Deep Metric Learning for the Hemodynamics Inference with Electrocardiogram Signals

Hyewon Jeong HYEWONJ@MIT.EDU
Collin M. Stultz CMSTULTZ@MIT.EDU
Marzyeh Ghassemi MGHASSEM@MIT.EDU
MIT CSAIL

Abstract

Heart failure is a debilitating condition that affects millions of people worldwide and has a significant impact on their quality of life and mortality rates. An objective assessment of cardiac pressures remains an important method for the diagnosis and treatment prognostication for patients with heart failure. Although cardiac catheterization is the gold standard for estimating central hemodynamic pressures, it is an invasive procedure that carries inherent risks, making it a potentially dangerous procedure for some patients. Approaches that leverage non-invasive signals – such as the electrocardiogram (ECG) – have the promise to make the routine estimation of cardiac pressures feasible in both inpatient and outpatient settings. Prior models trained to estimate intracardiac pressures (e.g., mean pulmonary capillary wedge pressure (mPCWP)) in a supervised fashion have shown good discriminatory ability but have been limited to the labeled dataset from the heart failure cohort. Furthermore, obtaining large datasets for diverse patient cohorts with intracardiac pressure labels is challenging due to the invasive nature of the procedure. To address these issues and build a robust representation, we apply traditional deep metric learning (DML) and propose a novel self-supervised DML with distance-based mining that improves the performance of a model with limited labels. We use a dataset that contains over 5.4 million ECGs without concomitant central pressure labels to pre-train a self-supervised DML model which showed improved classification of elevated mPCWP compared to self-supervised contrastive baselines. Additionally, the supervised DML model that uses ECGs with access to 8,172 mPCWP labels demonstrated significantly better performance on the mPCWP regression task compared to the supervised baseline. Moreover, our data suggest that DML yields models that are performant across patient subgroups, even when some patient subgroups are under-represented in the dataset.

1. Introduction

Heart failure is a chronic condition that occurs when the heart is unable to pump blood effectively, which affects people worldwide with significant morbidity and mortality rates (Roger, 2013). The diagnosis and management of heart failure are challenging due to the heterogeneity in the phenotype of patients (Heidenreich et al., 2022; Roger, 2013; Burkhoff et al., 2003; Komajda and Lam, 2014). Accurately measuring intracardiac pressure is crucial for diagnosing and following up on treatment response in heart failure patients, as hemodynamic assessment can predict the risk of various negative outcomes (Baudry et al., 2022). Intracardiac pressures play an important role in the assessment of hemodynamic severity in patients with heart failure (Heidenreich et al., 2022). The mean pulmonary

capillary wedge pressure (mPCWP) - an estimate of the left atrial pressure - is one intracardiac measurement that is used for treatment and prognostication in heart failure patients (Drazner et al., 2008; Heidenreich et al., 2022). Right heart catheterization is a procedure that is used to measure intracardiac and pulmonary pressures (e.g., mPCWP). However, it is an invasive procedure, where a pressure transducer is inserted into a great vessel and threaded into the right heart chambers (Drazner et al., 2008; Heidenreich et al., 2022). To provide readily available diagnostic tools to detect and understand heart failure, we need safe routine screening methods that reliably predict patient hemodynamics.

One widely available clinical datum is the electrocardiogram (ECG), which has been applied for detecting arrhythmia (Hannun et al., 2019), myocardial infarction (Acharya et al., 2017), and heart failure (Masetic and Subasi, 2016; Acharya et al., 2019). Prior works have targeted supervised models for non-invasive estimation of hemodynamics using ECGs (Schlesinger et al., 2022; Raghu et al., 2023b). However, the population of catheterized patients with labeled data is limited, and supervised models do not leverage the rich information contained in large, unlabeled ECG datasets. Self-supervised models pre-trained on unlabeled ECGs such as PCLR (Diamant et al., 2022), CLOCS (Kiyasseh et al., 2021), and SimCLR (Chen et al., 2020) have performed well in predicting cardiac abnormalities. However, our experiments found that contrastive learning methods perform worse than supervised models in regression tasks (Table 1). Further, we observed that supervised and contrastive baselines fail to achieve fair prediction across gender groups (Table 2).

In this work, we seek to improve models by leveraging Deep Metric Learning (DML) with mining methods using either label or similarity distance metrics. DML is similarity-based learning where the objective is formulated to employ formal distance metrics between encoded representations to capture semantic similarities between data. We apply DML to a real-world clinical task, creating a supervised DML baseline where positive instances are sampled by label. We further propose a novel self-supervised DML model for time series and signal datasets where positive instances are sampled by a self-supervised distance-based ranking (Figure 1). We also present a continuous label-based mining method for supervised DML. We focus on intracardiac pressure classification and regression using 12-lead ECGs, with 8,172 ECGs with matched mPCWP (*ECGs with Labels*) and 5,426,614 unlabeled ECGs (*ECGs without Labels*) from the data warehouse of Massachusetts General Hospital and Brigham and Women's Hospital. We compare the DML models to state-of-the-art supervised and self-supervised models (Chen et al., 2020; Kiyasseh et al., 2021; Diamant et al., 2022), and evaluate their ability to predict intracardiac pressure values and classes (e.g., elevated mPCWP).

We found that supervised DML showed significant improvements in the regression of hemodynamics while obtaining competitive performance in classification, compared to the supervised and contrastive learning baselines (Table 1). Self-supervised DML showed significant improvements in the discriminability of the model with the regression performance compared to random initialization and the performance was on par with baselines, which shows promise that the utilization of unlabeled datasets helps achieve more robustness (Table 1). Additionally, we found that both supervised baseline and contrastive learning showed high subgroup performance gaps, while supervised DML and self-supervised DML significantly reduced the subgroup gaps (Table 2, 8, Figure 2).

Generalizable Insights about Machine Learning in the Context of Healthcare

Invasive procedures such as right cardiac catheterization result in limited data availability. Contrastive learning has been used to address the problem of restricted labels, but in our dataset, it resulted in poor regression performance and inequitable prediction across gender groups (Table 1). To improve inference in the presence of a limited label, we propose using both supervised and self-supervised DML in intracardiac pressure inference tasks (classification and regression). The supervised DML and self-supervised DML models outperformed the supervised and contrastive learning baselines in elevated mPCWP classification and regression tasks and achieved fairer predictions across different gender groups. Overall, our findings suggest that DML approaches can be utilized to create more robust representations for hemodynamic inference, which has the potential to help overcome the barrier of limited data acquisition in healthcare.

Furthermore, we focus both on identifying abnormal mPCWP as a binary classification task and predicting the exact mPCWP values as a regression task, using either binary or continuous label-based mining as well as the DML training scheme. This is important, as medical datasets frequently comprise a mix of categorical and continuous data, necessitating both classification and regression tasks for model-based clinical status inference. For example, intracardiac pressure obtained through right heart catheterization is one typical technique in which clinicians seek to see if a patient has abnormal intracardiac pressure (e.g., increased mPCWP) or if the value changes before or after treatment.

2. Related Work

2.1. Clinical Prediction with ECGs

There is a large body of work developing a deep learning-based ECG classifier for inferring cardiac abnormalities (Hannun et al., 2019; Acharya et al., 2017; Masetic and Subasi, 2016; Acharya et al., 2019). Recent research has extended beyond ECG classification tasks to regression tasks that infer hemodynamics, such as cardiac output or intracardiac pressure including pulmonary artery pressure and mPCWP (Schlesinger et al., 2022; Raghu et al., 2023b). Additionally, recent research has shown that ECG signals also encode information about patient demographics by demonstrating age and sex prediction with ECGs (Attia et al., 2019). This shows that leveraging the encoded physiological and demographic information in ECGs could potentially improve model robustness and accuracy.

