PHAROS+: NON-INVASIVE LONGITUDINAL MONITORING OF PULMONARY HYPERTENSION WITH UNIMODAL AND MULTIMODAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Pulmonary hypertension is a disease characterized by elevated pressures in the blood vessels that supply the lungs. It is a progressive and incurable disease that can lead to right heart failure and premature death if improperly managed. Close monitoring plays an important role in management of patients with pulmonary hypertension, as it facilitates the timely detection of disease progression and enables the prompt administration of therapies that can alter the course of the disease. The gold standard for monitoring disease progression is a right heart catheterization (RHC) – a procedure that involves the insertion of a catheter, attached to a pressure transducer, into the pulmonary vasculature to measure the pulmonary pressures. This procedure is typically repeated several times during the course of a patient's life to monitor the response to therapies designed to reduce pulmonary pressures. Although RHC is an important tool that can help guide the care of patients with pulmonary hypertension, the procedure itself entails some risk to the patient and can only be performed in hospitals that have the needed equipment and trained personnel. Prior attempts to develop non-invasive alternatives for measuring pulmonary pressures have primarily focused on the task of initial diagnosis, rather than long-term monitoring of patients that have already been diagnosed. In this work, we propose a novel deep learning paradigm for the non-invasive assessment of pulmonary artery pressures. The method leverages electrocardiographic signals and, when available, cardiac ultrasound data to enable long-term monitoring in these patients. We demonstrate that our method achieves strong performance on an internal dataset from one hospital and generalizes well to the MIMIC-III Waveform Database from a different hospital. Our approach provides a cheap and accessible method that can be used to monitor patients with pulmonary hypertension at home. To the best of our knowledge, this work is the first to address the task of longitudinal monitoring in patients with pulmonary hypertension.

1 INTRODUCTION

Pulmonary hypertension is a chronic and progressive disease estimated to affect around 1% of the global population (Hoeper et al., 2016). It is characterized by high pressures in the pulmonary vasculature (blood vessels of the lungs) and is formally defined by a mean pulmonary artery pressure (mPAP) greater than 20 mmHg (Humbert et al., 2022). If the disease is not well-controlled by medications, it can lead to right heart failure and premature death. As a result, repeated hemodynamic evaluations (measuring pulmonary artery pressures) are crucial for the diagnosis and long-term monitoring and management. Currently, the gold standard for measuring these pressures is right heart catheterization (RHC), an invasive procedure that involves threading a catheter through a major vein in the body to the heart. This procedure entails risk, must be performed in a hospital setting, and is limited to hospitals that have a catheterization suite and trained personnel.

In recent years, deep learning approaches have shown promise in assessing cardiac hemodynamics from non-invasive data modalities. For instance, Tripathi et al. (2024) used cardiac magnetic resonance imaging and information from electronic health records (EHR) and Suvon et al. (2024) used electrocardiograms (ECG) and chest x-rays to detect elevated mean pulmonary capillary wedge pressures (mPCWP) for heart failure patients. For assessing pulmonary hypertension, Zhao et al.

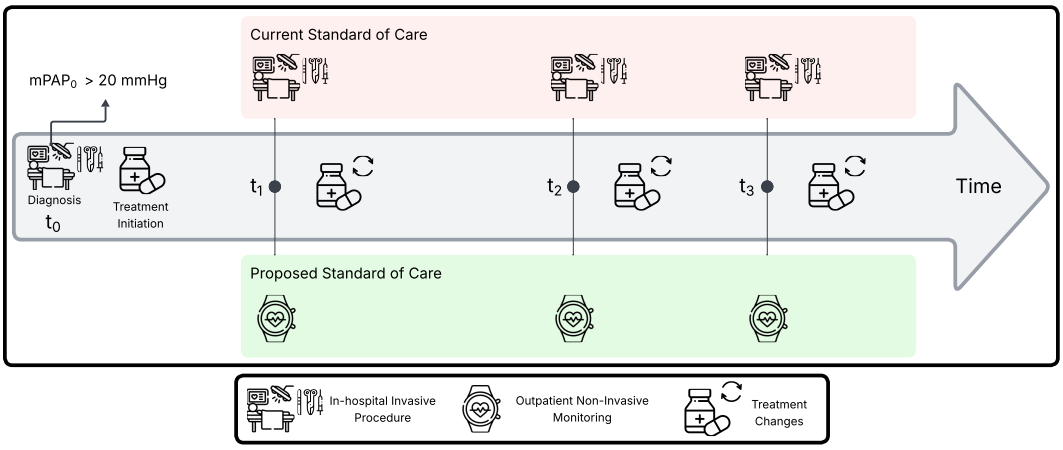

Figure 1: Clinical impact of our proposed work.

(2025) used a combination of tabular and textual data from EHRs along with chest x-rays while Liu et al. (2024b) used ECGs and chest x-rays to diagnose pulmonary hypertension. These methods, however, focus on identifying patients who have pulmonary hypertension (the diagnosis task) rather than longitudinally following patients who are known to have elevated pulmonary pressures.

In patients with pulmonary hypertension, the mPAP rarely normalizes, even with appropriate therapy. The goal in these patients is therefore to reduce the pulmonary pressure as much as possible in an attempt to prevent disease progression. Routine care therefore typically involves repeated RHCs to assess changes in the mPAP as therapy is adjusted (Figure 1). For effective longitudinal monitoring in patients with known pulmonary hypertension, rather than detecting elevated pressures, we propose a method to detect whether the mPAP has increased from its baseline value at any given point in time. As an increase in mPAP reflects disease progression, methods that identify such increases would directly influence therapy without the need for repeated invasive procedures. We envision the impact of our work to be as illustrated in Figure 1.

Our approach leverages ECGs to enable cheap, accessible, and outpatient or at-home monitoring, without the need for in-person hospital visits. In particular, we leverage single lead (Lead I) ECGs for its convenience and widespread availability in wearable and handheld electronic devices. We develop a flexible approach that can use ECGs as a single modality, when only ECGs are available, or use a combination of ECGs and cardiac ultrasound data, when both modalities are available.

The main contributions of this work are:

- We propose an ECG-based method, which we refer to as PHAROS (**P**ulmonary **H**ypertension **A**ssessment f**RO**m ECG **S**ignals), and a multi-modal ECG and echocardiography-based method, PHAROS+, for detecting pulmonary hypertension progression. To the best of our knowledge, our work is the first to address the task of longitudinal monitoring in patients with pulmonary hypertension.
- We evaluate the model on two independent datasets and demonstrate that strong performance is maintained despite a dataset shift.

2 Related Work

Data modalities for ML-based non-invasive pulmonary hypertension assessment. Prior studies suggest that echocardiograms (echo) can be used to estimate pulmonary artery pressure (Yock & Popp, 1984), and hence it can be used to follow patients with pulmonary hypertension (Humbert et al., 2022). However, measurements from cardiac ultrasound require a trained sonographer, and can vary depending on their skill level as well as the quality of the echocardiographic images obtained. As a result, estimates of pulmonary artery pressure from echos have large errors (Fisher et al., 2009; D'Alto et al., 2013; Slegg et al., 2021; Janda et al., 2011), making them unreliable

for long-term monitoring and therapeutic decision-making. Some machine learning methods have been introduced as attempts to improve the accuracy of echocardiography-guided estimates of pulmonary artery pressures (Anand et al., 2024; Salehi et al., 2025; Leha et al., 2019), while others have leveraged other data modalities, including cardiac magnetic resonance imaging (Cheng et al., 2025; Swift et al., 2020), computed tomography pulmonary angiography (Xie et al., 2025; Sharkey et al., 2022), and chest x-rays (Huang et al., 2025). Combinations of multiple modalities have also been explored, such as tabular, textual, and x-ray (Zhao et al., 2025) or ECG and x-ray (Liu et al., 2024b). Disadvantages of these methods are that they primarily address the diagnosis task and they require patients to make an in-person visit to a hospital or imaging center to obtain the scans. Other methods have used only readily available data such as electronic health records (Kogan et al., 2023), phonocardiograms (Guo et al., 2025), and ECGs (DuBrock et al., 2024; Raghu et al., 2023a; Aras et al., 2023; McLean et al., 2025; Suvon et al., 2025) to diagnose pulmonary hypertension. The goal of our work is also to use readily available data, primarily ECGs, to enable outpatient or at-home monitoring. Our method is flexible in the sense that it can be used when only an ECG available and when both ECG and echocardiographic measurements are available. Moreover, instead of addressing the diagnosis task as done by prior work, our proposed method aims to identify disease progression.

