Multimodal Variational Autoencoder for Low-cost Cardiac Hemodynamics Instability Detection

Mohammod N. I. Suvon[1,2(✉)], Prasun C. Tripathi[1,7], Wenrui Fan[1,2], Shuo Zhou[1,2], Xianyuan Liu[1,2], Samer Alabed[3,4,5], Venet Osmani[6], Andrew J. Swift[3,4,5], Chen Chen[1,8,9], and Haiping Lu[1,2,5]

[1] Department of Computer Science, University of Sheffield, Sheffield, UK
[2] Centre for Machine Intelligence, University of Sheffield, Sheffield, UK
[3] Department of Infection, Immunity and Cardiovascular Disease, University of Sheffield, Sheffield, UK
[4] Department of Clinical Radiology, Sheffield Teaching Hospitals, Sheffield, UK
[5] INSIGNEO, Institute for in Silico Medicine, University of Sheffield, Sheffield, UK
[6] Information School, University of Sheffield, Sheffield, UK
[7] Department of Computer Science, IITRAM, Gujarat, India
[8] Department of Engineering Science, University of Oxford, Oxford, UK
[9] Department of Computing, Imperial College London, London, UK
{m.suvon(✉), p.c.tripathi , wenrui.fan, shuo.zhou, xianyuan.liu, s.alabed, v.osmani, a.j.swift, chen.chen2, h.lu}@sheffield.ac.uk

Abstract. Recent advancements in non-invasive detection of cardiac hemodynamic instability (CHDI) primarily focus on applying machine learning techniques to a single data modality, e.g. cardiac magnetic resonance imaging (MRI). Despite their potential, these approaches often fall short especially when the size of labeled patient data is limited, a common challenge in the medical domain. Furthermore, only a few studies have explored multimodal methods to study CHDI, which mostly rely on costly modalities such as cardiac MRI and echocardiogram. In response to these limitations, we propose a novel multimodal variational autoencoder (CardioVAE$_{\mathrm{X,G}}$) to integrate low-cost chest X-ray (C**X**R) and electrocardiogram (EC**G**) modalities with pre-training on a large unlabeled dataset. Specifically, CardioVAE$_{\mathrm{X,G}}$ introduces a novel tri-stream pre-training strategy to learn both shared and modality-specific features, thus enabling fine-tuning with both unimodal and multimodal datasets. We pre-train CardioVAE$_{\mathrm{X,G}}$ on a large, unlabeled dataset of 50,982 subjects from a subset of MIMIC database and then fine-tune the pre-trained model on a labeled dataset of 795 subjects from the ASPIRE registry. Comprehensive evaluations against existing methods show that CardioVAE$_{\mathrm{X,G}}$ offers promising performance (AUROC = 0.79 and Accuracy = 0.77), representing a significant step forward in non-invasive prediction of CHDI. Our model also excels in producing fine interpretations of predictions directly associated with clinical features, thereby supporting clinical decision-making.

Keywords: Cardiac hemodynamics instability, Variational autoencoder, Multimodal learning, Interpretable model

1 Introduction

Cardiac hemodynamic instability (CHDI) can lead to unreliable and inefficient cardiovascular function and even heart failure. Pulmonary Artery Wedge Pressure (PAWP) is an important surrogate marker for detecting the severity of CHDI and heart failure. Elevated PAWP indicates left ventricular filling pressure and reduced contractility of the heart [3]. It can be precisely measured by invasive and expensive right heart catheterization (RHC). However, simpler and non-invasive methods are often required to monitor critical patients. In recent years, several machine learning-based and/or deep learning-based methods were developed for PAWP prediction from medical images acquired by non-invasive technique [3, 26, 27]. It has been shown that it is possible to predict PAWP not only from high-cost, high precision scans such as cardiac magnetic resonance imaging (MRI) [3] and echocardiography [26], but also from more affordable, accessible scans or measurements such as chest X-rays (CXR) [8, 15, 16] and electrocardiogram (ECG) [22, 23].

Most aforementioned studies focus mainly on extracting measurements from *one, single* data modality. More recently, multimodal learning has demonstrated superior diagnostic performance to unimodal learning approaches [25, 28, 31]. For example, Tripathi et al. [27] developed a tensor-based multimodal pipeline to combine features from cardiac MRI imaging with cardiac measurement for higher precision. In this work, we developed a multimodal model utilizing only low-cost modalities, e.g., CXR and ECG, to improve the performance of PAWP prediction. Such a low-cost approach is more relevant and feasible for broader adoption in low-income countries with limited access to MRI scans. The two main challenges in this low-cost approach are a) the limited information in 2D CXR images and 1D ECG signals compared to MRI and ultrasound and b) the scarcity of PAWP measurements in the labeled dataset relating to the difficulty and expense of obtaining precise and *invasive* PAWP.