Contrastive Learning with ECGs and Biomedical Signals Contrastive learning performs pre-training with pretext tasks to generate a global embedding over an unlabeled dataset, by encouraging the invariance to perturbations in the representations (Chen et al., 2020; He et al., 2020; Grill et al., 2020). There have been several attempts to leverage self-supervised representation learning with unlabeled biomedical signal datasets (Zhang et al., 2022; Cheng et al., 2020) including ECGs (Diamant et al., 2022; Kiyasseh et al., 2021; Gopal et al., 2021; Lan et al., 2022; Mehari and Strodthoff, 2022; Oh et al., 2022; Raghu et al., 2023a) and multimodal time-series data (Raghu et al., 2023a). However, to the best of our knowledge, self-supervised DML has not been applied to the downstream tasks involving ECGs. In this work, we propose a novel method to use DML on ECG datasets without labels (Self-supervised DML) along with conventional DML methods with labels

(Supervised DML) for downstream hemodynamics classification and regression tasks and compare the test performance with supervised learning and contrastive learning baselines.

2.2. Deep Metric Learning

Deep Metric Learning (DML) is a similarity-driven learning method for building a representation by learning a distance metric. DML has been demonstrated to improve generalization in representation learning (Roth et al., 2021; Milbich et al., 2021; Dullerud et al., 2021), including some domain-specific applications to ECG datasets (Yu et al., 2021; Zhu et al., 2022). DML works by optimizing the loss function defined on pairs or tuples generated from training examples, where the objective is to preserve the relative distances between pairs of data points in the training data. Triplet loss (Wu et al., 2017; Yu et al., 2018) is the representative DML objective function that pulls similar samples together and pushes the dissimilar samples away. Extending this objective, angular loss (Wang et al., 2017) introduces geometric restrictions to achieve a higher convergence rate with scale invariance. Margin loss (Wu et al., 2017) utilizes the learnable boundary across samples with the same and different labels, which extends the triplet loss.

While contrastive learning defines positive samples to be the perturbed views and negative samples to be the views from other samples, traditional DML defines positive and negative samples based on labels (positive samples are the ones with the same label as an anchor). Negative samples can be mined randomly (random mining (Hu et al., 2014)) from the pool of samples with different labels to the anchor, and can also be mined using several hard negative mining techniques. One such technique is semihard mining (Schroff et al., 2015), where the hard negatives are selected from the samples that exist within a certain distance margin from the positive sample. Another technique is softhard mining (Roth et al., 2020), where the hard negatives are selected between the maximum margin from positive samples and the minimum distance between the negative and anchor.

As DML generally requires the class of labels to form the tuplets of data, only a few studies have applied DML to regression problems (Huang et al., 2018). Furthermore, only a few recent works proposed ways for establishing DML without using labels to sample triplets (Dutta et al., 2020b; Iscen et al., 2018; Dutta et al., 2020a). Here we apply self-supervised DML to a real-world regression problem to determine if similarity information within ECG dataset as well as labels helps create an embedding space that improves the performance of downstream hemodynamics inference. Although there are several self-supervised DML works (Paixao et al., 2020; Fu et al., 2021b,a), to the best of our knowledge, our study is the first to use distance-based ranking to sample the triplets among the pool of unlabeled pertaining dataset.

3. Deep Metric Learning for Hemodynamics Inference

In this work, we evaluate whether DML methods can be used to build a robust representation of cardiac waveforms by leveraging labeled and unlabeled ECG data. We develop and assess two training strategies: supervised DML for the joint learning of hemodynamics classification and regression, and self-supervised DML pre-training for the downstream hemodynamics inference (classification and regression).

3.1. Supervised Deep Metric Learning for Hemodynamics Classification and Regression

Problem Setting: Given a sample size N labeled ECG dataset with N = 8,172 data points: $\mathcal{D}_{sup} = \{(x_i, y_i)|(x_1, y_1), \cdots, (x_N, y_N)\}$. Here, $x_i \in \mathbf{R}^{d \times \mathcal{T}}$ and $y_i \in \{0, 1\}$ with d being the number of ECG leads and $\mathcal{T}$ is the number of ECG timesteps. The binary label y describes whether a patient has elevated mPCWP or not, which is defined by the cutoff mPCWP threshold 18 mmHg (1 if the patient has elevated mPCWP ($>$ 18 mmHg)).

Binary Label Based Mining: For each anchor ECG x_a, we define the positive sample x_p to be the ECG having the same label as the anchor ECG and the negative sample x_n is defined to be the one that has a different label. The triplet (x_a, x_p, x_n) are then sampled from the batch $\mathcal{B}$ to calculate the triplet loss $\mathcal{L}_{triplet}$. Negative samples can be mined randomly from the batch of samples with different labels from anchors (*Random Mining*). We further applied hard negative mining methods (*Semihard Mining* (Schroff et al., 2015) and *Sofhard Mining* (Roth et al., 2020)) which select harder negatives around positive samples to build a robust DML representation. We introduce the formulations of hard negative mining in Appendix C.2.

Continuous Label Based Mining (*Label Based Mining*)**:** Given an anchor ECG x_a, we mine the triplet (x_a, x_p, x_n) from a batch based on the continuous label mPCWP value, where we designate the positive sample x_p as the ECG with an mPCWP value that is closest to the anchor label. Conversely, we define the negative sample, x_n, as the ECG with an mPCWP value that is furthest from the anchor label. The detailed method is in Appendix C.1.

Supervised DML Classification: The model f_Φ is an encoder has D-dimensional output, optimized with triplet loss ($\mathcal{L}_{Triplet} = \sum_{(x_a, x_p, x_n) \in \mathcal{B}} \{||f_\Phi(x_a) - f_\Phi(x_p)|| - ||f_\Phi(x_a) - f_\Phi(x_n)||$). The last fully connected layer g gets the input from the model embedding f_Φ that has not optimized with triplet loss, and $g \circ f_\Phi$ is optimized by the cross-entropy loss ($\mathcal{L}_{CE}$). The final joint training objective combines the two loss functions, for a batch size of $|\mathcal{B}|$ with the triplet loss scaling factor of α: $\mathcal{L}_{tot} = \mathcal{L}_{CE} + \alpha \mathcal{L}_{triplet}$ (Figure 1 (a), See Appendix A.1 for full loss function). We further experiment with two different objective functions (angular loss and margin loss) in replacement of triplet loss, where we introduce the learnable boundary between positive and negative pairs in margin loss and geometric constraints for angular loss. See Appendix B for detailed explanations of different loss functions.

Supervised DML Regression: Using the mined triplets (x_a, x_p, x_n), the model f_Φ and the final classifier output $g \circ f_\Phi$ are jointly trained in an end-to-end fashion with the objective adding triplet loss and Root Mean Square Error (RMSE) with the triplet scaling factor of α: $\mathcal{L}_{tot} = \mathcal{L}_{RMSE} + \alpha \mathcal{L}_{triplet}$. Unlike the classification task above, the regressor output predicts the mPCWP value itself. See Appendix Section A.2 for the full loss function. For both classification and regression joint training, we performed an exhaustive search to determine the optimal value of the triplet loss scaling factor α to enhance the performance of our model (Table 4 in Appendix Section A.3).