Non-invasive hemodynamics assessment from ECG. ECGs have been widely used in non-invasive, deep learning-based health assessment, including detecting demographics and health conditions (Abbaspourazad et al., 2024), sleep stages and disorders (Thapa et al., 2024), cardiac arrhythmias (Hannun et al., 2019; Liu et al., 2024a), heart attacks (Acharya et al., 2017), heart failure (Acharya et al., 2019), and hemodynamic abnormalities in patients with heart failure (Schlesinger et al., 2022; 2025; Raghu et al., 2023b) or pulmonary hypertension (Schlesinger et al., 2022; DuBrock et al., 2024; Aras et al., 2023; McLean et al., 2025; Suvon et al., 2025; Sadrawi et al., 2021). The hemodynamics assessment methods all aim to distinguish elevated from non-elevated pressures according to a predefined pressure threshold. The performance of these methods for estimating mPAP vary widely and, more importantly, these methods may not be useful for longitudinal monitoring as patients with pulmonary hypertension rarely have pressures that fall within the normal range. Instead of diagnosis, a few other methods have introduced hemodynamics estimation as regression tasks using physiologic signals as input (Klein et al., 2025; Jeong et al., 2023; Sadrawi et al., 2021). Klein et al. (2025) and Jeong et al. (2023) present methods that regress mPCWP and Sadrawi et al. (2021) regress mPAP. However, regression errors exceed variability in the RHC (Melillo et al., 2020) and clinically significant changes in pressure, making these methods unreliable for long-term monitoring. In contrast to prior work, we use ECGs to classify disease progression in pulmonary hypertension for the purpose of long-term monitoring and management.

3 METHOD

A RHC that definitively measures an elevated mPAP is required to make the diagnosis of pulmonary hypertension. Consequently, virtually all patients who have this diagnosis have had at least one RHC (Humbert et al., 2022). Using a dataset consisting of paired ECG and mPAP measurements from a RHC, our task is therefore to develop a method that uses ECGs to non-invasively determine whether there has been an increase in mPAP over time, thereby circumventing the need for repeated invasive procedures (Figure 1).

3.1 PROBLEM FORMULATION

We formulate the pulmonary hypertension monitoring problem as an optimization problem where the goal is to estimate the probability distribution of disease progression at time t_j relative to previous time t_i. More precisely, we estimate $p(Y_{ij}|\mathbf{s}_i, \mathbf{s}_j, m_i, \Delta t, \mathbf{f}_j) \in [0, 1]$, where $Y_{ij} \in \{0, 1\}$ is a binary label denoting whether the disease has progressed (i.e., the mPAP has increased) between times t_i and t_j. We define disease progression as an increase in mPAP that is larger than a certain threshold. That is, the ground truth label $Y_{ij}^{gt} = 1$ when $\Delta mPAP = m_j - m_i > threshold$

The time of the initial RHC and ECG is t_i; $\Delta t = t_j - t_i > 0$, is the time interval between the initial ECG at time t_i, and the current ECG at time t_j; $\mathbf{s}_i$ is the initial ECG obtained at time t_i; $\mathbf{s}_j$ is the current ECG obtained at time t_j; m_i is the known mPAP acquired via RHC at the time t_i; m_j is the

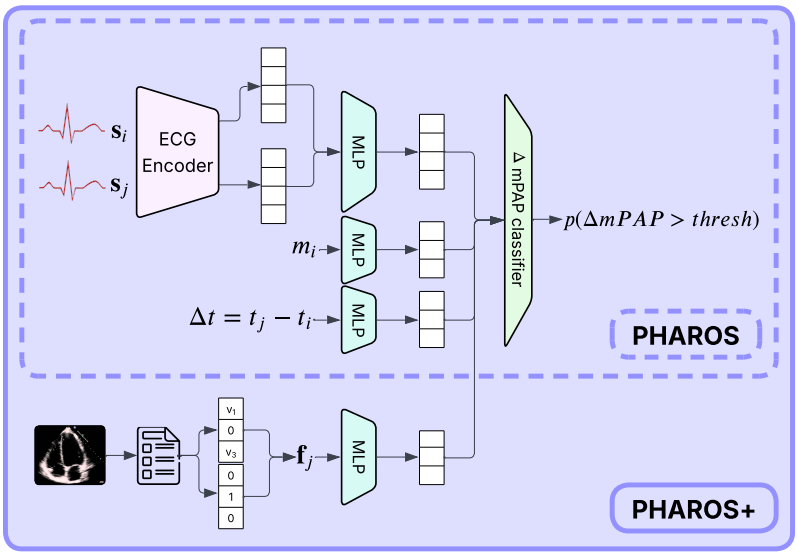

Figure 2: Overview of PHAROS (dashed box) and PHAROS+ (solid box). Merging arrows denote concatenation. Our models take two ECGs, $\mathbf{s}_i$ and $\mathbf{s}_j$, corresponding to time points t_i and t_j, respectively, the baseline mPAP m_i corresponding to the first ECG, and the time gap $\Delta t = t_j - t_i > 0$. In PHAROS+, we also input a vector of echo features $\mathbf{f}_j$ corresponding to the most recent echo relative to t_j. The embeddings of all the inputs are concatenated and passed to a final MLP to output the final prediction of disease progression.

mPAP at time t_j; $\mathbf{f}_j$ represents additional features of the patient corresponding to time t_j. In this work, $\mathbf{f}_j$ represents measurements from the echo most recent to time t_j if available for the patient.

3.2 Model Architecture

Our model architecture consists of an encoder that individually encodes $\mathbf{s}_i$ and $\mathbf{s}_j$, corresponding to two different time points, an encoder for the baseline mPAP m_i, and an encoder for the time gap between the ECGs Δt. We refer to this model as PHAROS. We additionally explore including echo measurements for patients who have echos, train a separate model, and compare performance. The model trained with echo parameters is referred to as PHAROS+. An overview of our architecture is shown in Figure 2.

Encoding ECGs. As input to our model, we use Lead I ECGs due to its convenience and availability on at-home and wearable health monitoring devices. Given two Lead I ECGs $\mathbf{s}_i$ and $\mathbf{s}_j$ from two different time points t_i and t_j, respectively, we encode each ECG using a ResNet-inspired (He et al., 2016) 1D convolutional neural network (CNN). We use the same encoder weights for both ECGs. The embeddings of $\mathbf{s}_i$ and $\mathbf{s}_j$ from the ECG encoder are concatenated and fed to a multi-layer perceptron (MLP) to fuse the embeddings.

Encoding the baseline mPAP m_i and time gap Δt. The baseline mPAP m_i and time gap Δt are each input to their own MLPs to obtain embeddings of each value. Prior to training, the mPAP values are normalized to have zero mean and unit variance in the training dataset. The values of Δt are scaled to be in units of months.