To tackle the above challenges, we first develop a cardiac multimodal variational autoencoder (CardioVAE$_{\mathrm{X,G}}$), aiming at maximizing the value from C**X**R images and EC**G**s through joint training. Specifically, CardioVAE$_{\mathrm{X,G}}$ is first pre-trained using a novel tri-stream multimodal pre-training strategy, leveraging the value from a large-scale public, *unlabeled* paired CXR and ECG data to reduce the need for large-scale, *labeled* data. We then fine-tune the pre-trained model on a relatively small dataset for the PAWP prediction task. The **main contributions** of this paper are three-fold: 1) We introduce CardioVAE$_{\mathrm{X,G}}$, a novel multimodal variational autoencoder (Fig. 1) for low-cost, non-invasive PAWP prediction. Unlike previous approaches that rely on expensive modalities such as cardiac MRI, CardioVAE$_{\mathrm{X,G}}$ efficiently integrates lower-cost CXR and ECG modalities. 2) We develop a novel tri-stream pre-training strategy with CardioVAE$_{\mathrm{X,G}}$ model to learn both shared and modality-specific features. This approach enables our CardioVAE$_{\mathrm{X,G}}$ model to be fine-tuned with unimodal or multimodal datasets, greatly improving usability. 3) We performed extensive experiments to show the promising performance of our model. Another unique feature is that CardioVAE$_{\mathrm{X,G}}$ is capable of providing explainable feature visu-

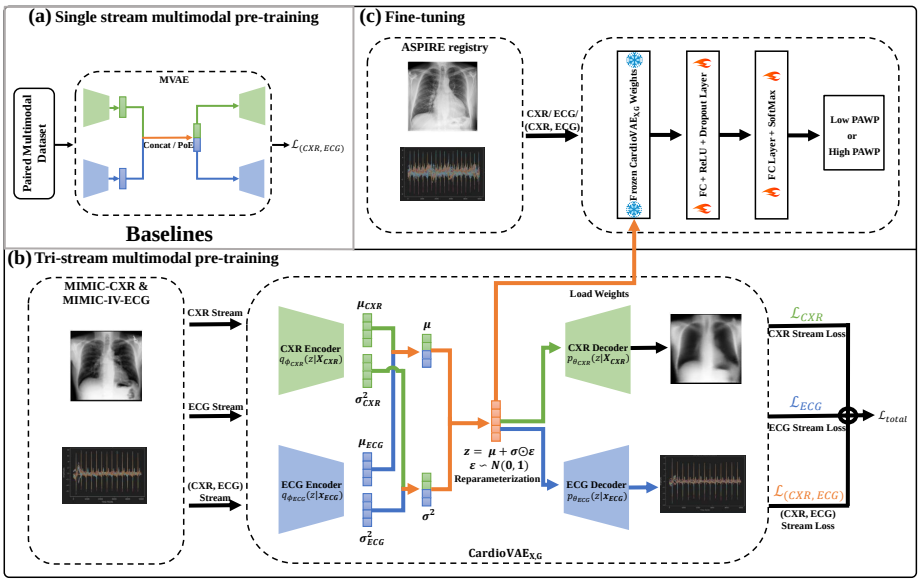


Fig. 1: The baselines and proposed CardioVAE$_{\text{X,G}}$ for PAWP prediction. (a) Top left: the single stream MVAE baselines [30, 18] require pre-training on paired data (CXR, ECG) and utilize only the shared features using concatenation or product of expert (PoE) [7] based multimodal fusion methods. (b) Bottom: Tri-stream multimodal pre-training that can learn both shared and modality-specific features. (c) Top right: fine-tuning on ASPIRE registry. Due to the tri-stream flow, our model can be fine-tuned on a single modality or both modalities.

alization for the interpretation of the clinical decisions (see Fig. 2), enhancing their applicability and reliability in real-world scenarios.

2 Methods

The proposed CardioVAE$_{\text{X,G}}$ for PAWP prediction is depicted in Fig. 1. It leverages low-cost CXR and ECG modalities for the prediction. Our model utilizes tri-stream multimodal pre-training, incorporating data streams from three sources: CXR, ECG, and the modality pair (CXR, ECG). Tri-stream setting allows our model to learn modality-specific features along with shared features in contrast with the baseline MVAE [30] that utilizes only the shared features of CXR and ECG. In the following, we discuss each building block of our model.