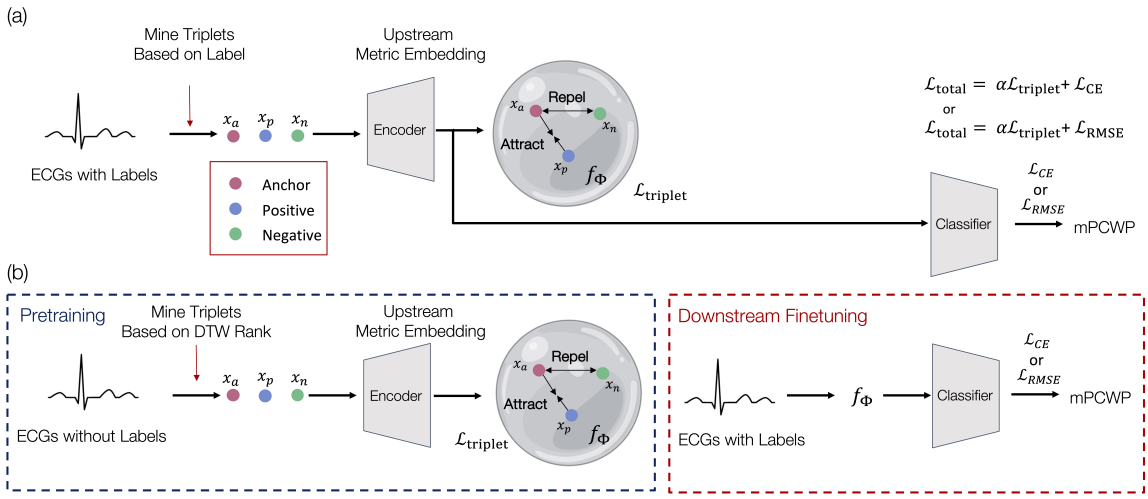

Figure 1: Intracardiac pressure inference using ECGs with DML. **(a) Supervised DML for Joint-Learning of Hemodynamics Classification and Regression.** ECG triplets are mined based on mPCWP labels (binary label indicating elevation of mPCWP $>$ 18 mmHg), which are used to build the upstream metric embedding f_Φ. Triplet loss from ECG triplets and classification/regression loss are jointly learned in an end-to-end fashion. **(b) Self-Supervised DML.** Using the unlabeled ECG corpus (ECGs without matched mPCWP), we learn upstream metric embedding f_Φ with deep metric learning objective (*Pretraining*). Then we finetune this embedding f_Φ with ECGs with labels to infer mPCWP (*Downstream Finetuning*).

3.2. Self-Supervised Deep Metric Learning

Problem Setting: Now we extend our approach to self-supervised DML where we have larger body of $M = 5.4$ million ECG data points without the matched mPCWP labels y: $\mathcal{D}_{self-sup} = \{x_j | x_1, \cdots, x_M\}$, where $x_j \in \mathbf{R}^{d \times \mathcal{T}}$.

Distance Based Mining: Given an anchor x_a in a batch, a positive sample corresponding to each anchor is determined as the one with the smallest distance metric which compares the ECG signal inputs. We sample triplets using a distance-based ranking, where the measured distance between two inputs represents the similarity between inputs. From the calculated measures of dynamic time warping (DTW (Müller, 2007), d_{DTW}) and Euclidean distance between the pairs of ECG samples in a batch, we sample the triplets based on the distance-based ranking. For instance, given an anchor x_a, positive sample x_p is the one with the minimum distance from x_a within the batch: $\mathcal{B}_{pos} = \{x_p | \arg\min_{x_p \in \mathcal{B}} d_{DTW}(x_p, x_a)\}$ for the case where the distance measure is DTW. Then we select the negative data points x_n randomly from the set that excludes the positive sample: $\mathcal{B}_{neg} = \{x_n | x_n \in \mathcal{B}, x_n \neq x_p\}$. Note that DTW has calculated for every pair of input batches as it is not a symmetric true distance and thus does not obey the triangle inequality.

Self-Supervised DML Pretraining and Downstream Hemodynamics Inference: The model f_Φ is optimized with the upstream DML with $\mathcal{L}_{triplet}$. (Figure 1 (a) Pretraining) The input triplets (x_a, x_p, x_n) are mapped on D-dimensional representation space f_Φ, where we finetune the pre-trained self-supervised DML model with the downstream hemodynamic inference task (See Appendix A.4, Figure 1 (b) Downstream Finetuning). The binary classification task is to infer the elevation of mPCWP with the threshold of 18 mmHg, and the regression task would be to predict the value of mPCWP.

4. Experimental Design

4.1. Experiments

To evaluate whether DML methods provide robust prediction on hemodynamics inference tasks (classification, regression), we first compare the performance of the DML models (supervised DML, self-supervised DML) against random initialization, supervised learning, and contrastive learning baselines (Table 1). To evaluate the quality of the representation, we evaluated the learned representation's ability to capture the samples with the same label within the 1-nearest neighbors using Recall@1 (Table 1). For supervised DML joint-trained models, we explored the mining methods (random, label-based, semihard, and softhard) and DML objectives (triplet, margin, and angular loss) (Table 6 and Table 5). Finally, we analyzed the performance gaps across different age and gender subgroups to measure subgroup bias in model performance (Table 2, 8, and Figure 2).[1]

4.2. Models and Implementation Details

4.2.1. DML Models

Supervised DML We used ResNet18 (He et al., 2016) backbone as an encoder f_Φ for generating D-dimensional space (with D of 128, 256, 512, 1024) to optimize the objective function, and f_Φ is fed into the classifier g with two fully-connected (FC) layers with batch normalization, ReLU activation, and a dropout rate of 0.3. We add a sigmoid activation function to the last fully connected layer for the classification task.

Self-Supervised DML We experimented with two similarity or distance measures (DTW, Euclidean distance) for the similarity-based ranking of triplet mining in self-supervised DML. We used the same ResNet18 (He et al., 2016) backbone as an encoder f_Φ to generate the embedding dimension D of 128, 256, 512, and 1024. Classifier (g) with the same architecture as SupDML (2 FC layers) is added after encoder f_Φ for finetuning.

4.2.2. Baselines

The baselines we used to compare against supervised DML or self-supervised DML are as below.

1. Random Initialization We randomly initialized f_Φ and fine-tuned the model on the downstream task. This serves as a standard baseline for self-supervised methods.

1. The code used for the experiment is available at https://github.com/mandiehyewon/ssldml.

2. Supervised Models We train *Supervised classification* and *Supervised Regression* models with ResNet18 (He et al., 2016) architecture that are trained for hemodynamics binary classification (elevated mPCWP) or regression.

3. Contrastive Learning Models We evaluate self-supervised contrastive learning models that are pre-trained with unlabeled ECGs (*ECGs without Labels*) used for finetuning on downstream tasks. Finetuning of the contrastive learning models followed the same step as the downstream hemodynamics inference steps in section 3.2.
(1) *SimCLR* (Chen et al., 2020) is a self-supervised learning approach that learns useful representations of images by maximizing the agreement between augmented views of the same image. We used the same encoder f_Φ as DML models (ResNet18), and the same classifier g except the projection head for calculating the contrastive loss. To reproduce the experiments for SimCLR in Kiyasseh et al. (2021), we used three perturbations: Gaussian noise (tested over noise variance hyperparameter 0.01, 1, 10, 100), flipping along the temporal axis, and flipping along the amplitude axis of ECGs.
(2) *CLOCS* (Kiyasseh et al., 2021) is a family of contrastive learning that exploits spatial and temporal invariance of ECG signals. We used Contrastive Multi-segment Coding (CMSC) for our CLOCS baseline, which was the best-performing model in the original paper.
(3) *PCLR* (Diamant et al., 2022) is patient-contrastive learning where the ECGs from the same patients were paired for self-supervised pre-training. We used a pre-trained PCLR model provided by the authors for finetuning, which is 320 dimensions.

4.2.3. Evaluation

For evaluating the performance of the mPCWP classification task, we utilized the Area Under the Receiver Operating Characteristic Curve (AUC) and the Area Under the Precision-Recall Curve (APR). Additionally, we adopted Root Mean Squared Error (RMSE) as the performance metric for the mPCWP regression task. To ensure the reliability of our results, we reported the mean and standard deviation of 1,000 bootstraps for both tasks. For statistical analysis of performance and the performance gap between models, we utilized the Kruskal-Wallis tests.

