gaudio [1,2,3], Miguel Coimbra [2,3], Aurélio Campilho [2,3], Asim Smailagic [1], Samuel E Schmidt [4], Francesco Renna [2,3]

[1] Carnegie Mellon University, Pittsburgh, United States
[2] INESC TEC, Porto, Portugal
[3] University of Porto, Porto, Portugal
[4] Aalborg University, Aalborg, Denmark

Abstract

Late diagnoses of patients affected by pulmonary artery hypertension (PH) have a poor outcome. This observation has led to a call for earlier, non-invasive PH detection. Cardiac auscultation offers a non-invasive and cost-effective alternative to both right heart catheterization and doppler analysis in analysis of PH. We propose to detect PH via analysis of digital heart sound recordings with over-parameterized deep neural networks. In contrast with previous approaches in the literature, we assess the impact of a pre-processing step aiming to separate S2 sound into the aortic (A2) and pulmonary (P2) components. We obtain an area under the ROC curve of .95, improving over our adaptation of a state-of-the-art Gaussian mixture model PH detector by +.17. Post-hoc explanations and analysis show that the availability of separated A2 and P2 components contributes significantly to prediction. Analysis of stethoscope heart sound recordings with deep networks is an effective, low-cost and non-invasive solution for the detection of pulmonary hypertension.

1. Introduction

Pulmonary artery hypertension (PH) is an under-recognized disease, with unmet need for diagnostic and treatment recommendations in low and middle-income regions [1]. PH disease has high mortality rate and early detection in screening programs can improve outcomes.

Existing tools for PH detection are not well optimized for the needs of low and middle income regions. Right heart catheterization is a gold standard for PH detection, but it is highly invasive and not suitable for screening programs. Doppler echocardiography is widely used for clinical screening of PH, but the noisy nature of its measurements requires additional modalities to improve reliability [2, 3]. Ultrasound technology also requires a trained technician and expensive machinery [4]. Other tests helpful to PH detection include blood gas analysis and imaging from cardiac magnetic resonance, chest x-ray, and pulmonary angiography [5]. Automated PH detection using cardiac auscultation data recently emerged as a non-invasive and low cost alternative that can outperform physicians [6]. Our approach to analyze heart sounds with deep networks has low resource cost and is suitable for early screening.

Detection of PH from heart sounds focuses on an analysis of the second heart sound, S2, which itself consists of two mixed sound signals: the Aortic valve closure (A2) and the Pulmonic valve closure (P2) [7]. Peak-to-peak analysis, in the time domain, shows that patients with PH disease present with larger distance and larger difference in amplitude between the A2 and P2 peaks [6].

Automated diagnosis of PH from heartsound includes handcrafted analysis [8] and traditional machine learning [6, 9]. In a related area, application of deep Convolutional Neural Networks (CNNs) is useful in heart murmur detection in children [4] and heart sound segmentation [10]. Our approach adopts deep convolutional neural networks (CNNs) [11, 12], typically used for image analysis, for the analysis of audio data. Post-hoc explanations of CNN predictions, with methods like IntegratedGradients [13], facilitate transparency of the black box model.

The novelty in our approach to PH detection is to propose deep networks on small data. We demonstrate: a) Applying deep networks to analysis of heart sound recordings gives strong predictive performance; b) Post-hoc explanations verify the role of proposed A2 and P2 components in the second heart sound.

2. Methods

Data Acquisition: We acquire a private dataset of 42 patients at Centro Hospitalar Universitário do Porto, Portugal. Summary statistics in Table 1 show 29 patients with

Table 1. Dataset Summary

Population	Male	Female	Age	HR (bpm)
Has PH	9	20	60±17	70±10
No PH	8	5	57±10	69±8
All Patients	17	25	59±15	69±9

PH and 13 without PH. Of diseased patients, the majority (20 of 29) are female. The age and heart rates of both positive and diseased populations are similar. Inclusion and exclusion criteria are unknown. PH is defined as positive when a patient has a Mean Pulmonary Arterial Pressure (MPAP) above 25 mm Hg, or Pulmonary Arterial Systolic Pressure (PASP) above 30 mm Hg. For each patient, we obtain the ground truth pulmonary artery pressure from a right heart catheterization, and an accompanying five minute PCG heart sound recording. The recording was obtained in a relatively quiet clinical setting with the patient supine and at rest. Auscultation was performed over the second left intercostal space using a custom cable stethoscope connected to a Rugloop Waves system. Heart sounds were recorded at a sample rate of 8 kHz and their amplitudes were quantized with 16-bit resolution. The dataset is not published to preserve privacy.