CXR Encoder: Our CXR encoder incorporates a convolutional neural network (CNN) consisting of three Convolutional (CONV) layers. It processes a CXR image $\mathbf{X}_{\text{CXR}} \in \mathbb{R}^{H \times W \times C}$, where H, W, and C represent the height, width, and number of channels, respectively. Each CONV layer is configured with a 3×3 kernel, a stride of 2, and padding of 1. We use channel depths of 16, 32, and

64 in CONV layers. Following this, the data is flattened and passed through two Fully Connected (FC) layers to produce the mean ($\boldsymbol{\mu}_{\text{CXR}}$) and variance ($\boldsymbol{\sigma^2}_{\text{CXR}}$) vectors, which are crucial for defining the CXR image's latent space representation and establishing its approximate posterior distribution parameters $q_{\phi_{\text{CXR}}}(\mathbf{z}_{\text{CXR}}|\mathbf{X}_{\text{CXR}})$.

ECG Encoder: The ECG encoder comprises three one-dimensional (1D) CONV layers. It processes the sequential ECG signal $\mathbf{x}_{\text{ECG}} \in \mathbb{R}^L$, where L denotes the length of the signal. Each 1D CONV layer is configured with a kernel size of 1×3, a stride of 2, and padding of 1. We use channel depths of 16, 32, and 64 in three CONV layers. Following this, the signal is flattened and processed through two FC layers, yielding the mean ($\boldsymbol{\mu}_{\text{ECG}}$) and variance ($\boldsymbol{\sigma^2}_{\text{ECG}}$) which are crucial for the ECG signal's latent space representation and its approximate posterior distribution $q_{\phi_{\text{ECG}}}(\mathbf{z}_{\text{ECG}}|\mathbf{x}_{\text{ECG}})$.

Multimodal Integration: In multimodal integration, we employ a Product of Experts (PoE) [7] approach to combine the approximate posterior distributions from the CXR and ECG encoders with a standard Gaussian prior $p(z) = \mathcal{N}(0, \mathbf{I})$ into a unified latent variable, effectively synthesizing the individual expert opinions, where $\mathbf{I}$ is the identity matrix. The combined mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma^2}$ of the latent variable [10] are computed as: $\boldsymbol{\mu} = (\sum_{m=1}^{M} \boldsymbol{\mu_m}/\boldsymbol{\sigma_m^2})/(\sum_{m=1}^{M} 1/\boldsymbol{\sigma_m^2})$ and $\boldsymbol{\sigma^2} = 1/(\sum_{m=1}^{M} 1/\boldsymbol{\sigma_m^2})$, where m represents the modality, ranging from 1 to M. This leads to the calculation of the latent space variable $\mathbf{z}$ using the reparameterization trick $\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \cdot \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is drawn from $\mathcal{N}(0, \mathbf{I})$. This step allows the decoders to generate diverse yet consistent reconstructions, effectively establishing the approximate joint posterior distribution $q_{\phi_{\text{CXR, ECG}}}(\mathbf{z}_{\text{CXR, ECG}}|\mathbf{X}_{\text{CXR}}, \mathbf{x}_{\text{ECG}})$ and reflecting a deep understanding of the merged input data.

Decoder Design: The CXR and ECG decoders aim to reconstruct inputs from the unified latent space $\mathbf{Z}$, which contains both modality-specific latent variables $\mathbf{z}_{\text{CXR}}$, $\mathbf{z}_{\text{ECG}}$, and shared latent variable $\mathbf{z}_{\text{CXR,ECG}}$. For CXR image reconstruction, the CXR decoder $p_{\theta_{\text{CXR}}}(\mathbf{X}_{\text{CXR}}|\mathbf{z}_{\text{CXR}})$ is a CNN consisting of three transposed convolutional layers, effectively restoring the sample $\mathbf{X}_{\text{CXR}}$ from the latent variable $\mathbf{z}_{\text{CXR}}$. Similarly, the ECG decoder $p_{\theta_{\text{ECG}}}(\mathbf{x}_{\text{ECG}}|\mathbf{z}_{\text{ECG}})$ utilizes three 1D transposed convolutional layers to convert $\mathbf{z}_{\text{ECG}}$ into the precise temporal waveform $\mathbf{x}_{\text{ECG}}$. This reconstruction process leverages the inherent flexibility of the Gaussian distribution within the latent space, enabling the production of diverse, high-fidelity reconstructions by sampling various $\mathbf{z}$ points. For the unimodal streams, the decoder uses their respective modality-specific latent variables, but for the multimodal stream, the decoder uses both modality-specific and shared latent variables for reconstruction.