Encoding echo measurements as optional features. We train a version of our model, PHAROS+, that optionally takes echo measurements as input when they are available for a patient. Each feature type in the vector is normalized across the dataset to have zero mean and unit variance. Not all patients have echos, and not all measurements are taken for the patients that do have echos. As a result, there can be missing values. To account for this, we vectorize all possible features and fill in missing values with zeros. We concatenate the features with a mask that identifies which elements are missing. The result forms the echo input $\mathbf{f}_j$, which is passed through a MLP to obtain an echo embedding.

Classifying disease progression. Finally, for determining disease progression, the fused ECG embedding, the embeddings of m_i and Δt, and the echo embedding (for PHAROS+) are concatenated and fed to a MLP that outputs the estimated probability of disease progression at t_j. For the training loss, we use the binary cross-entropy loss.

4 EXPERIMENTS

4.1 DATASETS

To train our model, a large source of time-aligned Lead I ECGs and mPAP measurements is required. Although obtaining ECGs is typically routine procedure in preparation for RHCs, the ECG and RHC are not done at the exact same time and could even be weeks apart. This data would have noisy ground truth labels due to inherent variability in mPAP even within the same day (Rich et al., 1985) or hour (Melillo et al., 2020). Therefore, large quantites of such well-aligned data is difficult to obtain from the RHC procedures done in a catheterization lab of a hospital. Alternatively, patients are sometimes admitted to intensive care units (ICU) with a pulmonary artery catheter left inside the pulmonary artery to continuously monitor the mPAP in real-time. Essentially all ICU patients also have continuous ECG monitoring. We therefore take advantage of this ICU data to train our model. In our experiments, we verify that our model performance is maintained at larger time gaps by evaluating on the subsets of data corresponding to longer ICU stays.

Massachusetts General Hospital (MGH) Dataset. We train, validate, and test our model on a large private dataset from MGH. This dataset consists of waveform data recorded from ICU bedside monitors, including Lead I ECG and pulmonary artery pressure waveforms. There are also echo measurements for patients that had echos. More details about the echos are described in Appendix A. We randomly split the dataset by patient into 80%, 10%, and 10% splits for training, validation, and testing, respectively.

MIMIC-III Waveform Database (Moody et al., 2020; Johnson et al., 2016). We externally evaluate our model on the publicly available MIMIC-III Waveform Database from PhysioNet (Goldberger et al., 2000). This dataset was collected from ICUs at Beth Israel Deaconess Medical Center. We use the subset of recordings that contain aligned Lead I ECG and pulmonary artery pressure waveforms. There are no echo measurements in this dataset, so when evaluating PHAROS+, we input fully masked vectors.

4.2 DATA PRE-PROCESSING

We prepare the data by splitting up aligned continuous Lead I ECG and pulmonary artery pressure waveforms into 10-second segments, the standard duration of outpatient ECGs (Sattar & Chhabra, 2023). The aligned waveforms are then filtered to ensure they have good signal quality and alignment, and poor quality waveforms are discarded. Further details are discussed in Appendix B. The ECGs in the MGH dataset have a frequency of 120 Hz or 240 Hz and those in the MIMIC-III dataset have a frequency of 125 Hz. Prior to inputting ECGs to our model, they are upsampled such that $\mathbf{s}_i, \mathbf{s}_j \in \mathbb{R}^{4096}$. The statistics of both datasets after pre-processing are shown in Table 1.

For validation and testing on the MGH dataset and for external evaluation on MIMIC-III, the pairs of ECGs, $\mathbf{s}_i$ and $\mathbf{s}_j$, are randomly sampled offline and saved to ensure we evaluate on the exact same pairs each time. During training, we randomly sample different pairs in an online fashion, allowing pairs to vary at each epoch to make full use of the large training data. We validate every approximately 5.4 million pairs sampled across all ICU stays and refer to this as one epoch. For all data splits, ECGs in a sampled pair always come from the same ICU stay and the number of sampled pairs per ICU stay is proportional to its duration. We compute the mPAPs m_i and m_j from the aligned pulmonary artery pressure waveform corresponding to $\mathbf{s}_i$ and $\mathbf{s}_j$, respectively.

4.3 IMPLEMENTATION, TRAINING, AND EVALUATION DETAILS

Implementation. The ECG encoder is a 1D CNN that begins with a convolutional layer of size (1, 64) followed by a residual block of size (64, 256) and a global pooling layer, leading to embeddings of $\mathbf{s}_i$ and $\mathbf{s}_j$ that have a size of 256 each. The two embeddings are concatenated and a two-layer

	MGH Dataset	MIMIC-III
# patients	7894	N/A
# ICU stays	8213	295
# 10-sec aligned Lead I ECG and mPAP	63.3 million	1.1 million
Average aligned waveform duration per ICU stay	1.7 days	0.7 days
# patients with echos	3885	N/A
# ICU stays with echos	4121	N/A
# echos	7993	N/A
Age (years, mean $\pm$ std)	64 ± 14	N/A
Sex (female, %)	30.3%	N/A

Table 1: Statistics of the parts of the MGH and MIMIC-III Waveform (Moody et al., 2020) datasets that contain aligned Lead I ECG and pulmonary artery pressure waveforms. Numbers shown correspond to the data after pre-processing.

MLP with layer sizes of (512, 256) is used to combine them into an embedding of size 256. The MLPs for m_i and Δt have the same architecture. Both are two-layer MLPs with layer sizes of (32, 32) that map their respective scalar inputs to embeddings of size 32. The echo inputs consist of 57 measurements and a mask of the same size to indicate missing values. They are concatenated to give $\mathbf{f}_j \in \mathbb{R}^{114}$. The MLP that embeds $\mathbf{f}_j$ consists of two layers with sizes of (64, 32), yielding an embedding of size 32. Finally, all the embeddings are concatenated and fed to a three-layer MLP with layer sizes (256, 256, 1) to obtain the final classification. More details about the architecture are presented in Appendix C.

Measurements of mPAP are generally presumed to have inherent variability, so small increases in mPAP should not be considered disease progression. In fact, Melillo et al. (2020) have shown that the standard deviation of differences between two mPAP measurements in a single RHC procedure is 3.9 mmHg. Considering this, we use a $\Delta mPAP$ threshold of 4 mmHg to define the ground truth positive class. That is, for every ECG pair corresponding to $\Delta mPAP > 4$ the ground truth label is $Y_{ij}^{gt} = 1$. This threshold results in a positive class prevalence of 0.1990 in the MGH dataset and 0.1841 in MIMIC-III.

Training. We train our models using binary cross-entropy loss and the AdamW optimizer (Loshchilov & Hutter, 2019) with a batch size of 256, learning rate of 1×10^{-6}, and weight decay of 0.01. Each model is trained for 20 epochs or until the validation loss converged as defined by early stopping after at least 5 epochs. Models were trained on a single NVIDIA RTX A6000 GPU, with training time taking around 32 hours. We use the validation AUC to select the best model and report results on the test split of the MGH dataset and externally evaluate on MIMIC-III.

Evaluation. We evaluate our models on several common binary classification metrics, including Area Under the Receiver Operating Characteristic Curve (AUC), Brier score, Area Under the Precision-Recall Curve (AUPRC), positive predictive value (PPV, also known as precision), and negative predictive value (NPV). The probability threshold used to binarize model outputs for the PPV and NPV was the threshold corresponding to a sensitivity (also known as recall or true positive rate) of 80%, as commonly done in related work (Schlesinger et al., 2022; Raghu et al., 2023b). To calculate statistical measures of uncertainty for all metrics, sets of inputs were randomly drawn with replacement from the test set to produce 10 bootstrapped datasets with equal size to the original test dataset. The reported errors represent one standard deviation computed using the results across the 10 bootstraps.