To evaluate the quality of the embedding space, we employed the Recall@k (Jegou et al., 2011) performance metric, which quantifies the number of instances with identical labels within the k-nearest neighbor of the sample. Specifically, we set k=1 to assess the proportion of instances with matching labels within the 1-nearest neighbor and report Recall@1 (Table 1).

To investigate the difference in task inference quality between demographic groups, we tested pre-trained or trained models listed in Section 4.2.2 on age and gender subgroups of the test dataset. For the performance gap across gender subgroups, we presented the relative difference in performance (AUC, APR, RMSE) calculated by subtracting the performance of the female group from the performance of the male group (Table 2). For analyzing the performance gap across age subgroups, we calculated the absolute value of the averaged pairwise relative difference in performance between four distinct age subgroups: $18 \leq$ age < 35, $35 \leq$ age < 50, $50 \leq$ age < 75, $75 \leq$ age.

The proportion of k-nearest neighbors of the embedding space belonging to the same minority gender group (female) is then calculated to measure the degree of minority sub-

group segregation in the representation space, with $k = 2, 3, 5$ (Table 3). We performed the k-nearest neighbor computation with scikit-learn (Pedregosa et al., 2011) and used seaborn (Waskom, 2021) for the bar plot.

4.3. Datasets

We trained and evaluated the models with two sets of datasets with distinct patient cohorts: *ECGs with Labels* and *ECGs without Labels*. Both datasets include primarily 10 total seconds of trace per patient, where these ECGs were acquired under optimal conditions while the patients were at rest. We utilized the dataset *ECGs with Labels* for supervised DML joint-training (Figure Figure 1 (a)), supervised models, and for the finetuning of pre-trained models (self-supervised DML, contrastive learning, Figure 1 (b) Downstream Finetuning). We then used *ECGs without Labels* for self-supervised DML and self-supervised contrastive learning pre-training (Figure 1 (b) Pretraining). Both *labeled* and *unlabeled* ECG datasets below were obtained as part of an IRB-approved protocol.

Dataset with mPCWP Labels (*ECGs with Labels*) We construct the dataset containing 12 lead ECGs with corresponding right heart catheterization procedures from the data warehouse of Massachusetts General Hospital. Our dataset consists of 8,172 ECGs and right heart catheterization procedure data from 4,051 patients who underwent ECG recordings on the same date they had right heart catheterization from 2010 to 2020. The ECG data contains 10 seconds of signal data, sampled at a rate of 250 Hz, and the label mPCWP was obtained from the right heart catheterization procedure. We partitioned the dataset into the train, valid, and testing sets, which include 4,680, 1,667, and 1,825 data points, as well as 2,430, 810, and 811 patients, respectively. Our training, validation, and testing sets were partitioned ensuring that no patient overlapped across these different sets. Among 811 test set patients, 344 were female and 467 were male, showing a higher representation of male patients compared to females. We have 20 patients in the age group 18-35, 77 patients in the 35-50 group, 428 in 50-75, and 297 in patients aged over 75. Because age and gender is the primary demographic group available in our dataset, we assessed the fairness of models in different age and gender groupings.

Dataset without mPCWP Labels (*ECGs without Labels*) The larger 12 lead ECG dataset without mPCWP mapping has been provided by the agreement from Massachusetts General Hospital and Brigham and Women's Hospital where we have a total of $5,426,614$ ECGs with $1,195,268$ patients who had ECG recordings while their daytime visit or admission. This dataset includes a more general population of patients and was not only limited to heart failure patients. Of all the ECGs of patients, 917,395 patients have ECG signals exceeding 10 seconds in length. In these particular instances, we selected and isolated 10-second segments from the total ECG signal to generate multiple distinct ECG traces. We implemented this method to ensure a standardized length across all ECG samples in our dataset. We removed patient cohorts overlapping with the *ECG datasets with mPCWP labels* and excluded abnormal ECG signals with continued zero amplitude (mV) at any point of recording at any lead. The other preprocessing and preparation of the ECG dataset follow the same procedure used for the *ECGs with Labels*.

Table 1: Performance of downstream mPCWP classification task with supervised baseline, supervised DML, and unsupervised DML pretraining according to each embedding dimension (Dim). The supervised baseline model utilized an encoder with 128 dimensions. We report the mean with a standard deviation based on 1,000 bootstraps, and models with the best performance are boldfaced.

			Classification		Regression	Embedding
	Models	Dim	AUC	APR	RMSE	Recall@1
Random Initialization		128	74.3 ± 1.9	55.7 ± 3.3	7.88 ± 0.26	68.4
Supervised	Supervised	128	78.4 ± 1.1	59.5 ± 2.2	7.45 ± 0.15	-
	Supervised DML (Random Mining)	128	76.7 ± 1.6	56.1 ± 3.3	**7.15** ± 0.21	**69.6**
		256	76.5 ± 1.7	57.2 ± 3.2	7.19 ± 0.20	66.5
		512	77.1 ± 1.8	59.6 ± 3.2	7.21 ± 0.18	67.5
		1024	77.5 ± 1.7	58.1 ± 2.9	7.20 ± 0.20	66.3
Self-Supervised	SimCLR	128	75.0 ± 1.9	58.2 ± 3.2	7.58 ± 0.22	68.6
	CLOCS	128	74.9 ± 1.7	54.0 ± 3.2	7.61 ± 0.22	68.3
	PCLR	320	75.3 ± 1.5	56.0 ± 1.5	10.04 ± 0.27	63.2
	Self-supervised DML (DTW)	128	77.2 ± 1.7	**60.5** ± 3.1	7.65 ± 0.21	**69.3**
		256	76.4 ± 1.8	57.6 ± 3.1	7.69 ± 0.17	**69.4**
		512	75.3 ± 1.8	55.2 ± 3.3	7.71 ± 0.26	67.2
		1024	74.8 ± 1.7	55.4 ± 3.1	7.70 ± 0.21	63.8
	Self-supervised DML (Euclidean)	128	77.4 ± 1.8	58.2 ± 3.5	7.80 ± 0.22	64.2
		256	77.2 ± 1.7	**60.7** ± 3.2	7.59 ± 0.20	61.4
		512	**79.4 ± 1.1**	59.9 ± 2.2	7.64 ± 0.17	67.2
		1024	76.3 ± 1.6	58.5 ± 3.3	7.70 ± 0.24	67.7

5. Results

5.1. Supervised and Self-supervised DML on Hemodynamics Inference

We summarize the test performance of DML models and baselines on downstream hemodynamics inference task in Table 1. We find that training with DML showed significantly improved performance compared to the baselines. Supervised DML (best RMSE of 7.15 with embedding dimension 128) significantly outperformed random initialization (7.88), supervised (7.38), and contrastive baselines (lowest 7.58 with SimCLR) on the regression task. The regression performance of supervised DML on every embedding dimension (128, 256, 512, 1024) achieved significantly low ($p < 0.0001$) RMSE loss compared to any other baselines. Supervised DML exhibited comparable performance on classification using the APR metric (59.6 with embedding dimension 1024), where its performance was not significantly different from that of the supervised baseline (59.5), while the AUC was significantly higher for the supervised baseline (78.4 for supervised and 77.5 for supervised DML with embedding dimension 1024). This implies that the underlying similarity information in the ECG signal pattern captured by the labels helps infer the downstream wedge pressure value. Supervised DML results were reported with the best triplet loss scaling factor α after

hyperparameter tuning on various α values (0.1, 1.0, 2.0, 3.0, 10.0): α of 3.0 and 2.0 were used for classification and regression, respectively, for the embedding dimension of 128. For detailed results on other embedding dimensions (256, 512, 1024) please refer to Table 4 in Appendix Section A.