Tri-stream Pre-training Strategy: For pre-training, we extend [12, 30] to introduce a tri-stream strategy that processes CXR and ECG data both separately and jointly to capture the modality-specific and shared features of each modality in the latent space. We pass three data streams through our CardioVAE$_{\text{X, G}}$ model: 1) CXR only, 2) ECG only, and 3) paired CXR and ECG, and calculate three separate losses for each stream by incorporating Evidence Lower Bound (ELBO) [14]. We define $\mathcal{L}_{\text{CXR}}$ for CXR, $\mathcal{L}_{\text{ECG}}$ for ECG, and $\mathcal{L}_{(\text{CXR,ECG})}$ for

joint loss, as follows:

$$\mathcal{L}_{\text{CXR}} = \mathbb{E}_{q_{\phi_{\text{CXR}}}(\mathbf{z}_{\text{CXR}}|\mathbf{X}_{\text{CXR}})}[\lambda_{\text{CXR}} \log p_{\theta_{\text{CXR}}}(\mathbf{X}_{\text{CXR}}|\mathbf{z}_{\text{CXR}})] - \beta D_{\text{KL}}[q_{\phi_{\text{CXR}}}(\mathbf{z}_{\text{CXR}}|\mathbf{X}_{\text{CXR}})||p(z)], \quad (1)$$

$$\mathcal{L}_{\text{ECG}} = \mathbb{E}_{q_{\phi_{\text{ECG}}}(\mathbf{z}_{\text{ECG}}|\mathbf{x}_{\text{ECG}})}[\lambda_{\text{ECG}} \log p_{\theta_{\text{ECG}}}(\mathbf{x}_{\text{ECG}}|\mathbf{z}_{\text{ECG}})] - \beta D_{\text{KL}}[q_{\phi_{\text{ECG}}}(\mathbf{z}_{\text{ECG}}|\mathbf{x}_{\text{ECG}})||p(z)], \quad (2)$$

$$\begin{aligned}\mathcal{L}_{(\text{CXR,ECG})} = {} & \mathbb{E}_{q_{\phi_{\text{CXR}}}(\mathbf{z}_{\text{CXR}}|\mathbf{X}_{\text{CXR}})}[\lambda_{\text{CXR}} \log p_{\theta_{\text{CXR}}}(\mathbf{X}_{\text{CXR}}|\mathbf{z}_{\text{CXR}})] \\ & + \mathbb{E}_{q_{\phi_{\text{ECG}}}(\mathbf{z}_{\text{ECG}}|\mathbf{x}_{\text{ECG}})}[\lambda_{\text{ECG}} \log p_{\theta_{\text{ECG}}}(\mathbf{x}_{\text{ECG}}|\mathbf{z}_{\text{ECG}})] \\ & - \beta D_{\text{KL}}[q_{\phi_{(\text{CXR,ECG})}}(\mathbf{z}_{(\text{CXR, ECG})}|\mathbf{X}_{\text{CXR}}, \mathbf{x}_{\text{ECG}})||p(z)], \end{aligned} \quad (3)$$

where the first part $\mathbb{E}_{q_\phi}$ of the losses is the expected conditional log-likelihood of the data given the latent variable $\mathbf{z}$, indicating how well the model reconstructs the data. This part takes the reconstructed output from the decoder and finds the reconstruction error. The second part of the losses, D_{KL}, is the Kullback-Leibler (KL) divergence [2] between the approximate posterior (CXR or ECG or CXR-ECG) and the prior distribution $p(z)$, following the Gaussian distribution, over the latent variables, calculated in encoders and multimodal integration phase. These losses aim to maximize the likelihood of the data while minimizing the difference between the approximate posterior and the prior. Moreover, in our losses, we incorporate two important hyperparameters λ [17] to balance the weights of modalities and β [6] to balance the trade-off between reconstruction loss and KL divergence. In practice, $\lambda = 1$ and β are slowly annealed from 0 to 1 to form a valid lower bound on the evidence [1]. After calculating three losses in Eq. (1-3) for the three streams, we combine them to obtain a total loss $\mathcal{L}_{\text{total}}$ as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CXR}} + \mathcal{L}_{\text{ECG}} + \mathcal{L}_{(\text{CXR,ECG})}. \quad (4)$$

The tri-stream methodology enhances the model's capability to accurately represent and reconstruct multimodal data, leveraging the strengths of each modality towards a comprehensive representation of the multimodal inputs for improved overall performance.

Fine-tuning Strategy: Applying the pre-trained model to the CXR-ECG classification task is straightforward, as tri-stream pre-training enables the model to learn both modality-specific and shared features. During fine-tuning, the new downstream dataset is passed through the frozen model to extract features which are subsequently processed by two FC layers. The binary cross-entropy loss [4] is used for fine-tuning our model.

