Multi-Encoder Autoencoders for Unsupervised Separation of the Second Heart Sound: An Application to Pulmonary Hypertension Diagnosis

Francisco da Ana[1], Noemi Giordano[2], Miguel Martins[1], Francesco Renna[1]

[1] INESC TEC, Faculdade de Ciências da Universidade do Porto, Porto, Portugal
[2] Aalborg University, Aalborg, Denmark

Abstract

Aims: *The second heart sound (S2) provides clinically relevant information for the diagnosis of pulmonary hypertension (PH) through the relative timing and amplitude of its aortic (A2) and pulmonary (P2) components. Algorithms for estimating these components traditionally rely on iterative optimization frameworks with predefined assumptions about their waveforms. This work proposes a data-driven alternative based on a multi-encoder convolutional autoencoder that learns to separate single-channel S2 mixtures in an unsupervised process.* ***Methods:*** *Two parallel encoder branches share one single decoder, and each encoder is encouraged to specialize in one source subspace via a custom loss function. The model was assessed on 12,831 S2 sounds from 42 patients (29 with PH) diagnosed by right heart catheterization, using eight statistical features aggregated per patient to train a classifier under nested leave-one-out cross-validation.* ***Results:*** *The best run achieved an AUROC of 0.836 and an average precision of 0.938, competitive with an established alternating-optimization baseline.* ***Conclusion:*** *These results suggest that deep learning-based blind source separation of the second heart sound can support the diagnosis of PH with a discriminative acoustic biomarker.*

1. Introduction

Pulmonary hypertension (PH) is a pathophysiological disorder characterized by elevated pressure in the pulmonary artery, affecting both cardiac and pulmonary function, which in an advanced stage may lead to right heart failure and, ultimately, death. Its gold-standard method for clinical diagnosis is right heart catheterization which, although highly precise, is naturally invasive and requires specialized equipment and trained operators, limiting its accessibility for widespread screening. Transthoracic echocardiography is non-invasive but operator-dependent, leaving a cost-effective acoustic route to early detection as an open clinical need [1].

The second heart sound (S2) results from two nearly simultaneous events: the closure of the aortic (A2) and of the pulmonary (P2) valves, commonly separated in healthy individuals by only 20 to 80 ms [2]. In general, A2 occurs earlier and louder, with a P2/A2 energy ratio of about 0.22 [3]. In the presence of PH, these two properties tend to be altered: the time split widens and P2 becomes stronger, motivating the study of these features as potential acoustic biomarkers for the disease [2, 3].

End-to-end classifiers can detect PH from heart sounds with strong accuracy, yet they act as a "black box", giving no physiological explanation for the decision [1]. More interpretable strategies in the literature involve systems that separate A2 and P2 before PH classification, either fitting mathematical waveform templates [2, 4] or running a per-signal iterative optimization [5], then performing PH detection from features extracted from the separated components [5].

This work adapts the multi-encoder autoencoder of Webster and Lee [6] to the second heart sound, as a new, data-driven approach to the separation of the A2 and P2 components. The proposed model is assessed by testing whether its two branches specialize, without supervision, into waveforms consistent with the two valve closures, and whether the separated components support PH detection on a clinical cohort referenced against right heart catheterization. The comparison of the diagnostic task against the iterative baseline of Renna et al. [5] uses the same biomarkers and classifiers, isolating the effect of the separation stage.

2. Methods

2.1. Problem formulation

The task is addressed as a single-channel blind source separation problem. Formally, an observed S2 segment is modelled as the superposition of two physiological sources plus residual noise,

$$x_{S2}(t) = s_{A2}(t) + s_{P2}(t) + n(t), \tag{1}$$

where $s_{A2}(t)$ and $s_{P2}(t)$ are the waveforms of the aortic and pulmonary components. They both must be recovered from x_{S2} alone, since on real recordings the isolated components are never observed and no training objective can compare an estimate with its true source.