Baseline Comparison. We compare PHAROS and PHAROS+ to the current non-invasive standard for longitudinal monitoring – pressure estimation using echo data (Humbert et al., 2022). In an echo, the peak velocity of tricuspid regurgitation (the abnormal backflow of blood from the right ventricle to the right atrium) can be often be measured in patients with pulmonary hypertension. This velocity is used to compute the pressure gradient between the right atrium and ventricle using $g_i = 4v_i^2$, where g_i is the pressure gradient and v_i is the peak tricuspid regurgitation velocity from the echo at time t_i (Yock & Popp, 1984). From this, the pulmonary artery systolic pressure (PASP), which is approximately equal to the right ventricular systolic pressure (RVSP), can be computed as

$PASP_i \approx RVSP_i = g_i + RAP$, where RAP is the pressure in the right atrium. RAP cannot be measured on an echo, so a constant value is commonly assumed and simply added to g to estimate $RVSP$. Likewise, we assume that RAP is some constant and does not change in time. Therefore, given two velocities v_i and v_j from t_i and t_j, respectively, we can compute the change in PASP as $\Delta PASP = PASP_j - PASP_i = (g_j + RAP_j) - (g_i + RAP_i) = g_j - g_i$. To convert $\Delta PASP$ to $\Delta mPAP$ we use the widely accepted equation proposed by Chemla et al. (2004) to get $\Delta mPAP = 0.61 \times \Delta PASP$. To compute the binary classification metrics described above, we use `sigmoid(`$\Delta mPAP$ `- thresh)` to convert our computations to probabilities, where `thresh=4` is the $\Delta mPAP$ threshold used to define the positive class.

5 Results and Discussion

5.1 Evaluation on the MGH Dataset

Our results on the MGH dataset are shown in Table 2. PHAROS achieves a strong performance, significantly outperforming the echo baseline on all metrics. In PHAROS+, by including echo data when available, the AUC, Brier score, AUPRC, and PPV improve. This shows that echo measurements can help the model achieve better performance when it is available. The NPV of PHAROS+ decreases slightly compared to PHAROS, but remains significantly higher than the echo baseline. Overall, our method significantly outperforms the current standard for non-invasive longitudinal monitoring. We also note that the performance of PHAROS+ when only using ECG data (i.e., masking out echo data for all inputs) is similar to the performance of PHAROS which was trained without echo.

Model	Eval Inputs	AUC (↑)	Brier (↓)	AUPRC (↑)	PPV (↑)	NPV (↑)
Echo Baseline	–	0.5147 ± 0.0789	0.2523 ± 0.0256	0.1932 ± 0.0676	0.1765 ± 0.0390	0.8750 ± 0.0720
PHAROS	ECG	0.7803 ± 0.0010	0.1312 ± 0.0003	0.4932 ± 0.0023	0.3289 ± 0.0010	**0.9236** ± 0.0006
PHAROS+	ECG only	0.7809 ± 0.0009	0.1307 ± 0.0002	0.4895 ± 0.0021	0.3267 ± 0.0010	0.9235 ± 0.0005
PHAROS+	ECG + Echo	**0.7827** ± 0.0009	**0.1296** ± 0.0002	**0.4979** ± 0.0019	**0.3305** ± 0.0009	0.9211 ± 0.0004

Table 2: Performance of PHAROS and PHAROS+ on the MGH dataset. (↑ indicates higher is better and ↓ indicates lower is better.)

5.2 Evaluation on the MGH Dataset for Larger Time Gaps

We examine the effect of large time gaps on performance. Plots for PHAROS and PHAROS+ are shown in Figure 3. In general, the models maintain strong performance and slightly improves, albeit with larger standard error, as the time gap increases up to a duration of 45 days. This is a promising finding, as it indicates the potential for PHAROS and PHAROS+ to perform well on larger time gaps beyond the span of typical ICU lengths of stay.

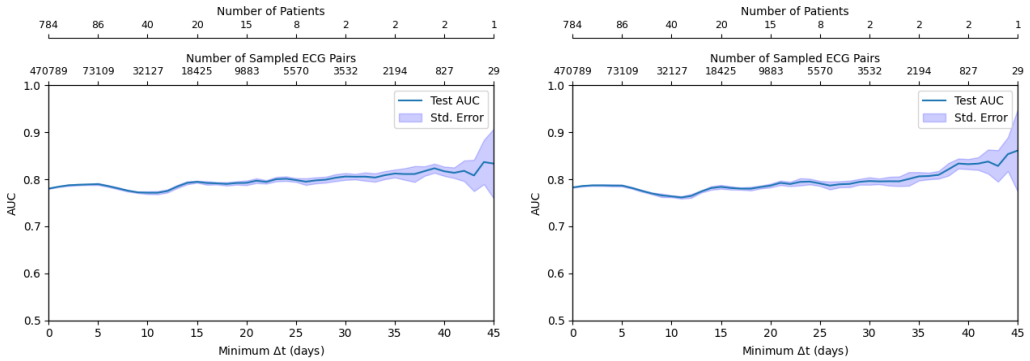

Figure 3: AUC for $0 < min(\Delta t) <= 45$ on the MGH dataset for PHAROS (left) and PHAROS+ (right).

5.3 External Evaluation on the MIMIC-III Waveform Database

In our previous analysis (Table 2), we found that when given ECGs only, PHAROS+ has a similar performance to PHAROS on our internal dataset, and when echo data is available, PHAROS+ performs better. PHAROS+ would therefore be the model to use in practice, so we focus on this model for external evaluation. We use the MIMIC-III Waveform Database (Moody et al., 2020), which was collected from a different hospital, to verify that performance translates to other patient populations. The waveforms in MIMIC-III do not have associated echo measurements, so we can only evaluate PHAROS+ using ECGs and need to fully mask the input vector $\mathbf{f}_j$. Our results are shown in Table 3. We find that despite a performance gap, our model still performs well with an AUC above 0.7 and a slight increase in NPV. These results demonstrate good generalization to unseen data distributions.

Dataset	Eval Inputs	AUC (↑)	Brier (↓)	AUPRC (↑)	PPV (↑)	NPV (↑)
MGH	ECG only	0.7809 $\pm$ 0.0009	0.1307 $\pm$ 0.0002	0.4895 $\pm$ 0.0021	0.3267 $\pm$ 0.0010	0.9235 $\pm$ 0.0005
MIMIC-III	ECG only	0.7052 $\pm$ 0.0007	0.1649 $\pm$ 0.0002	0.3145 $\pm$ 0.0011	0.2375 $\pm$ 0.0006	0.9550 $\pm$ 0.0004

Table 3: Performance of PHAROS+ on MIMIC-III compared to MGH. (↑ indicates higher is better and ↓ indicates lower is better.)

6 Conclusion

In this work, we proposed a multi-modal approach that uses ECGs and echos for addressing the novel task of non-invasively identifying pulmonary hypertension progression, defined by an increase in mPAP between two timepoints of $\Delta mPAP > 4$ mmHg. The performance of PHAROS, trained using only ECGs, was improved upon by PHAROS+, which used ECGs and was also able to incorporate echo data when available to boost performance. Our models demonstrated strong performance across various metrics, even for time gaps of up to 45 days. The discriminatory ability of PHAROS and PHAROS+ is significantly above what is obtained used echocardiographic measurements alone. Indeed, the AUC for the estimating pulmonary artery pressures using echocardiography is close to random (0.5), suggesting that this is not reliable method for estimating changes in pulmonary artery pressures. By contrast, our results argue that more reliable estimates of pulmonary artery pressure changes can be obtained with single-lead ECG signals, which do not require a highly trained operator for data acquisition. The inadequacy of echocardiographic information for assessing changes in pulmonary artery pressures is further highlighted by the fact that the PHAROS+, which uses ECG and echo data, has only modest improvement over PHAROS, which only uses ECG data.