Table 1: Patients characteristics of included patients in ASPIRE registry dataset. p-values were obtained using t-test [29].

	Low PAWP(≤ 15)	High PAWP(> 15)	p-value
Number of patients	560	235	-
Age (in years)	58.98 ± 15.30	62.62 ± 13.75	0.0017
Body Surface Area (BSA)	1.92 ± 0.26	1.98 ± 0.24	0.0025
Heart Rate (bpm)	78.87 ± 13.87	74.75 ± 14.93	0.0002
Pulmonary Artery Pressure	42.78 ± 14.01	44.86 ± 12.31	0.0483
Pulmonary Artery Systolic Pressure	71.35 ± 24.00	75.86 ± 23.75	0.0155
Pulmonary Artery Diastolic Pressure	25.06 ± 10.13	26.82 ± 7.87	0.0176
PAWP (mmHg)	9.94 ± 3.06	19.42 ± 3.51	< 0.0001

3 Experimental Results and Analysis

Dataset for Pre-training: We pre-trained our CardioVAE$_{\text{X,G}}$ model using two datasets, MIMIC-CXR [11] and MIMIC-IV-ECG [5], by pairing them via unique patient ID and time. This gave us $50,982$ pairs of CXR-ECG samples.
Study Population and Dataset for Downstream Task: We evaluated all models using a dataset from the ASPIRE registry [9] for the detection of CHDI via PAWP prediction in patients with suspected pulmonary hypertension. The local institutional review board and ethics committee approved this study. A total of 795 patients who underwent RHC, CXR, and ECG were included in this study. Based on the measurements from RHC (using a balloon-tipped 7.5 French thermodilution catheter), we found 560 patients with normal PAWP (≤ 15 mmHg), and 235 with elevated PAWP (> 15 mmHg). Table 1 summarizes the patient characteristics of the used ASPIRE registry dataset.
Experimental Design: We converted CXR and 12-lead ECG data to 2D images (224×224) and 1D signals ($1 \times 60,000$), respectively. We pre-trained our model with unlabeled MIMIC subset [5, 11] by partitioning it with a $90 : 10$ ratio for training and validation sets. The hyperparameters for λ_{CXR} and λ_{ECG} were selected using grid search. The optimal hyperparameters were then used to pre-train CardioVAE$_{\text{X,G}}$ model on the whole MIMIC subset. We used the Fréchet inception distance [20] to assess the performance of pre-training. For pre-training, we used Adam optimizer with a learning rate of 0.001 and a batch size of 128, and trained for 100 epochs to ensure the model convergence.

For fine-tuning, we froze the layers in the encoders and fine-tuned FC layers on the ASPIRE registry dataset. We used 128 nodes in FC layer and used a dropout of 0.5. We evaluated the prediction performance using 10-fold cross-validation with a training and validation ratio of $80 : 20$. We set the learning rate to 0.001 and the batch size to 32, and trained the model for 50 epochs. We used Area Under Receiver Operating Curve (AUROC) and accuracy metrics to assess classification performance. Moreover, we calculated the p-values of our best-performing model against other models to show the statistical significance of the results. For a fair comparison, we used the same training settings and data partitioning for comparing methods [16, 18, 23, 30]. An Nvidia RTX4090

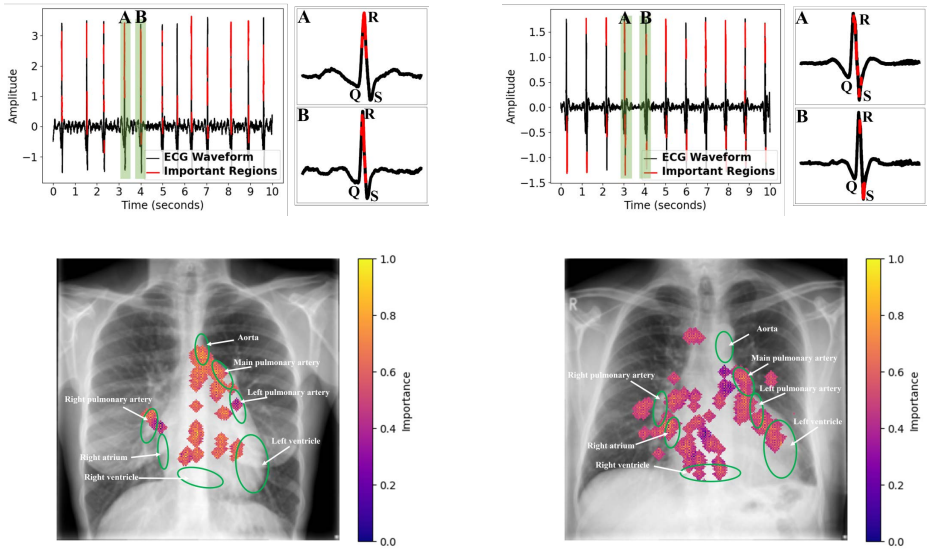
Table 2: Performance comparison using two metrics (with **best** in bold). CM: Cardiac Measurements from cardiac MRI, CMRI (4ch): Four-chamber cardiac MRI. Garg et al. [3]$^{\blacklozenge}$ and Tripathi et al. [27]$^{\blacklozenge}$ were tested on a different cohort and included here for reference only.