We evaluated the classification performance of self-supervised DML and compared it to that of contrastive learning baselines. Our findings show that both DTW-based and Euclidean-based ranking in self-supervised DML achieved significantly better classification performance (Best AUC of 79.4 in supervised DML with Euclidean distance, APR of 60.5 for self-supervised DML with DTW, and 60.7 for Euclidean) compared to the contrastive learning baselines (best performance AUC 75.3 with PCLR and APR of 58.2 with SimCLR) and random initialization (AUC 74.3 and APR 55.7) ($p < 0.0001$ against all baselines). On the other hand, we found no significant difference in regression performance between the contrastive learning baselines (lowest RMSE loss with SimCLR, 7.58) and self-supervised methods (lowest with Euclidean-based ranking, 7.59), while observing a significant improvement over PCLR (10.04) and random initialization (7.88) ($p < 0.0001$). Overall, self-supervised DML demonstrated the optimal trade-off between classification and regression tasks, achieving the best performance in the classification task while maintaining a comparable level of performance in the regression task.

5.2. Evaluation of the Representation Quality

To evaluate the discriminative ability of representation space, we compare the label consistency of close samples using the Recall@1 metric (Table 1). The best performance was observed in supervised DML with random mining and softhard mining (Table 6), as well as self-supervised DML with DTW-based ranking. On the other hand, supervised DML with label based mining, semihard mining (Table 6), and self-supervised DML with the Euclidean-based ranking (Table 1) showed worse performance compared to the contrastive learning (SimLCR and CLOCS). Furthermore, increasing the embedding dimension improved joint-learned task performance in supervised DML (Table 1), which is consistent with prior DML studies (Roth et al., 2020). However, the increase in embedding dimension did not affect the downstream task performance in supervised DML with hard negative mining (Table 6 in Appendix Section C) and self-supervised DML methods (Table 1).

5.3. DML Models Reduce Subgroup Performance Gaps

We investigated the differences in mPCWP inference among demographic groups (Table 2, 8, and Figure 2). For the gender groups, we reported the absolute performance difference between male and female groups of 1,000 bootstraps (Table 2, rows 'Gap'), and for the age groups, we calculated the average pairwise difference in performance between any two groups. These performance gaps are relative differences in performance.

Supervised and self-supervised DML helped achieve a significantly fairer prediction of downstream tasks across different demographic subgroups against the baselines. Among supervised models (supervised baseline, supervised DML), supervised DML showed low gaps for all the performance metrics (AUC, APR, and RMSE) compared to the supervised baseline and random initialization, achieving fair prediction across different gender subgroups. Supervised DML also achieved lower AUC and RMSE gaps compared to random

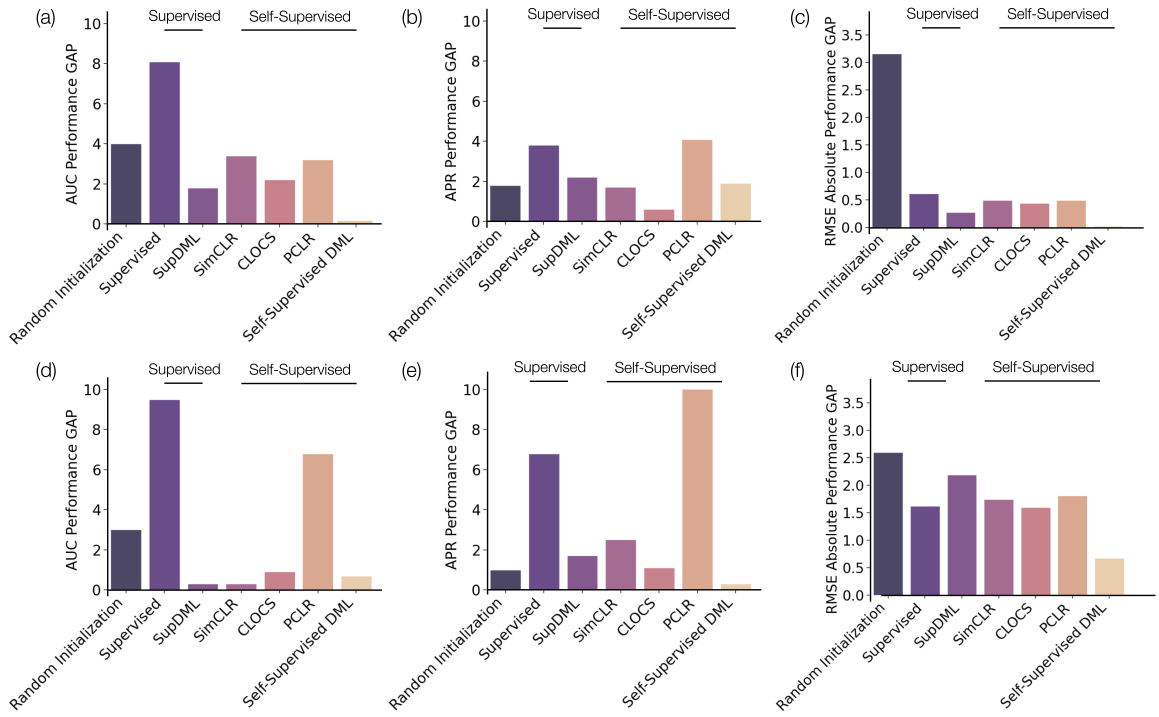

Figure 2: The absolute performance gap of mPCWP inference tasks (classification ((a), (d) AUC, (b), (e) APR) and regression ((c), (f) RMSE)) across demographic subgroups ((a)-(c) gender subgroup and (d)-(f) age subgroup).

initialization and self-supervised contrastive learning baselines (SimCLR, CLOCS, PCLR). Self-supervised learning showed the lowest performance gap on both AUC and RMSE with statistical significance with any other baselines ($p < 0.0001$), and SimCLR and CLOCS achieved a lower performance gap on APR compared to supervised and self-supervised DML. A comprehensive discussion of the performance gaps across age subgroups can be found in Appendix Section D.2.

The improvements in fair classification and regression imply that the pre-trained embedding learned with DML objectives, either using the triplets sampled based on label (supervised DML) or distance-based similarity metric (self-supervised DML), represent the physiological difference related to demographic groups, supporting the previous work that ECG signal can provide information to infer demographic groups (Attia et al., 2019).

5.4. Impact of varying DML Objectives

In Sections 5.4 and 5.5, we explore the effect of varying DML objectives and mining techniques, and the results show that our supervised DML model shows robust performance and achieves equitable prediction.

We investigated the impact of different DML objectives on the classification and regression tasks of mPCWP inference (Figure 3 (a)-(c)). We utilized margin loss and angular loss,

Table 2: Gender subgroup performance gap of downstream mPCWP classification task has decreased on either Supervised or Self-Supervised DML compared to supervised and contrastive baselines (an embedding dimension of 128 used for DML and contrastive learning encoder). We report the mean with a standard deviation based on 1,000 bootstraps, and models with the lowest performance gaps are boldfaced.