2.2. Data

The clinical dataset was collected at Centro Hospitalar Universitário do Porto, Portugal, with the heart sounds acquired by a custom stethoscope connected to a Rugloop Waves system at a sampling frequency of 8 kHz. These signals belong to 42 patients, 29 of them were diagnosed with PH by right heart catheterization, defined as a mean pulmonary arterial pressure above 25 mmHg or a systolic pulmonary arterial pressure above 30 mmHg. The remaining 13 patients were healthy individuals, with respect to PH. There are in total 12,831 S2 sounds, segmented by the U-Net-based model of Renna et al. [7], each spanning 200 timesteps resampled at 1 kHz, peak-normalized and front-zero-padded to 256 (a length compatible with the six pooling stages of the encoder). The unsupervised separation model is trained on the entire dataset; the patient-level cross-validation used for the diagnostic evaluation is described in Section 2.6.

2.3. Multi-encoder convolutional autoencoder

The separation model is a multi-encoder convolutional autoencoder: instead of coupling one encoder to the decoder, one parallel branch per source expected in the mixture feeds a single shared decoder [6], which here amounts to two branches, one for each valve closure. Formally, each encoder maps the mixture to a latent vector, $\mathbf{z}_i = E_i(\mathbf{x}_{S2})$ for $i \in \{1, 2\}$, and the decoder reconstructs the mixture from their concatenation, $\hat{\mathbf{x}}_{S2} = D([\mathbf{z}_1; \mathbf{z}_2])$.

The parallel pathways act as an architectural inductive bias towards specialized source representations, and the channel budget is divided evenly across branches, keeping the parameter count close to that of an equivalent single-encoder autoencoder. Figure 1 illustrates both regimes.

Each encoder stacks convolutional blocks (Conv1d $\rightarrow$ ReLU $\rightarrow$ BatchNorm $\rightarrow$ Conv1d $\rightarrow$ ReLU $\rightarrow$ MaxPool1d $\rightarrow$ BatchNorm) whose channels progress as $[16, 32, 64, 128, 256, 256]$. All convolutions use a kernel of 21 samples, and a final 1×1 convolution compresses each branch to its share of a latent width of 32. The decoder mirrors this structure with upsampling and transposed convolutions, applying group normalization with two groups so that each encoder's channel block is normalized independently, reinforcing the separation between pathways.

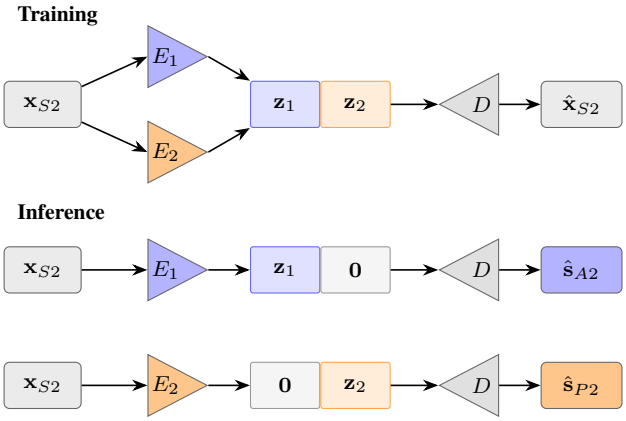


Figure 1: Multi-encoder convolutional autoencoder with one branch per valve closure. Training reconstructs $\mathbf{x}_{S2}$ from the concatenated latents; inference isolates one source by zeroing the other.