Study Type	Modality(s)	Method	AUROC	p-value$_{\text{AUC}}$	Accuracy
Unimodal	CXR	CNN	0.638 ± 0.06	< 0.0001	0.675 ± 0.03
		Kusunose et al. [16]	0.581 ± 0.04	< 0.0001	0.695 ± 0.06
		CardioVAE$_{\text{X,G}}$ (ours)	$\mathbf{0.681 \pm 0.05}$	< 0.0001	$\mathbf{0.709 \pm 0.04}$
	ECG	CNN	0.724 ± 0.05	< 0.0021	0.707 ± 0.05
		Schlesinger et al. [23]	0.670 ± 0.03	< 0.0001	0.703 ± 0.04
		CardioVAE$_{\text{X,G}}$ (ours)	$\mathbf{0.744 \pm 0.05}$	0.0226	$\mathbf{0.727 \pm 0.04}$
	CM	Garg et al. [3]$^{\blacklozenge}$	0.730 ± 0.04	-	0.740 ± 0.03
Multimodal	CMRI (4ch) & CM	Tripathi et al. [27]$^{\blacklozenge}$	0.813 ± 0.02	-	0.792 ± 0.02
	CXR & ECG	CNN	0.748 ± 0.05	0.0352	0.735 ± 0.03
		Li et al. [18]	0.737 ± 0.05	0.0101	0.724 ± 0.04
		Wu et al. [30]	0.758 ± 0.03	0.0283	0.756 ± 0.04
		CardioVAE$_{\text{X,G}}$ (ours)	$\mathbf{0.790 \pm 0.03}$	-	$\mathbf{0.772 \pm 0.04}$

GPU was used for all experiments. The implementation of all the models was carried out in Python (version 3.10) with PyTorch [21]. The source code can be found at: `https://github.com/Shef-AIRE/AI4Cardiothoracic-CardioVAE`.

Unimodal Study: Table 2 compared the results of unimodal models for CXR and ECG in rows $2-7$. We considered Kusunose et. al.'s method [16] as a baseline for CXR modality and Schlesinger et al.'s method [23] as a baseline for ECG. Our CardioVAE$_{\text{X,G}}$ fine-tuning on unimodal data outperformed other unimodal methods for both modalities. The results show that CardioVAE$_{\text{X,G}}$ obtains improvements of ΔAUROC $= 0.100$ and ΔAccuracy $= 0.014$ over the CXR baseline [16], and the improvements of ΔAUROC $= 0.074$ and ΔAccuracy$= 0.024$ over ECG baseline [23]. The baseline methods do not use the unsupervised pre-training of models. Our CardioVAE$_{\text{X,G}}$ leverages pre-training from a large unlabeled dataset, and learns modality-specific features for the inference, enabling it to achieve better performance. To show the effect of pre-training, unimodal CNN models (row 1 and 4) are included in Table 2 which use the same encoders and classification layers as in our CardioVAE$_{\text{X,G}}$ without pre-training. The results show that pre-training is important in achieving higher performance. We also included the current baseline model for high-cost cardiac MRI unimodal for PAWP prediction (row 7), which was tested on a different cardiac MRI cohort and not for direct comparison with our low-cost unimodal models.

Multimodal Study: We compared CardioVAE$_{\text{X,G}}$ fine-tuned with multimodal (CXR & ECG) data against four competing methods (rows $8-11$) in Table 2. Li et al. [18] used feature concatenation to combine two modalities in their multimodal variational autoencoder. Wu et al. [30] utilized Product of Expert (PoE) based fusion. These two methods are based on a single-stream approach. Our CardioVAE$_{\text{X,G}}$ outperformed these two methods. Therefore, the tri-stream strategy in our model is effective for learning unique modality-specific along with shared features for prediction. Additionally, obtained p-values show that