Models		Subgroup	Classification		Regression
			AUC	APR	RMSE
Random Initialization		Full	74.3 ± 1.9	55.7 ± 3.3	7.88 ± 0.26
		Male	77.3 ± 1.5	54.3 ± 3.1	8.68 ± 0.80
		Female	73.4 ± 1.4	56.1 ± 2.3	11.84 ± 2.52
		Gap	4.0	1.8	3.16
Supervised	Supervised Baseline	Full	78.4 ± 1.1	59.5 ± 2.2	7.45 ± 0.15
		Male	81.3 ± 1.4	61.6 ± 2.9	7.17 ± 0.18
		Female	73.2 ± 2.0	57.8 ± 3.5	7.88 ± 0.29
		Gap	8.1	3.8	0.71
	Supervised DML (Random Mining)	Full	76.7 ± 1.6	56.1 ± 3.3	7.15 ± 0.21
		Male	77.4 ± 1.5	54.2 ± 2.8	7.07 ± 0.19
		Female	75.6 ± 1.2	56.4 ± 2.4	7.34 ± 0.17
		Gap	**1.8**	**2.2**	**0.27**
Self-Supervised	SimCLR	Full	75.0 ± 1.9	58.2 ± 3.2	7.58 ± 0.22
		Male	76.2 ± 1.7	54.8 ± 3.2	7.18 ± 0.17
		Female	72.8 ± 1.4	56.5 ± 2.4	7.67 ± 0.16
		Gap	3.4	1.7	0.49
	CLOCS	Full	75.9 ± 1.7	56.2 ± 3.4	7.45 ± 0.22
		Male	76.5 ± 1.6	55.5 ± 2.9	7.22 ± 0.17
		Female	74.3 ± 1.4	56.1 ± 2.3	7.66 ± 0.17
		Gap	2.2	**0.6**	0.44
	PCLR	Full	75.3 ± 1.5	56.0 ± 2.4	10.04 ± 0.27
		Male	75.1 ± 1.5	53.9 ± 2.3	11.14 ± 0.24
		Female	71.9 ± 1.7	58.0 ± 2.7	11.63 ± 0.34
		Gap	3.2	4.1	0.49
	Self-supervised DML (DTW)	Full	77.2 ± 1.7	60.5 ± 3.1	7.65 ± 0.21
		Male	75.0 ± 1.0	57.1 ± 3.4	7.82 ± 0.20
		Female	75.0 ± 1.3	59.0 ± 2.3	7.84 ± 0.17
		Gap	**0.0**	1.9	**0.02**

along with the triplet loss, and trained them jointly with the loss for mPCWP inference (as described in Section 3.1). A detailed description of the objectives can be found in Appendix Section B. Our results showed that the angular loss performed nonsignificantly similar to the triplet loss on the classification task, significantly outperforming the performance with margin loss (Table 5). However, the triplet loss outperformed the angular loss in terms of RMSE on the regression task, for the supervised DML method.

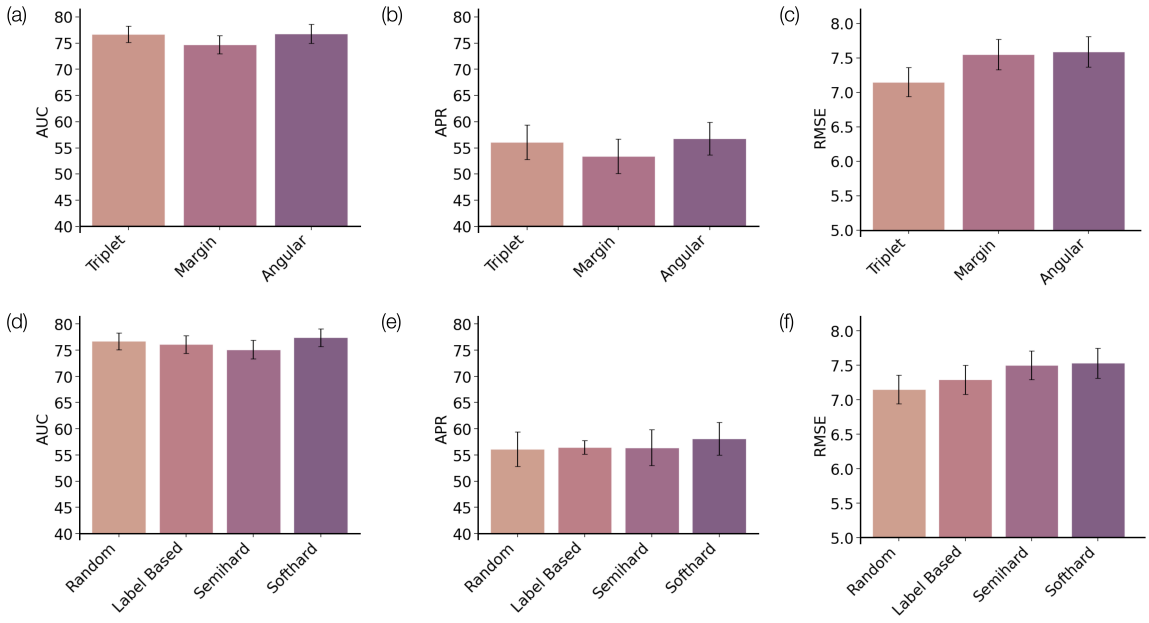

Figure 3: Evaluation of performance on various DML objectives ((a)-(c)) and mining strategies ((d)-(f)). AUC (a), APR (b), and RMSE of triplet, margin, and angular loss (c). AUC (d), APR (e), and RMSE of random, semihard, and softhard mining.

5.5. Impact of varying Supervised DML Mining

We evaluated the performance of label based mining and different hard negative mining methods (semihard and softhard mining) against the random mining method in supervised DML for mPCWP elevation classification and regression tasks across all embedding dimensions. The results are summarized in Figure 3 and Table 6 of Appendix C. We found that label based mining with 1024 dimensions showed AUC performance on par with random mining. Furthermore, softhard mining with 128 dimensions showed comparable classification performance to the result with random mining, while semihard mining showed regression performance similar to random mining. Interestingly, the trend of classification performance improvement with increased embedding dimension, observed in random mining, was not clearly visible in the explored label based mining and hard negative mining methods (semihard and softhard). This result indicates that the impact of embedding dimension on performance is not consistent when using label based or hard negative mining methods.

Subgroup Performance with Label Based Mining and Hard Negative Mining We report the subgroup performance gap of Supervised DML joint trained with label based and hard negative mining methods in Figure 4 (d)-(f) and Table 7 of Appendix C. Our results indicate that although softhard mining showed a higher performance gap compared to random mining in the classification task, the performance gap of APR and RMSE has decreased in Semihard mining. Also, label based mining showed the lowest gap in AUC and

showed gaps in other performance metrics comparable to random mining. This implies that Supervised DML embedding built with triplets with harder negatives helps achieve fairer prediction, while not harming the performance too much. The detailed performance report is on Table 7 of Appendix C.

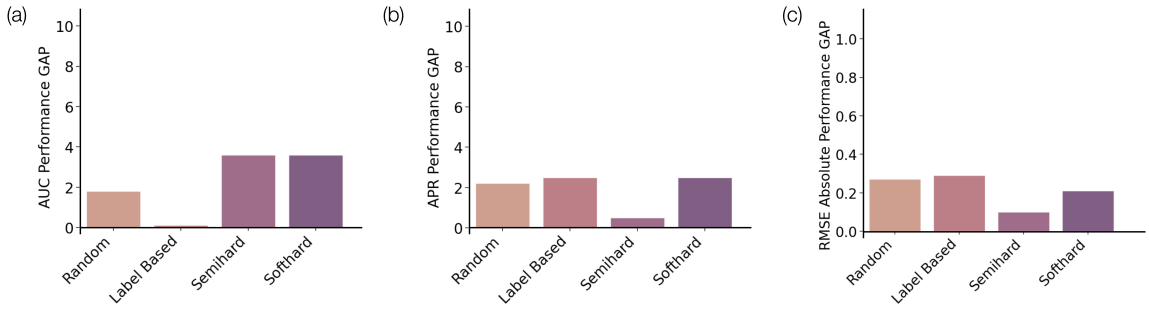

Figure 4: Effect of supDML hard negative mining on performance gap. Performance gap of male and female test subgroups on mPCWP inference tasks (Classification ((a) AUC, (b) APR) and Regression ((c) RMSE)) are reported with a bar plot. A positive value indicates that the male group performed better than the female group, and a negative value indicates the other way around. We report the absolute difference for (c) RMSE.