Pre-Processing: In each five minute audio signal, we segment the heatbeats and extract a 200 ms window for each heartbeat's S2 sound, where start time of the window is chosen so the peaks of all S2 sounds for that patient are aligned in time. The S2 signal is filtered with second order Butterworth filters with cut-off frequencies of 25 Hz and 400 Hz, re-sampled to 1 kHz, cleaned by removing spikes via the method in [14], and separated into proposed A2 and P2 components according to [15]. Source separation assumes the Aortic and Pulmonic components maintain approximately the same waveform across heartbeats, and assumes the delay between the components within a heartbeat varies due to change in thoracic pressure at different respiratory phases. The two components are retrieved via alternating optimization of a least-squares problem.

Alignment and segmentation results in a multi-channel 2-D representation of the audio data containing S2, proposed A2, and proposed P2 components. Each 2-D channel has 200 columns (representing a 200 ms window) and as many rows as there are heartbeats. We then make channels for all patients of the same shape by zero padding to 454 rows, and independently normalize each of the three channels per patient to unit variance. Normalizing to unit variance helps stabilize gradient backprogation by reducing risk of vanishing or exploding gradients.

Deep CNN Models and Optimization: We consider DenseNet121, ResNet18 and EfficientNet-b0 architectures. Pre-trained deep network initialization can improve performance for small datasets. Random and ImageNet initializations were considered. We report DenseNet trained from random initialization, ResNet18 from ImageNet initialization and EfficientNet-b0 from standard adversarial ImageNet initialization. The models were all trained with batch Gradient Descent (learning rate 0.0001, momentum 0.5) for 150 epochs. Deep networks typically train on large datasets with stochastic minibatch gradient descent. To stabilize gradient updates, we use a batch size equal to the dataset size. The loss is weighted binary cross entropy with the positive class balancing weight $\frac{8+5}{20+9}$.

GMM and SVM Baseline: To benchmark the predictive performance of the deep networks against classical methods, we implement a Gaussian Mixture Model (GMM) and Support Vector Machine (SVM). Our GMM implementation adapts the state-of-the-art work of [6], where one GMM was trained for positive classes, and another for negative classes. The class of a test sample is the GMM model with higher posterior negative log likelihood. To get best performance with this baseline, we develop a different pre-processing pipeline, and accordingly optimized the GMM models to have two components and spherical covariance. The SVM uses an RBF kernel and slack parameter $C = 1$. For pre-processing, we used only the S2 channel. The addition of proposed A2 and P2 channels negatively impacts performance due to overfitting. Each of the heartbeats (each row of the S2 channel) was transformed with a 1-d Short Time Fourier Transform, using an FFT window of 64 samples and hop length of two samples, and computing the energy spectrum via absolute value. The patient data, a tensor of shape (H,33,101), was reduced to (33,101) by computing a 98% quantile over the H heartbeats. The channel was zero padded to 454 rows and normalized to unit variance, then flattened as a vector and subsequently passed to the SVM and GMM models.

Evaluation All models were evaluated using 10-fold stratified cross validation. To report performance, we store validation set prediction probabilities from each fold. There is one prediction probability for each patient. We report the area under the ROC curve (ROC AUC) and standard classification metrics. Classification metrics require choosing a threshold to convert the probabilities into classes. We choose a threshold T_k for each k^{th} fold that maximizes the difference of true positive rate minus the false positive rate on the k^{th} fold training set ROC curve. This threshold optimizes the training set balanced accuracy score. We compute validation performance within each fold and then aggregate the metrics by an average across folds and epochs 100 to 150.

Post-hoc Explainability To better understand which parts of the proposed A2 and proposed P2 channels contribute to PH detection, we apply the IntegratedGradients attribution method [13]. In particular, after training the DenseNet121 model on ten folds, we have ten independently trained models. Therefore, we compute ten attributions to each heartbeat in the dataset and then average

Table 2. Deep Networks Give State-of-the-art Results

Model	ROC AUC	MCC	BAcc	Precision	Recall
GMM	0.78	0.57	0.78	0.92	0.82
SVM	0.88	0.55	0.78	0.97	0.65
DenseNet121	**0.95**	**0.82**	**0.91**	0.96	**0.90**
EfficientNet-b0	0.93	0.79	0.90	**1.00**	0.81
ResNet18	0.92	0.53	0.77	0.88	0.59
DenseNet121 (S2)	0.93	0.69	0.85	0.94	0.81
EfficientNet-b0 (S2)	0.89	0.52	0.76	0.85	0.84

them to get one attribution per channel. For better visualization, the attribution is converted to a magnitude via absolute value and then clipped to 1% and 99% of its values. Clipping aids visualization because gradient-based attribution methods generate some outlier points.