External validation of PHAROS+ on the MIMIC-III Waveform Database demonstrated good discriminatory ability, albeit worse than what was observed in the MGH dataset. Nevertheless, the NPV (at a threshold corresponding to 80% recall) remains strong ($> 95\%$) in this cohort. The high NPVs in both the MGH and MIMIC datasets argue that a negative result from PHAROS/PHAROS+ does not suggest that corresponding pulmonary artery pressure is elevated, at a threshold that ensures a true positive rate of 80%.

References

Salar Abbaspourazad, Oussama Elachqar, Andrew Miller, Saba Emrani, Udhyakumar Nallasamy, and Ian Shapiro. Large-scale training of foundation models for wearable biosignals. In *ICLR*, 2024.

U Rajendra Acharya, Hamido Fujita, Shu Lih Oh, Yuki Hagiwara, Jen Hong Tan, and Muhammad Adam. Application of deep convolutional neural network for automated detection of myocardial infarction using ecg signals. *Information Sciences*, pp. 190–198, 2017.

U Rajendra Acharya, Hamido Fujita, Shu Lih Oh, Yuki Hagiwara, Jen Hong Tan, Muhammad Adam, and Ru San Tan. Deep convolutional neural network for the automated diagnosis of congestive heart failure using ecg signals. *Applied Intelligence*, 49:16–27, 2019.

Vidhu Anand, Alexander D. Weston, Christopher G. Scott, Garvan C. Kane, Patricia A. Pellikka, and Rickey E. Carter. Machine learning for diagnosis of pulmonary hypertension by echocardiography. *Mayo Clinic Proceedings*, 99(2):260–270, 2024.

Mandar A. Aras, Sean Abreau, Hunter Mills, Lakshmi Radhakrishnan, Liviu Klein, Neha Mantri, Benjamin Rubin, Joshua Barrios, Christel Chehoud, Emily Kogan, Xavier Gitton, Anderson Nnewihe, Deborah Quinn, Charles Bridges, Atul J. Butte, Jeffrey E. Olgin, and Geoffrey H. Tison. Electrocardiogram detection of pulmonary hypertension using deep learning. *Journal of Cardiac Failure*, 29(7):1017–1028, 2023.

Denis Chemla, Vincent Castelain, Marc Humbert, Jean-Louis Hébert, Gérald Simonneau, Yves Lecarpentier, and Philippe Hervé. New formula for predicting mean pulmonary artery pressure using systolic pulmonary artery pressure. *CHEST*, 126(4):1313–1317, 2004.

Li-Hsin Cheng, Xiaowu Sun, Charlie Elliot, Robin Condliffe, David G. Kiely, Samer Alabed, Andrew J. Swift, Rob J. van der Geest, David G Kiely, Lisa Watson, Iain Armstrong, Catherine Billings, Athanasios Charalampopoulos, Robin Condliffe, Charlie Elliot, Abdul Hameed, Neil Hamilton, Judith Hurdman, Allan Lawrie, Robert A Lewis, Smitha Rajaram, Alex Rothman, Andy J. Swift, Steven Wood, AA Roger Thompson, and Jim Wild. Mean pulmonary artery pressure prediction with explainable multi-view cardiovascular magnetic resonance cine series deep learning model. *Journal of Cardiovascular Magnetic Resonance*, 27(1):101133, 2025.

Michele D'Alto, Emanuele Romeo, Paola Argiento, Antonello D'Andrea, Rebecca Vanderpool, Anna Correra, Eduardo Bossone, Berardo Sarubbi, Raffaele Calabrò, Maria Giovanna Russo, and Robert Naeije. Accuracy and precision of echocardiography versus right heart catheterization for the assessment of pulmonary hypertension. *International Journal of Cardiology*, 168(4), 2013.

Hilary M. DuBrock, Tyler E. Wagner, Katherine Carlson, Corinne L. Carpenter, Samir Awasthi, Zachi I. Attia, Robert P. Frantz, Paul A. Friedman, Suraj Kapa, Jeffrey Annis, Evan L. Brittain, Anna R. Hemnes, Samuel J. Asirvatham, Melwin Babu, Ashim Prasad, Unice Yoo, Rakesh Barve, Mona Selej, Peter Agron, Emily Kogan, Deborah Quinn, Preston Dunnmon, Najat Khan, and Venky Soundararajan. An electrocardiogram-based ai algorithm for early detection of pulmonary hypertension. *European Respiratory Journal*, 64(1), 2024.

Micah R. Fisher, Paul R. Forfia, Elzbieta Chamera, Traci Housten-Harris, Hunter C. Champion, Reda E. Girgis, Mary C. Corretti, and Paul M. Hassoun. Accuracy of doppler echocardiography in the hemodynamic assessment of pulmonary hypertension. *American Journal of Respiratory and Critical Care Medicine*, 179(7), 2009.

A. Goldberger, L. Amaral, L. Glass, J. Hausdorff, P. C. Ivanov, R. Mark, J.E. Mietus, G.B. Moody, C.K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation [Online]*, 101(23):e215–e220, 2000. Accessed: Sept 16, 2025.

Ling Guo, Nivedita Khobragade, Spencer Kieu, Suleman Ilyas, Preston N. Nicely, Emmanuel K. Asiedu, Fabio V. Lima, Caroline Currie, Emileigh Lastowski, and Gaurav Choudhary. Development and evaluation of a deep learning–based pulmonary hypertension screening algorithm using a digital stethoscope. *Journal of the American Heart Association*, 14(3), 2025.

Awni Y Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H Tison, Codie Bourn, Mintu P Turakhia, and Andrew Y Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25(1): 65–69, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, 2016.

Marius M Hoeper, Marc Humbert, Rogerio Souza, Majdy Idrees, Steven M Kawut, Karen Sliwa-Hahnle, Zhi-Cheng Jing, and J Simon R Gibbs. A global view of pulmonary hypertension. *The Lancet Respiratory Medicine*, 4(4):306–322, 2016.

Zhihua Huang, Xiaolin Diao, Yanni Huo, Zhihui Zhao, Jiahui Geng, Qing Zhao, Jia Liu, Qunying Xi, Yun Xia, Ou Xu, Xin Li, Anqi Duan, Sicheng Zhang, Luyang Gao, Yijia Wang, Sicong Li, Qin Luo, Zhihong Liu, and Wei Zhao. Deep learning-enhanced noninvasive detection of pulmonary hypertension and subtypes via chest radiographs, validated by catheterization. *CHEST*, 2025.

Marc Humbert, Gabor Kovacs, Marius M Hoeper, Roberto Badagliacca, Rolf M F Berger, Margarita Brida, Jørn Carlsen, Andrew J S Coats, Pilar Escribano-Subias, Pisana Ferrari, Diogenes S Ferreira, Hossein Ardeschir Ghofrani, George Giannakoulas, David G Kiely, Eckhard Mayer, Gergely Meszaros, Blin Nagavci, Karen M Olsson, Joanna Pepke-Zaba, Jennifer K Quint, Göran Rådegran, Gerald Simonneau, Olivier Sitbon, Thomy Tonia, Mark Toshner, Jean Luc Vachiery, Anton Vonk Noordegraaf, Marion Delcroix, Stephan Rosenkranz, and ESC/ERS Scientific Document Group. 2022 ESC/ERS Guidelines for the diagnosis and treatment of pulmonary hypertension: Developed by the task force for the diagnosis and treatment of pulmonary hypertension of the European Society of Cardiology (ESC) and the European Respiratory Society (ERS). Endorsed by the International Society for Heart and Lung Transplantation (ISHLT) and the European Reference Network on rare respiratory diseases (ERN-LUNG). *European Heart Journal*, 43(38): 3618–3731, 08 2022.

Surinder Janda, Neal Shahidi, Kenneth Gin, and John Swiston. Diagnostic accuracy of echocardiography for pulmonary hypertension: a systematic review and meta-analysis. *Heart*, 97(8):612–622, 2011.