2.4. Training objective

The baseline loss function of this architecture [6] combines the mean squared error between the jointly decoded latent ($\hat{\mathbf{x}}_{S2}$) and the mixture ($\mathbf{x}_{S2}$) with three regularizers: an L_1 penalty $\mathcal{L}_{\text{sep}}$ pushing the decoder weights towards a block-diagonal structure, a zero-reconstruction term $\mathcal{L}_{\text{zero}} = \|D(\mathbf{0})\|^2$ that enforces the decoder to generate silence from a null latent, and an L_2 penalty $\mathcal{L}_{\text{enc}}$ on the latents to keep their values stable. All these terms act on the joint reconstruction, the decoder weights or the latent space, so the individual estimates are never expressed in the objective. With this training objective, two degenerate solutions were commonly reached in practice: "single-branch collapse", where one component captures the whole mixture while the other is silenced, and "redundant overlap", where both encoders return scaled copies of the mixture whose sum reconstructs the input, yet neither isolates the expected distinct sources.

Three domain-informed terms are therefore added, taking advantage of the physiological constraints from this specific application, and joined to the inherited regularizers in the objective

$$\begin{aligned}\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{add}} + \lambda_{\text{sep}}\mathcal{L}_{\text{sep}} + \lambda_{\text{zero}}\mathcal{L}_{\text{zero}} + \lambda_{\text{enc}}\mathcal{L}_{\text{enc}} \\ + \lambda_{\text{null}}\mathcal{L}_{\text{null}} + \lambda_{\text{decor}}\mathcal{L}_{\text{decor}},\end{aligned} \tag{2}$$

in which $\lambda_{\text{sep}} = \lambda_{\text{decor}} = 0.5$, $\lambda_{\text{zero}} = 0.05$, $\lambda_{\text{enc}} = 10^{-4}$ and $\lambda_{\text{null}} = 1.0$.

The additive reconstruction $\mathcal{L}_{\text{add}} = \|D([\mathbf{z}_1; \mathbf{0}]) + D([\mathbf{0}; \mathbf{z}_2]) - \mathbf{x}_{S2}\|^2$ replaces the original reconstruction term, encoding the fact that S2 is the sum of the two valve closures rather than trusting a decoding of the concatenated latents. The other two act on the estimates themselves, one

against each degenerate solution. $\mathcal{L}_{\text{null}}$ is an anti-collapse hinge, $\mathrm{ReLU}(\varepsilon - \mathrm{E_i}/\mathrm{E_{S2}})^2$ summed over both branches, whose energy floor $\varepsilon = 0.1$ sits deliberately below the typical ratio of P2 because both valve closures are always present, so neither estimate may vanish. $\mathcal{L}_{\text{decor}}$ penalizes the squared cosine similarity between the estimates: cosine is scale-invariant, so the A2–P2 amplitude gap does not dominate, enforcing the waveforms to follow distinct temporal patterns.

Training runs 250 epochs of Adam at a learning rate of 10^{-3}, with batch size 256, gradients clipped at 0.5 and the final-epoch checkpoint retained, every configuration being trained from scratch across ten seeds with deterministic initialization parameters.

2.5. Inference and source labelling

At inference the decoder is fed one branch at a time, with the other latent vector replaced by zeros, so that $D([\mathbf{z}_1; \mathbf{0}])$ and $D([\mathbf{0}; \mathbf{z}_2])$ return one estimate each in two forward passes. The branches are not tied to a physiological component during training, so each estimate is labelled by the time-weighted centroid of its homomorphic envelope, obtained with a second-order Butterworth low-pass filter at 50 Hz. The centroid measures when, on average, the source releases its energy, and since the aortic valve is expected to close first, the earlier one is labelled A2.

2.6. Pulmonary hypertension detection

To assess the utility of the proposed separation method for PH diagnosis, three biomarkers are extracted from every separated beat: the A2–P2 split, taken as the distance between the two envelope centroids, the energy ratio E_{P2}/E_{S2} and the peak-to-peak amplitude ratio $\mathrm{PtP_{P2}}/\mathrm{PtP_{S2}}$. Since the label is defined per patient, the beat-wise series are collapsed into an eight-dimensional descriptor: mean, standard deviation, minimum and maximum of the split, and mean and standard deviation of each ratio.