3. Results

Deep Networks Improve Detection Performance. The results in Table 2 show that the DenseNet121 and EfficientNet-b0 deep networks outperform traditional machine learning models on the considered PH dataset by large margins. The DenseNet121 model has the highest performance of 0.95 ROC AUC, the highest Balanced Accuracy (BAcc), and highest Matthew's Correlation Coefficient (MCC). The two best performing models are DenseNet121 and EfficientNet-b0.

Separating S2 into A2 and P2 Improves Performance. The bottom rows of Table 2 show that availability of S2, A2 and P2 channels improves performance over using only the S2. A motivation of deep learning is to negate need for pre-processing via data-driven feature generation and larger datasets. In the small data regime, as is the case here, we observe that pre-processing improves performance. Moreover, the over-parameterized nature of deep networks requires rethinking traditional interpretations of underfitting and overfitting. Classical methods like the SVM and GMM overfit with additional parameters from the A2 and P2 channels while deep networks improve.

Proposed A2 and P2 Agree with Domain Knowledge. The top three rows of Figure 1 visualize one patient's heart sound data. Each line is a single heartbeat. The top row shows the S2 signal. The second and third rows show the proposed source separated signals A2 and P2. We visualize the signals after normalizing them to unit variance to represent the input as passed to the predictive model. We found empirically that the normalization improved performance; normalization makes the quieter P2 have similar amplitude to the louder A2. We observe the A2 signal is very clearly defined, due to the fact that the heartbeats have been aligned based on their peak. The distance between A2 and P2 components varies depending factors such as whether the patient is inhaling or exhaling, as well as presence of PH. Thus, current domain knowledge agrees with the visual that an average P2 signal should be less well located in time. In this patient, we observe the P2 has most varied behavior between 30 ms to 60 ms. Current domain knowledge expects PH to be related to changes in the timing and amplitude of the P2.

Post-Hoc Explanation Validates Domain Knowledge and Utility of A2 and P2 Segmentations. The bottom plot in Figure 1 shows the average attribution over all heartbeats and a 99.9% confidence interval. The attribution to P2 dominates for this patient, and also coincides with the period between 30 ms to 60 ms of most varied P2 behavior. Both observations suggest Deep Networks agree with domain knowledge. The attribution to A2 is strongest at the peak, just before 25 ms. Attribution shows the availability of separated components facilitates prediction.

4. Conclusions

Our main contribution is to advance the state-of-the-art in automated detection of pulmonary artery hypertension from heart sounds. We show that deep networks trained on a private dataset of pre-processed digital stethoscope recordings achieve ROC AUC scores of 0.95 and 0.93, giving improvements of +.17 and +.15 over our adaptation of a previous state-of-the-art based on a Gaussian Mixture Model, and improvements of +.07 and +.05 over our best traditional machine learning implementation. Post-hoc explanations and improved performance show that the separation of the S2 sound into proposed A2 and P2 components aids detection. Analysis of stethoscope heart sound data with deep networks is an effective, low-cost and non-invasive solution for detection 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.

References

[1] Hasan B, Hansmann G, Budts W, Heath A, Hoodbhoy Z, Jing ZC, Koestenberger M, Meinel K, Mocumbi AO, Radchenko GD, et al. Challenges and special aspects of pulmonary hypertension in middle-to low-income regions: Jacc state-of-the-art review. Journal of the American College of Cardiology 2020;75(19):2463–2477.

[2] Lau EM, Humbert M, Celermajer DS. Early detection of pulmonary arterial hypertension. Nature Reviews Cardiology 2015;12(3):143–155.

[3] Taleb M, Khuder S, Tinkel J, Khouri SJ. The diagnostic accuracy of Doppler echocardiography in assessment of pulmonary artery systolic pressure: A meta-analysis. Echocardiography 2013;30(3):258–265.