Hyewon Jeong, Collin M. Stultz, and Marzyeh Ghassemi. Deep metric learning for the hemodynamics inference with electrocardiogram signals. In *Machine Learning for Healthcare Conference*, 2023.

A. E. W. Johnson, T. J. Pollard, L. Shen, L. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(160035), 2016.

Liviu Klein, Marat Fudim, Mozziyar Etemadi, Robert Gordon, Anjan Tibrewala, Jaime Hernandez-Montfort, Patrick McCann, Lucas Zier, Kevin Shah, Allman Rollins, Darshak Karia, Arshed Quyyumi, Shweta R. Motiwala, Nikolaos Diakos, John Rommel, Andrew P. Ambrosy, Venu G. Ganti, Priyanka Soni, Karen Larimer, Andrew M. Carek, and Omer T. Inan. Noninvasive pulmonary capillary wedge pressure estimation in heart failure patients with the use of wearable sensing and ai. *JACC: Heart Failure*, 13(8):102513, 2025.

Emily Kogan, Eva-Maria Didden, Eileen Lee, Anderson Nnewihe, Dimitri Stamatiadis, Samson Mataraso, Deborah Quinn, Daniel Rosenberg, Christel Chehoud, and Charles Bridges. A machine learning approach to identifying patients with pulmonary hypertension using real-world electronic health records. *International Journal of Cardiology*, 374:95–99, 2023.

Andreas Leha, Kristian Hellenkamp, Bernhard Unsöld, Sitali Mushemi-Blake, Ajay M. Shah, Gerd Hasenfuß, and Tim Seidler. A machine learning approach for the prediction of pulmonary hypertension. *PLoS ONE*, 14(10), 2019.

Che Liu, Zhongwei Wan, Cheng Ouyang, Anand Shah, Wenjia Bai, and Rossella Arcucci. Zero-shot ECG classification with multimodal learning and test-time clinical knowledge enhancement. In *ICML*, 2024a.

Pang-Yen Liu, Shi-Chue Hsing, Dung-Jang Tsai, Chin Lin, Chin-Sheng Lin, Chih-Hung Wang, and Wen-Hui Fang. A deep-learning-enabled electrocardiogram and chest x-ray for detecting pulmonary arterial hypertension. *Journal of Imaging Informatics in Medicine*, 38:747–756, 2024b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

Dalton McLean, John Rommel, John A. Steuter, William S. Carroll, Mark Rabbat, Sudarshan Rajagopal, Venkatraman Srinivasan, Dean J. Kereiakes, Michael C. Roberts, Abhijit Raval, Navid Nemati, Farhad Fathieh, Timothy Burton, Horace R. Gillins, Ian Shadforth, Shyam Ramchandani, Charles R. Bridges, and Vallerie V. McLaughlin. Clinical validation of a machine-learned, point-of-care system to identify pulmonary hypertension. *European Respiratory Journal Open Research*, 2025.

Celia A. Melillo, James E. Lane, Kulwant S. Aulak, Allaa Almoushref, Raed A. Dweik, and Adriano R. Tonelli. Repeatability of pulmonary pressure measurements in patients with pulmonary hypertension. *Annals of the American Thoracic Society*, 17(8), 2020.

B. Moody, G. Moody, M. Villarroel, G. D. Clifford, and I. Silva. MIMIC-III Waveform Database (version 1.0). PhysioNet. RRID:SCR_007345., 2020. Accessed: Sept 16, 2025.

Aniruddh Raghu, Payal Chandak, Ridwan Alam, John Guttag, and Collin M. Stultz. Sequential multi-dimensional self-supervised learning for clinical time series. In *ICML*, 2023a.

Aniruddh Raghu, Daphne Schlesinger, Eugene Pomerantsev, Srikanth Devireddy, Pinak Shah, Joseph Garasic, John Guttag, and Collin M. Stultz. Ecg-guided non-invasive estimation of pulmonary congestion in patients with heart failure. *Nature Scientific Reports*, 13(3923), 2023b.

Stuart Rich, Gilbert E. D'alonzo, David R. Dantzker, and Paul S. Levy. Magnitude and implications of spontaneous hemodynamic variability in primary pulmonary hypertension. *The American Journal of Cardiology*, 55(1):159–163, 1985.

Muammar Sadrawi, Yin-Tsong Lin, Chien-Hung Lin, Bhekumuzi Mathunjwa, Ho-Tsung Hsin, Shou-Zen Fan, Maysam F. Abbod, and Jiann-Shing Shieh. Non-invasive hemodynamics monitoring system based on electrocardiography via deep convolutional autoencoder. *Sensors*, 21 (18), 2021.

Mahan Salehi, Samer Alabed, Michael Sharkey, Ahmed Maiter, Krit Dwivedi, Tarik Yardibi, Mona Selej, Abdul Hameed, Athanasios Charalampopoulos, David G. Kiely, and Andrew J. Swift. Artificial intelligence-based echocardiography assessment to detect pulmonary hypertension. *ERJ Open Research*, 11(3), 2025.

Yasar Sattar and Lovely Chhabra. Electrocardiogram. *StatPearls [Internet]*, 2023. URL `https://www.ncbi.nlm.nih.gov/books/NBK549803/`.

Daphne E. Schlesinger, Nathaniel Diamant, Aniruddh Raghu, Erik Reinertsen, Katherine Young, Puneet Batra, Eugene Pomerantsev, and Collin M. Stultz. A deep learning model for inferring elevated pulmonary capillary wedge pressures from the 12-lead electrocardiogram. *JACC: Advances*, 1(1), 2022.

Daphne E. Schlesinger, Ridwan Alam, Roey Ringel, Eugene Pomerantsev, Srikanth Devireddy, Pinak Shah, Joseph Garasic, and Collin M. Stultz. Artificial intelligence for hemodynamic monitoring with a wearable electrocardiogram monitor. *Nature Communications Medicine*, 5(4), 2025.

M Sharkey, K Karunasaagarar, C S Johns, S Rajaram, K Dwivedi, P Garg, D Alkhanfar, S Alabed, P Metherall, J Taylor, R J Van Der Geest, R Condliffe, M Mamalakis, D G Kiely, A J Swift, and K Dwivedi. Fully automatic deep learning pulmonary hypertension diagnosis using ct pulmonary angiography. *European Respiratory Journal*, 60(suppl 66), 2022.

O Slegg, J A Willis, C Gibson, A Kendler-Rhodes, F Wilkinson, J Rossdale, P Charters, R MacKenzie Ross, J D Pauling, J Easaw, K Carson, S R Kandan, G Robinson, J Suntharalingam, and D X Augustine. Accuracy of echocardiographic doppler measures of pulmonary hypertension compared with right heart catheterisation in a real world population referred to a specialist centre. *European Heart Journal*, 42, 10 2021.

Mohammod N. I. Suvon, Prasun C. Tripathi, Wenrui Fan, Shuo Zhou, Xianyuan Liu, Samer Alabed, Venet Osmani, Andrew J. Swift, Chen Chen, and Haiping Lu. Multimodal variational autoencoder for low-cost cardiac hemodynamics instability detection. In *MICCAI*, 2024.

Mohammod N. I. Suvon, Shuo Zhou, Prasun C. Tripathi, Wenrui Fan, Samer Alabed, Bishesh Khanal, Venet Osmani, Andrew J. Swift, Chen Chen, and Haiping Lu. Multimodal latent fusion of ecg leads for early assessment of pulmonary hypertension. *IEEE Journal of Biomedical and Health Informatics*, 2025.