6. Discussion

In this work, we proposed a DML approach for the robust representations of electrocardiogram (ECG) signals for the hemodynamics inference task. Our proposed DML approach significantly outperforms random initialization, supervised baseline and existing contrastive learning methods, where supervised DML achieved the best regression performance and self-supervised DML methods achieved the best classification performance. This result demonstrates its ability to learn contextual representations preexisting in ECGs for better generalizability and robustness. More importantly, we showed with our experiment that supervised and self-supervised DML achieves the fair inference of mPCWP across different age and gender subgroups, which significantly improves the subgroup performance gaps in contrastive and supervised baselines. Our proposed method captures the changes in hemodynamics based on individual medical history, promising a potential to be applied for personalized physiological tracking. Our work highlights the potential of DML approaches for improving inference using ECG signals and other biomedical signals.

In our analysis of fairness in supervised and self-supervised DML models, we have observed some interesting trends (Table 3). In the supervised baseline and contrastive learning baselines (PCLR, SimCLR, CLOCS), a majority of instances in the minority gender group (Female) are surrounded by instances of the same gender group, indicating less uniformity in the embedding space. However, this trend is less apparent in both supervised and self-supervised DML, where fewer same-gender instances are found in the k-nearest neighbors of the minority group (Female). Supervised and self-supervised DML models have

fewer instances of same-gender subgroups surrounding the data points (e.g., female groups are surrounded by 68.21% of the same gender in 2-NN for supervised DML. Females are surrounded by 69.27% same-gender instances in self-supervised DML). This could be an explanation for the decreased performance disparity between subgroups in supervised and self-supervised DML models.

Table 3: The proportion (%) of k-nearest neighbors of the embedding space belonging to the same gender with k = 2, 3, and 5, to assess the degree of gender-based subgroup segregation in the representation space.

Models	k=2	k=3	k=5
Supervised Baseline	81.08	73.84	65.18
Supervised DML	68.21	59.31	51.81
PCLR	69.87	60.18	51.21
SimCLR	71.73	60.71	51.84
CLOCS	73.84	64.37	56.01
Self-Supervised DML	69.27	58.60	50.80

As DML effectively leverages the notion of similarity, it improves the quality of embedding by locating similar samples together although not explicitly ensuring fair prediction. In the context of fairness in DML, previous studies have made significant strides. Ilvento (2019) proposed a method for a metric approximation that specifically targets individual fairness. On the other hand, Dullerud et al. (2021) pioneered new fairness metrics in DML, and proposed a way of de-correlating the propagated bias in DML representation. Our proposed DML approach also pushed toward fairness in DML. Supervised and self-supervised DML not only provided robust representations for ECG signals in hemodynamics inference tasks but also demonstrated improved fairness in gender subgroup predictions. Our results suggest that DML models are capable of achieving uniformity of demographics in the embedding space, and this could help mitigate subgroup performance gaps, contributing to more equitable inference. These findings support the application of DML approaches for ECG signals and other biomedical signals, ultimately promoting fair and personalized physiological surveillance for diverse patient populations.

Limitations Despite the promising results, our study has several limitations. First, our analysis relies on internal validation. Leveraging the secondary source of ECG data with matched mPCWP for external validation can further solidify the application of DML models on mPCWP inference. Second, we have only focused on a limited number of problems related to intracardiac pressure classification and regression. Further research is needed to investigate the effectiveness of DML pretraining for other types of downstream tasks detecting cardiac conditions such as arrhythmia (Hannun et al., 2019), myocardial infarction (Acharya et al., 2017), and heart failure (Masetic and Subasi, 2016; Acharya et al., 2019). The generalizability of the DML pre-trained model can be further validated by finetuning it on the public benchmark datasets with larger labeled data corpus, such as PhysioNet 2021 Reyna et al. (2021), Chapman (Zheng et al., 2020a,b), and PTB-XL (Wagner et al., 2020) datasets. Auditing the models on different subgroups (e.g., self-reported racial groups,

insurance plans) available in those datasets would help validate the fairness gain of the model. Furthermore, we can assess the intersections of various sensitive features (Morina et al., 2019) to gain greater insight into achieving equitable prediction of the model.

7. Acknowledgements

This research was supported in part by funds from Quanta Computer, Inc. We would like to thank members of the Computational Cardiovascular Research Group and HealthyML Lab at MIT, and the reviewers for their invaluable feedback.

References

William T Abraham, Philip B Adamson, Robert C Bourge, Mark F Aaron, Maria Rosa Costanzo, Lynne W Stevenson, Warren Strickland, Suresh Neelagaru, Nirav Raval, Steven Krueger, et al. Wireless pulmonary artery haemodynamic monitoring in chronic heart failure: a randomised controlled trial. *The Lancet*, 377(9766):658–666, 2011.

William T Abraham, Lynne W Stevenson, Robert C Bourge, Jo Ann Lindenfeld, Jordan G Bauman, and Philip B Adamson. Sustained efficacy of pulmonary artery pressure to guide adjustment of chronic heart failure therapy: complete follow-up results from the champion randomised trial. *The Lancet*, 387(10017):453–461, 2016.

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*, 415:190–198, 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.

Zachi I Attia, Paul A Friedman, Peter A Noseworthy, Francisco Lopez-Jimenez, Dorothy J Ladewig, Gaurav Satam, Patricia A Pellikka, Thomas M Munger, Samuel J Asirvatham, Christopher G Scott, et al. Age and sex estimation using artificial intelligence from standard 12-lead ecgs. *Circulation: Arrhythmia and Electrophysiology*, 12(9):e007284, 2019.

Guillaume Baudry, Guillaume Coutance, Richard Dorent, Fabrice Bauer, Katrien Blanchart, Aude Boignard, Céline Chabanne, Clément Delmas, Nicolas d'Ostrevy, Eric Epailly, et al. Prognosis value of forrester's classification in advanced heart failure patients awaiting heart transplantation. *ESC Heart Failure*, 9(5):3287–3297, 2022.

Daniel Burkhoff, Mathew S Maurer, and Milton Packer. Heart failure with a normal ejection fraction: is it really a disorder of diastolic function?, 2003.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

Joseph Y Cheng, Hanlin Goh, Kaan Dogrusoz, Oncel Tuzel, and Erdrin Azemi. Subject-aware contrastive learning for biosignals. *arXiv preprint arXiv:2007.04871*, 2020.

Nathaniel Diamant, Erik Reinertsen, Steven Song, Aaron D Aguirre, Collin M Stultz, and Puneet Batra. Patient contrastive learning: A performant, expressive, and practical approach to electrocardiogram modeling. *PLoS Computational Biology*, 18(2):e1009862, 2022.

Mark H Drazner, Anne S Hellkamp, Carl V Leier, Monica R Shah, Leslie W Miller, Stuart D Russell, James B Young, Robert M Califf, and Anju Nohria. Value of clinician assessment of hemodynamics in advanced heart failure: the escape trial. *Circulation: Heart Failure*, 1(3):170–177, 2008.

Natalie Dullerud, Karsten Roth, Kimia Hamidieh, Nicolas Papernot, and Marzyeh Ghassemi. Is fairness only metric deep? evaluating and addressing subgroup gaps in deep metric learning. In *International Conference on Learning Representations*, 2021.

Ujjal Kr Dutta, Mehrtash Harandi, and C Chandra Sekhar. Unsupervised metric learning with synthetic examples. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3834–3841, 2020a.

Ujjal Kr Dutta, Mehrtash Harandi, and Chellu Chandra Sekhar. Unsupervised deep metric learning via orthogonality based probabilistic loss. *IEEE Transactions on Artificial Intelligence*, 1(1):74–84, 2020b.

Zheren Fu, Yan Li, Zhendong Mao, Quan Wang, and Yongdong Zhang. Deep metric learning with self-supervised ranking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1370–1378, 2021a.