The descriptors are classified by two class-weighted models, a support vector machine with a radial basis function kernel and a random forest, under a nested leave-one-patient-out cross-validation. The outer loop holds out one patient at a time, so each of the 42 is predicted exactly once by a model fitted on the other 41. Inside each fold, the standardization is estimated on those 41 patients and the hyperparameters are chosen by a grid search scored with 3-fold cross-validated AUROC. The decision threshold is set afterwards, as the cut-off maximizing the macro-averaged F1 score of the selected model on the same 41 patients, leaving the held-out one to be scored and classified without contributing to any of these choices. The classification performance is reported as AUROC, average precision, balanced accuracy and macro-averaged F1. The baseline for comparison is given by the identical classification pipeline, fed instead by the estimates of the alternating-optimization separation of [5].

3. Results

Since the isolated components are never observed in clinical recordings, the assessment of the separation is necessarily qualitative. Figure 2 shows the encoders specializing as intended on a real S2 beat: one branch isolating an earlier and louder transient consistent with A2, the other a later, weaker and more oscillatory waveform consistent with P2. These separation patterns showed to be robust across the dataset.

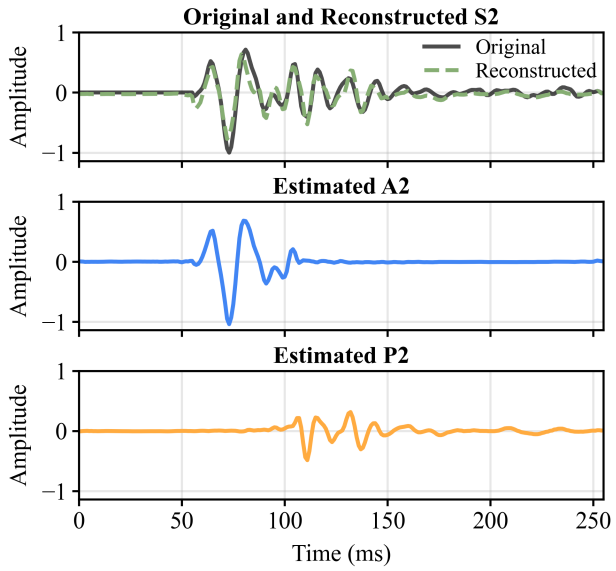


Figure 2: Separation of a real S2 beat: mixture and its reconstruction (top), estimated A2 (middle) and P2 (bottom), resolving an earlier, louder A2 and a later, weaker P2.

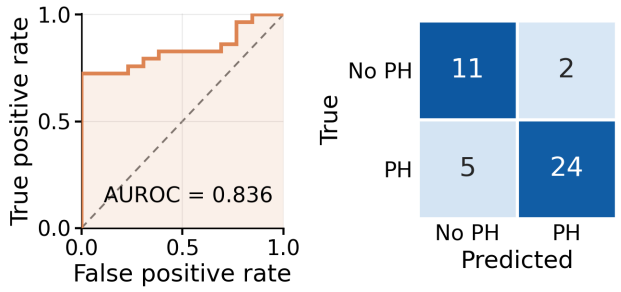
The clinical capability of the developed system is reported in Table 1, showing how well the separated waveforms support the diagnosis. On average, the learned separator feeding the classification pipeline performs below the alternating-optimization baseline: its best mean AUROC, about 0.69 for the random forest, trails the baseline's 0.825 and 0.764, with the remaining metrics following the same order. The gap narrows at the top of the distribution, where the strongest seed matches or exceeds the baseline, reaching an AUROC of 0.836 and a balanced accuracy of 0.837 — 35 of 42 patients are correctly classified, with the errors spread across both classes rather than concentrated in the minority controls. Figure 3 shows the corresponding operating characteristic, which rises steeply and recovers most of the PH-positive patients at a low false-positive rate, to-

gether with the confusion matrix of the decisions taken at the tuned threshold.