Andrew J Swift, Haiping Lu, Johanna Uthoff, Pankaj Garg, Marcella Cogliano, Jonathan Taylor, Peter Metherall, Shuo Zhou, Christopher S Johns, Samer Alabed, Robin A Condliffe, Allan Lawrie, Jim M Wild, and David G Kiely. A machine learning cardiac magnetic resonance approach to extract disease features and automate pulmonary arterial hypertension diagnosis. *European Heart Journal - Cardiovascular Imaging*, 22(2):236–245, 01 2020.

Rahul Thapa, Bryan He, Magnus Ruud Kjaer, Hyatt Moore Iv, Gauri Ganjoo, Emmanuel Mignot, and James Zou. SleepFM: Multi-modal representation learning for sleep across brain activity, ECG and respiratory signals. In *ICML*, 2024.

Prasun C Tripathi, Sina Tabakhi, Mohammod N I Suvon, Lawrence Schöb, Samer Alabed, Andrew J Swift, Shuo Zhou, and Haiping Lu. Interpretable multimodal learning for cardiovascular hemodynamics assessment. *arXiv:2404.04718*, 2024.

Hongyan Xie, Xin Zhao, Nan Zhang, Jiayi Liu, Guang Yang, Yunshan Cao, Jialin Xu, Lei Xu, Zhonghua Sun, Zhaoying Wen, Senchun Chai, and Dongting Liu. Machine learning–based hemodynamics quantitative assessment of pulmonary circulation using computed tomographic pulmonary angiography. *Internation Journal of Cardiology*, 437:133457, 2025.

P G Yock and R L Popp. Noninvasive estimation of right ventricular systolic pressure by doppler ultrasound in patients with tricuspid regurgitation. *Circulation*, 70(4):657–662, 1984.

Wei Zhao, Zhihua Huang, Xiaolin Diao, Zhan Yang, Zhihui Zhao, Yun Xia, Qing Zhao, Zhaohong Sun, Qunying Xi, Yanni Huo, Ou Xu, Jiahui Geng, Xin Li, Anqi Duan, Sicheng Zhang, Luyang Gao, Yijia Wang, Sicong Li, Qin Luo, and Zhihong Liu. Development and validation of multimodal deep learning algorithms for detecting pulmonary hypertension. *npj Digital Medicine*, 8 (198), 2025.

A Echo Parameters

To train PHAROS+, we incorporated the echo parameters shown in Table 4. This table also shows the number of times each feature appeared in our dataset. When a parameter had more than one value recorded within a single echo procedure, the mean of the values was used as input.

B Filtering ECGs and Pulmonary Artery Pressure Waveforms for Quality and Alignment

Both the MGH and MIMIC-III datasets consisted of raw waveforms obtained from ICU bedside monitors. These waveforms can be noisy due to external factors such as motion from the patient or other artifacts. Sometimes patients could be disconnected from the monitor resulting in a flat or nonsensical signal. As a result, after separating the continuous waveforms into 10-second segments, we processed the data as follows:

- We removed all segments where the magnitude of the ECG voltage exceeded 5 mV or the max - min voltage was less than 0.1 mV.
- We removed all segments where the mPAP computed from the pulmonary artery pressure (PAP) waveform exceeded 80 mmHg, the minimum of the waveform was less than -10 mmHg, the mPAP was less than 0, the pulse pressure (max - min) was greater than 80, or the pulse pressure was less than 2 mmHg.
- We removed all segments where the number of peaks in the ECG did not match that of the PAP waveform.
  1. We performed R-peak detection on the ECG segment using NeuroKit2, specifically the `ecg_clean` and `ecg_peaks` functions with the default keyword arguments.
  2. For the PAP waveform, we used the function `scipy.signal.find_peaks()` with keyword arguments `prominence=10` and `distance=sampling_rate / 4`, where `sampling_rate` is the signal frequency in each dataset and 4 is an estimate of the maximum possible beats per second obtained from 240 beats per minute divided by 60 seconds per minute.
  3. If the difference in number of peaks between the ECG and PAP waves exceeded 1, we discarded the segment.
- We removed all segments where the mPAP range was too large.
  1. We found the R-peaks in the ECG as previously described.

Measurement	Count
AORTIC SINUS INDEX POST 1	11946
RIGHT ATRIUM INDEX MEDIAL-LATERAL (1)	8795
BODY SURFACE AREA (1)	7528
EJECTION FRACTION (1)	7499
LEFT VENTRICLE INTERNAL DIAMETER END DIASTOLE (1)	7447
LEFT VENTRICLE INTERNAL DIAMETER END SYSTOLE (1)	7390
LEFT VENTRICLE EJECTION FRACTION - QUINES RAW (1)	7389
INTERVENTRICULAR SEPTUM THICKNESS (1) MM	7260
LEFT VENTRICLE POSTERIOR WALL THICKNESS (1) MM	7259
LEFT ATRIUM DIMENSION ANTERIOR-POSTERIOR (1)	6640
LEFT VENTRICLE APICAL CONTRIBUTION (1)	5949
WEIGHT	5929
HEIGHT	5914
PHS CV ECHO AV SINUS INDEX 1	5806
ASCENDING AORTA DIAMETER (1) MM	5173
RIGHT ATRIUM PRESSURE ESTIMATED (1)	5096
RIGHT VENTRICLE TO RIGHT ATRIUM PRESSURE GRADIENT (1)	4995
TRICUSPID VALVE PEAK VELOCITY (1)	4995
RIGHT VENTRICLE PEAK SYSTOLIC PRESSURE (1)	4966
ASCENDING AORTA INDEX 1	4902
ASCENDING AORTA INDEX POST 1	4902
LEFT ATRIUM DIMENSION SUPERIOR-INFERIOR (1) MM	4567
LEFT ATRIUM DIMENSION MEDIAL-LATERAL (1) MM	4537
RIGHT ATRIUM DIMENSION SUPERIOR-INFERIOR (1)	4533
RIGHT ATRIUM INDEX SUPERIOR-INFERIOR POST	4340
RIGHT ATRIUM INDEX SUPERIOR-INFERIOR (1)	4340
RIGHT ATRIUM INDEX MEDIAL-LATERAL POST	4301
RIGHT VENTRICLE LINEAR DIMENSION (1) MM	3451
LEFT ATRIAL VOLUME (1)	2652
LEFT ATRIAL VOLUME INDEX (1)	2632
INFERIOR VENA CAVA DIAMETER (1) MM	1877
LEFT VENTRICULAR OUTFLOW TRACT VELOCITY (1)	1300
RIGHT VENTRICLE PULSE DOPPLER S WAVE (1)	1284
RIGHT VENTRICLE TAPSE (1)	1251
AORTIC VALVE PEAK GRADIENT (1)	810
AORTIC VALVE MEAN GRADIENT (1)	780
MITRAL VALVE PEAK GRADIENT (1)	737
MITRAL VALVE MEAN GRADIENT (1)	734
AORTIC VALVE PROSTHETIC PEAK GRADIENT (1)	640
AORTIC VALVE PROSTHETIC MEAN GRADIENT (1)	635
MITRAL VALVE GRADIENT HR (1)	614
LEFT VENTRICULAR OUTFLOW TRACT DIAMETER (1)	359
AORTIC VALVE AREA (1)	320
AORTIC VALVE AREA INDEX (1)	298
RIGHT VENTRICLE BASAL DIAMETER (1)	281
MITRAL VALVE PROSTHETIC MEAN GRADIENT (1)	280
MITRAL VALVE PROSTHETIC PEAK GRADIENT (1)	280
RIGHT VENTRICLE FRACTIONAL AREA CHANGE (FAC) (1)	271
PERICARDIUM EFFUSION DIMENSION 1 (1) MM	261
AORTIC VALVE DISTANCE TO CANNULA TIP MM	249
TRICUSPID VALVE PROSTHETIC MEAN GRADIENT (1)	207
TRICUSPID VALVE PROSTHETIC PEAK GRADIENT (1)	207
LEFT VENTRICULAR OUTFLOW TRACT TIME VELOCITY INTEGRAL (1)	156
PULMONARY ARTERY MAIN - DIMENSION (1)	137
PERICARDIUM EFFUSION DIMENSION 2 (1) MM	126
PULMONARY ARTERY RIGHT - DIMENSION (1)	121
PULMONARY ARTERY LEFT - DIMENSION (1)	108

Table 4: Echo measurements used in PHAROS+ and the number of measurements for each feature across the full dataset.