(a) Normal PAWP Patient

(b) Elevated PAWP Patient

Fig. 2: Interpretability of CardioVAE$_{X,G}$ for two subjects using integrated gradients method [24]. (a) 1D ECG (top left) and CXR (bottom left) for normal PAWP, (b) 1D ECG (top right) and CXR (bottom right) for elevated PAWP. Green annotations on CXRs highlight seven regions of the heart and lungs, marked by an expert clinician for enhanced visualization of key areas. The 1D ECG signal was smoothed with NeuroKit2 [19] library for better visualization.

our best-performing model produces statistically significant results against other models. We also included the comparison with the multimodal pipeline in Tripathi et al. [27] which was tested on a different cardiac MRI cohort. The results show that our model produces very competitive performance using low-cost data modalities. The scanning cost and time of cardiac MRI are very high in comparison to CXR and ECG modalities. Faster and easier scanning is vital for critical patients in clinical settings. Thus, PAWP prediction using CXR and ECG will be beneficial for clinicians. Next, our model also produces better results than multimodal CNN model (row 9) which uses the same backbone as our model without multimodal pre-training, showing the potential of multimodal pre-training.

Model Interpretation: We used the integrated gradients method [24] to demonstrate the interpretability of our best performing CardioVAE$_{X,G}$ model with both CXR and ECG modalities. Fig. 2 depicts important regions for a normal subject and an abnormal subject, as identified by our model's decisions. For ECG, the model focuses on R peak (for normal PAWP) as shown in zoomed-in segments, whereas the model relies on R and S peaks for abnormal PAWP. This indicates that our model performs the prediction based on QRS complex region [13]. The distinct alterations in the QRS complex enable the identification of left ventric-

ular structural changes and conduction abnormalities, which are closely linked to variations in PAWP, reflecting the heart's response to altered cardiac hemodynamic states. In CXR images, the model focuses on cardiac regions, i.e. the left and right ventricles and arteries. By examining these regions in CXR, the model identifies their enlargement or structural changes, important indicators of cardiac function and fluid status. This offers important insights into PAWP levels by detecting subtle radiographic features of cardiac hemodynamic shifts and ventricular pressure alterations.

4 Conclusion and Future Work

This paper presented a multimodal variational autoencoder for CHDI detection from CXR and ECG. We showed that 1) the low-cost medical modalities (i.e., CXR and ECG) can be used to detect CHDI and are comparable to [3, 27] from high-cost modalities such as cardiac MRI, 2) the employed tri-stream unsupervised pre-training improved the performance of unimodal and multimodal models compared to [16, 18, 23, 30], and 3) interpretations made by our model are relevant for clinical decision-making as confirmed by a clinician. Future work can extend CardioVAE$_{\mathrm{X,G}}$ to other cardiac hemodynamics prediction tasks.

Disclosure of Interests: The authors have no competing interests to declare.

References

1. Alemi, A., Poole, B., Fischer, I., Dillon, J., Saurous, R.A., Murphy, K.: Fixing a broken ELBO. In: International Conference on Machine Learning. pp. 159–168. PMLR (2018)
2. Bu, Y., Zou, S., Liang, Y., Veeravalli, V.V.: Estimation of KL divergence: Optimal minimax rate. IEEE Transactions on Information Theory **64**(4), 2648–2674 (2018)
3. Garg, P., Gosling, R., Swoboda, P., Jones, R., Rothman, A., Wild, J.M., Kiely, D.G., Condliffe, R., Alabed, S., Swift, A.J.: Cardiac magnetic resonance identifies raised left ventricular filling pressure: prognostic implications. European Heart Journal **43**(26), 2511–2522 (2022)
4. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016)
5. Gow, B., Pollard, T., Nathanson, L.A., Johnson, A., Moody, B., Fernandes, C., Greenbaum, N., Berkowitz, S., Moukheiber, D., Eslami, P., et al.: Mimic-iv-ecg-diagnostic electrocardiogram matched subset. Type: dataset (2023)
6. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: Beta-vae: Learning basic visual concepts with a constrained variational framework. In: International Conference on Learning Representations (2016)
7. Hinton, G.E.: Training products of experts by minimizing contrastive divergence. Neural Computation **14**(8), 1771–1800 (2002)
8. Hirata, Y., Kusunose, K., Tsuji, T., Fujimori, K., Kotoku, J., Sata, M.: Deep learning for detection of elevated pulmonary artery wedge pressure using standard chest X-ray. Canadian Journal of Cardiology **37**(8), 1198–1206 (2021)