Table 1: PH detection on the 42-patient cohort, leave-one-patient-out. *Proposed*: learned separation, mean ± standard deviation over ten seeds, best classifier in bold. *Best seed*: seed 256, SVM. *Baseline*: same pipeline on the deterministic separation of [5].

	Proposed		**Best**	**Baseline** [5]	
Metric	**SVM-RBF**	**RF**	**seed**	**SVM**	**RF**
AUROC	0.675 ± 0.204	$\mathbf{0.686 \pm 0.088}$	0.836	0.825	0.764
AP	0.817 ± 0.121	$\mathbf{0.830 \pm 0.071}$	0.938	0.843	0.815
BalAcc	0.635 ± 0.118	$\mathbf{0.646 \pm 0.059}$	0.837	0.700	0.798
F1-macro	0.624 ± 0.111	$\mathbf{0.633 \pm 0.063}$	0.816	0.708	0.786

The random forest is the more stable of the two classifiers, attaining both the highest mean and the lowest standard deviation on every metric. The main weakness of the learned separator is the variability between seeds: a single good run should not be taken as the rule. However, favourable runs exist and, when reached, the separation is competitive with the iterative method at one forward pass per beat rather than a per-signal optimization.



(a) ROC curve.

(b) Confusion matrix.

Figure 3: Best-balanced run (SVM, seed 256): AUROC 0.836, balanced accuracy 0.837, F1-macro 0.816.

4. Discussion and Conclusions

This work is a first adaptation of the multi-encoder masking architecture [6] to the second heart sound, guided by an objective grounded on the physiology of the two valve closures: an additive reconstruction, an anti-collapse penalty keeping both closures present, and a scale-invariant cosine term that discourages redundancy. On clinical recordings the two branches settle, without supervision, on an earlier and louder transient and a later, weaker and more oscillatory one, consistent with the aortic and the pulmonary closure.

Fed to the diagnostic pipeline, this learned separator is competitive with the alternating-optimization baseline, although on average below it, with only its strongest seeds matching the iterative method. Several limitations are associated with these findings: the cohort holds only 42 patients and is imbalanced; the separation quality on the real recordings can only be judged indirectly, through visual inspection and the diagnosis task; and the model is stochastic where the baseline is deterministic, so a favourable seed may report more than the training recipe can reliably deliver.

Acknowledgements

This work is co-funded 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) (https://doi.org/10.54499/17172 (COMPETE2030-FEDER-00864900)).

References

[1] Gaudio A, Giordano N, Elhilali M, Schmidt S, Renna F. Pulmonary hypertension detection from heart sound analysis. IEEE Trans Biomed Eng 2025;72(10):2902–2914.

[2] Sæderup R, Hoang P, Winther S, Bøttcher M, Struijk J, Schmidt S, Østergaard J. Estimation of the second heart sound split using windowed sinusoidal models. Biomed Signal Process Control 2018;44:229–236.

[3] Barma S, Chen BW, Man K, Wang JF. Quantitative measurement of split of the second heart sound (S2). IEEEACM Trans Comput Biol Bioinform 2015;12(4):851–860.

[4] Muramatsu S, Yamamoto M, Takamatsu S, Itoh T. Estimation of splitting interval in second heart sound by optimizing a demixing vector. In 2023 IEEE SENSORS. Vienna, Austria, 2023; 1–4.

[5] Renna F, Gaudio A, Mattos S, Plumbley M, Coimbra M. Separation of the aortic and pulmonary components of the second heart sound via alternating optimization. IEEE Access 2024;12:34632–34643.

[6] Webster M, Lee J. Blind source separation of single-channel mixtures via multi-encoder autoencoders. Neurocomputing 2025;652:131008.

[7] Renna F, Oliveira J, Coimbra M. Deep convolutional neural networks for heart sound segmentation. IEEE J Biomed Health Inform 2019;23(6):2435–2445.

Address for correspondence:

Francisco da Ana

INESC TEC, Campus da FEUP, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal

francisco.p.ana@inesctec.pt