2. Using the indices of a consecutive pair of R-peaks, we found the corresponding sections in the PAP waveform. Each section corresponds to the PAP wave for one heart beat.
3. Using the PAP waveform sections, we computed the mPAP for each heart beat.
4. If any of the mPAPs were outside the range of $\pm 3\sigma_{mPAP}$, where σ_{mPAP} is the standard deviation, we discarded the segment.

- We removed all segments where the systolic PAP, diastolic PAP, and mPAP were inconsistent.
  1. We found the R-peaks in the ECG as previously described.
  2. Using the indices of a consecutive pair of R-peaks, we found the corresponding sections in the PAP waveform. Each section corresponds to the PAP wave for one heart beat.
  3. Using the PAP waveform sections, we computed the mPAP and systolic and diastolic PAPs for each heart beat.
  4. If any heart beat had a corresponding systolic PAP less than or equal to the mPAP or diastolic PAP greater than or equal to the mPAP, the segment was discarded.

C Model Architecture

The detailed architecture of PHAROS+ is shown below. PHAROS follows the same architecture, but without the MLP encoding the echo parameters, and with a MLP decoder that takes in 320 features instead of 352.

```
PHAROSPlus(
  (encoder): ECGPairPADeltatEncoder(
    (ecg_encoder): CNN1D(
      (conv1): Conv1d(1, 64, kernel_size=(17,), stride=(1,), padding=same, bias=False)
      (bn1): BatchNorm1d(64, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
      (res_blocks): ModuleList(
        (0): ResidualBlock(
          (conv1): Conv1d(64, 256, kernel_size=(17,), stride=(1,), padding=same, bias=False)
          (bn1): BatchNorm1d(256, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
          (conv2): Conv1d(256, 256, kernel_size=(17,), stride=(12,), bias=False)
          (bn2): BatchNorm1d(256, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
          (pool): MaxPool1d(kernel_size=12, stride=12, padding=0, dilation=1, ceil_mode=False)
          (identity_layer): Conv1d(64, 256, kernel_size=(1,), stride=(1,), bias=False)
        )
      )
      (global_avg_pool): AdaptiveAvgPool1d(output_size=1)
    )
    (ecg_fuser): Sequential(
      (0): Linear(in_features=512, out_features=512, bias=True)
      (1): ReLU()
      (2): Dropout(p=0.4, inplace=False)
      (3): Linear(in_features=512, out_features=256, bias=True)
    )
    (hemodynamics_encoder): Sequential(
      (0): Linear(in_features=1, out_features=32, bias=True)
      (1): ReLU()
      (2): Dropout(p=0.4, inplace=False)
      (3): Linear(in_features=32, out_features=32, bias=True)
    )
    (time_encoder): Sequential(
```

```
      (0): Linear(in_features=1, out_features=32, bias=True)
      (1): ReLU()
      (2): Dropout(p=0.4, inplace=False)
      (3): Linear(in_features=32, out_features=32, bias=True)
    )
    (echo_parameters_encoder): Sequential(
      (0): Linear(in_features=114, out_features=64, bias=True)
      (1): ReLU()
      (2): Dropout(p=0.4, inplace=False)
      (3): Linear(in_features=64, out_features=32, bias=True)
    )
  )
  (decoder): MLP(
    (layers): Sequential(
      (0): Linear(in_features=352, out_features=256, bias=True)
      (1): ReLU()
      (2): Dropout(p=0.4, inplace=False)
      (3): Linear(in_features=256, out_features=256, bias=True)
      (4): ReLU()
      (5): Dropout(p=0.4, inplace=False)
      (6): Linear(in_features=256, out_features=1, bias=True)
    )
  )
)
```

D Model Performance at Different $\Delta mPAP$ Thresholds

In this work, we define disease progression as an increase in mPAP greater than 4 mmHg and trained our models with this threshold. In practice, however, other thresholds can be informative as well. For instance, in some cases, an increase in mPAP can expected, such as when withdrawing medication (e.g., due to adverse events). A model trained with a larger threshold could be useful for ruling out increases in mPAP beyond a certain value. We studied the impact of the $\Delta mPAP$ classification threshold on performance by training different PHAROS+ models with thresholds ranging from 0 to 7 mmHg. The results are plotted in Figure 4. As the threshold increases, AUC increases. This is likely because larger changes in mPAP are more distinguishable in the ECGs, leading to better performance.

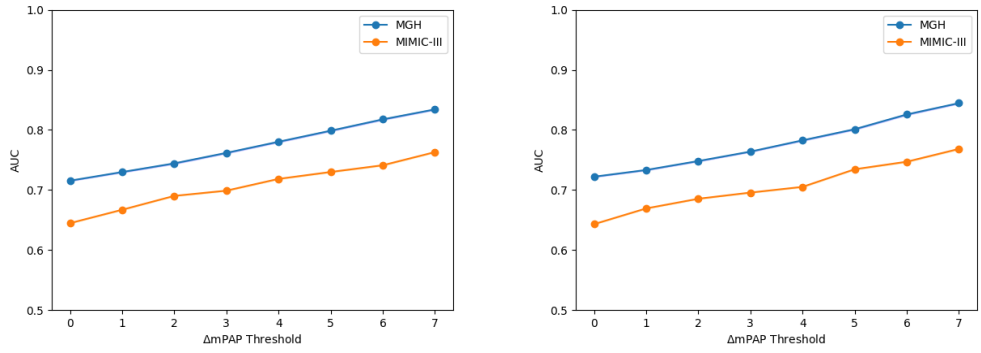

Figure 4: AUC of PHAROS (left) and PHAROS+ (right) when trained at different $\Delta mPAP$ thresholds. Error bars are present but are too small to see.

E Ablation Study of Δt and m_i

An ablation study of Δt and the baseline mPAP m_i was conducted. The results are presented in Table 5. We see that both Δt and m_i improved AUC, with m_i contributing much more to the results

than Δt. This makes sense because the dataset we used to train our models consisted of ICU stays with short time scales, an average duration of 1.7 days, as shown in Table 1. Within such close time intervals, the baseline m_i would be a stronger predictor of changes in mPAP than Δt.

Δt	m_i	AUC (↑)	Brier (↓)	AUPRC (↑)	PPV (↑)	NPV (↑)
✗	✗	0.6081 $\pm$ 0.0007	0.1567 $\pm$ 0.0003	0.2704 $\pm$ 0.0010	0.2281 $\pm$ 0.0005	0.8671 $\pm$ 0.0008
✓	✗	0.6101 $\pm$ 0.0009	0.1565 $\pm$ 0.0004	0.2720 $\pm$ 0.0015	0.2297 $\pm$ 0.0009	0.8681 $\pm$ 0.0006
✗	✓	0.7812 $\pm$ 0.0004	0.1299 $\pm$ 0.0003	0.5025 $\pm$ 0.0014	0.3315 $\pm$ 0.0012	0.9219 $\pm$ 0.0005
✓	✓	0.7827 $\pm$ 0.0009	0.1296 $\pm$ 0.0002	0.4979 $\pm$ 0.0019	0.3305 $\pm$ 0.0009	0.9211 $\pm$ 0.0004

Table 5: Ablation study of the Δt and baseline mPAP (m_i) model inputs in PHAROS+. The results shown are for the MGH dataset. (↑ indicates higher is better and ↓ indicates lower is better.)