9. Hurdman, J., Condliffe, R., Elliot, C., Davies, C., Hill, C., Wild, J., Capener, D., Sephton, P., Hamilton, N., Armstrong, I., et al.: Aspire registry: assessing the spectrum of pulmonary hypertension identified at a referral centre. European Respiratory Journal **39**(4), 945–955 (2012)
10. Hwang, H., Kim, G.H., Hong, S., Kim, K.E.: Multi-view representation learning via total correlation objective. Advances in Neural Information Processing Systems **34**, 12194–12207 (2021)
11. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data **6**(1), 317 (2019)
12. Joy, T., Shi, Y., Torr, P.H., Rainforth, T., Schmon, S.M., Siddharth, N.: Learning multimodal vaes through mutual supervision. arXiv preprint arXiv:2106.12570 (2021)
13. Kashani, A., Barold, S.S.: Significance of qrs complex duration in patients with heart failure. Journal of the American College of Cardiology **46**(12), 2183–2192 (2005)
14. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
15. Kusunose, K., Hirata, Y., Tsuji, T., Kotoku, J., Sata, M.: Deep learning to predict elevated pulmonary artery pressure in patients with suspected pulmonary hypertension using standard chest X ray. Scientific Reports **10**(1), 19311 (2020)
16. Kusunose, K., Hirata, Y., Yamaguchi, N., Kosaka, Y., Tsuji, T., Kotoku, J., Sata, M.: Deep learning approach for analyzing chest X-rays to predict cardiac events in heart failure. Frontiers in Cardiovascular Medicine **10**, 1081628 (2023)
17. Lawry Aguila, A., Chapman, J., Altmann, A.: Multi-modal variational autoencoders for normative modelling across multiple imaging modalities. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 425–434. Springer (2023)
18. Li, L., Camps, J., Wang, Z., Banerjee, A., Rodriguez, B., Grau, V.: Towards enabling cardiac digital twins of myocardial infarction using deep computational models for inverse inference. arXiv preprint arXiv:2307.04421 (2023)
19. Makowski, D., Pham, T., Lau, Z.J., Brammer, J.C., Lespinasse, F., Pham, H., Schölzel, C., Chen, S.A.: Neurokit2: A python toolbox for neurophysiological signal processing. Behavior Research Methods pp. 1–8 (2021)
20. Obukhov, A., Krasnyanskiy, M.: Quality assessment method for GAN based on modified metrics inception score and fréchet inception distance. In: Proceedings of Computational Methods in Systems and Software, Vol. 14. pp. 102–114. Springer (2020)
21. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems **32** (2019)
22. Raghu, A., Schlesinger, D., Pomerantsev, E., Devireddy, S., Shah, P., Garasic, J., Guttag, J., Stultz, C.M.: ECG-guided non-invasive estimation of pulmonary congestion in patients with heart failure. Scientific Reports **13**(1), 3923 (2023)
23. Schlesinger, D.E., Diamant, N., Raghu, A., Reinertsen, E., Young, K., Batra, P., Pomerantsev, E., Stultz, C.M.: A deep learning model for inferring elevated pulmonary capillary wedge pressures from the 12-lead electrocardiogram. JACC: Advances **1**(1), 100003 (2022)

24. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International Conference on Machine Learning. pp. 3319–3328. PMLR (2017)
25. Suvon, M.N., Tripathi, P.C., Alabed, S., Swift, A.J., Lu, H.: Multimodal learning for predicting mortality in patients with pulmonary arterial hypertension. In: International Conference on Bioinformatics and Biomedicine. pp. 2704–2710. IEEE (2022)
26. Traversi, E., Cobelli, F., Pozzoli, M.: Doppler echocardiography reliably predicts pulmonary artery wedge pressure in patients with chronic heart failure even when atrial fibrillation is present. European Journal of Heart Failure **3**(2), 173–181 (2001)
27. Tripathi, P.C., Suvon, M.N., Schobs, L., Zhou, S., Alabed, S., Swift, A.J., Lu, H.: Tensor-based multimodal learning for prediction of pulmonary arterial wedge pressure from cardiac MRI. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 206–215. Springer (2023)
28. Vafaii, H., Mandino, F., Desrosiers-Grégoire, G., O'Connor, D., Markicevic, M., Shen, X., Ge, X., Herman, P., Hyder, F., Papademetris, X., et al.: Multimodal measures of spontaneous brain activity reveal both common and divergent patterns of cortical functional organization. Nature Communications **15**(1), 229 (2024)
29. Welch, B.L.: The generalization of 'student's'problem when several different population varlances are involved. Biometrika **34**(1-2), 28–35 (1947)
30. Wu, M., Goodman, N.: Multimodal generative models for scalable weakly-supervised learning. Advances in Neural Information Processing Systems **31** (2018)
31. Zhou, H.Y., Yu, Y., Wang, C., Zhang, S., Gao, Y., Pan, J., Shao, J., Lu, G., Zhang, K., Li, W.: A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. Nature Biomedical Engineering pp. 1–13 (2023)