Zheren Fu, Zhendong Mao, Chenggang Yan, An-An Liu, Hongtao Xie, and Yongdong Zhang. Self-supervised synthesis ranking for deep metric learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7):4736–4750, 2021b.

Bryan Gopal, Ryan W Han, Gautham Raghupathi, Andrew Y Ng, Geoffrey H Tison, and Pranav Rajpurkar. 3kg: Contrastive learning of 12-lead electrocardiograms using physiologically-inspired augmentations. *arXiv preprint arXiv:2106.04452*, 2021.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

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. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

Paul A Heidenreich, Biykem Bozkurt, David Aguilar, Larry A Allen, Joni J Byun, Monica M Colvin, Anita Deswal, Mark H Drazner, Shannon M Dunlay, Linda R Evers, et al. 2022 aha/acc/hfsa guideline for the management of heart failure: a report of the american college of cardiology/american heart association joint committee on clinical practice guidelines. *Journal of the American College of Cardiology*, 79(17):e263–e421, 2022.

Junlin Hu, Jiwen Lu, and Yap-Peng Tan. Discriminative deep metric learning for face verification in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1875–1882, 2014.

Zheng-Yi Huang, Qing-Qing Mu, Mian Hu, Ying Zou, and Jiang-Wen Xiao. Regression-based metric learning. In *2018 37th Chinese Control Conference (CCC)*, pages 9107–9112. IEEE, 2018.

Christina Ilvento. Metric learning for individual fairness. *arXiv preprint arXiv:1906.00250*, 2019.

Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. Mining on manifolds: Metric learning without labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7642–7651, 2018.

Herve Jegou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *IEEE transactions on pattern analysis and machine intelligence*, 33(1): 117–128, 2011.

Dani Kiyasseh, Tingting Zhu, and David A Clifton. Clocs: Contrastive learning of cardiac signals across space, time, and patients. In *International Conference on Machine Learning*, pages 5606–5615. PMLR, 2021.

Michel Komajda and Carolyn SP Lam. Heart failure with preserved ejection fraction: a clinical dilemma. *European heart journal*, 35(16):1022–1032, 2014.

Xiang Lan, Dianwen Ng, Shenda Hong, and Mengling Feng. Intra-inter subject self-supervised learning for multivariate cardiac signals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4532–4540, 2022.

Zerina Masetic and Abdulhamit Subasi. Congestive heart failure detection using random forest classifier. *Computer methods and programs in biomedicine*, 130:54–64, 2016.

Temesgen Mehari and Nils Strodthoff. Self-supervised representation learning from 12-lead ecg data. *Computers in Biology and Medicine*, 141:105114, 2022.

Timo Milbich, Karsten Roth, Samarth Sinha, Ludwig Schmidt, Marzyeh Ghassemi, and Bjorn Ommer. Characterizing generalization under out-of-distribution shifts in deep metric learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Giulio Morina, Viktoriia Oliinyk, Julian Waton, Ines Marusic, and Konstantinos Georgatzis. Auditing and achieving intersectional fairness in classification problems. *arXiv preprint arXiv:1911.01468*, 2019.

Meinard Müller. Dynamic time warping. *Information retrieval for music and motion*, pages 69–84, 2007.

Sherif F Nagueh, Katherine J Middleton, Helen A Kopelen, William A Zoghbi, and Miguel A Quiñones. Doppler tissue imaging: a noninvasive technique for evaluation of left ventricular relaxation and estimation of filling pressures. *Journal of the American College of Cardiology*, 30(6):1527–1533, 1997.

Sherif F Nagueh, Christopher P Appleton, Thierry C Gillebert, Paolo N Marino, Jae K Oh, Otto A Smiseth, Alan D Waggoner, Frank A Flachskampf, Patricia A Pellikka, and Arturo Evangelisa. Recommendations for the evaluation of left ventricular diastolic function by echocardiography. *European journal of echocardiography*, 10(2):165–193, 2009.

Jungwoo Oh, Hyunseung Chung, Joon-myoung Kwon, Dong-gyun Hong, and Edward Choi. Lead-agnostic self-supervised learning for local and global representations of electrocardiogram. In *Conference on Health, Inference, and Learning*, pages 338–353. PMLR, 2022.

Thiago M Paixao, Rodrigo F Berriel, Maria Boeres, Alessandro L Koerich, Claudine Badue, Alberto F De Souza, and Thiago Oliveira-Santos. Fast (er) reconstruction of shredded text documents via self-supervised deep asymmetric metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14343–14351, 2020.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Aniruddh Raghu, Payal Chandak, Ridwan Alam, John Guttag, and Collin Stultz. Sequential multi-dimensional self-supervised learning for clinical time series. In *International Conference on Machine Learning*. PMLR, 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. *Scientific Reports*, 13(1): 3923, 2023b.

Matthew A Reyna, Nadi Sadr, Erick A Perez Alday, Annie Gu, Amit J Shah, Chad Robichaux, Ali Bahrami Rad, Andoni Elola, Salman Seyedi, Sardar Ansari, et al. Will two do? varying dimensions in electrocardiography: the physionet/computing in cardiology challenge 2021. In *2021 Computing in Cardiology (CinC)*, volume 48, pages 1–4. IEEE, 2021.

Véronique L Roger. Epidemiology of heart failure. *Circulation research*, 113(6):646–659, 2013.

Karsten Roth, Timo Milbich, Samarth Sinha, Prateek Gupta, Bjorn Ommer, and Joseph Paul Cohen. Revisiting training strategies and generalization performance in deep metric learning. In *International Conference on Machine Learning*, pages 8242–8252. PMLR, 2020.

Karsten Roth, Timo Milbich, Bjorn Ommer, Joseph Paul Cohen, and Marzyeh Ghassemi. Simultaneous similarity-based self-distillation for deep metric learning. In *International Conference on Machine Learning*, pages 9095–9106. PMLR, 2021.

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):100003, 2022.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016.

Sunit Tolia, Zubair Khan, Gunjan Gholkar, and Marcel Zughaib. Validating left ventricular filling pressure measurements in patients with congestive heart failure: Cardiomems™ pulmonary arterial diastolic pressure versus left atrial pressure measurement by transthoracic echocardiography. *Cardiology Research and Practice*, 2018, 2018.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I Lunze, Wojciech Samek, and Tobias Schaeffter. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific data*, 7(1):1–15, 2020.

Jian Wang, Feng Zhou, Shilei Wen, Xiao Liu, and Yuanqing Lin. Deep metric learning with angular loss. In *Proceedings of the IEEE international conference on computer vision*, pages 2593–2601, 2017.

Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021. doi: 10.21105/joss.03021. URL https://doi.org/10.21105/joss.03021.

Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krahenbuhl. Sampling matters in deep embedding learning. In *Proceedings of the IEEE international conference on computer vision*, pages 2840–2848, 2017.

Baosheng Yu, Tongliang Liu, Mingming Gong, Changxing Ding, and Dacheng Tao. Correcting the triplet selection bias for triplet loss. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 71–87, 2018.

Junsheng Yu, Xiangqing Wang, Xiaodong Chen, and Jinglin Guo. Automatic premature ventricular contraction detection using deep metric learning and knn. *Biosensors*, 11(3): 69, 2021.

Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik. Self-supervised contrastive pre-training for time series via time-frequency consistency. *Advances in Neural Information Processing Systems*, 35:3988–4003, 2022.

Jianwei Zheng, Huimin Chu, Daniele Struppa, Jianming Zhang, Magdi Yacoub, Hesham El-Askary, Anthony Chang, Louis Ehwerhemuepha, Islam Abudayyeh, Alexander Barrett, et al. Optimal multi-stage arrhythmia classification approach. *Scientific reports*, 10(1): 1–17, 2020a.

Jianwei Zheng, Jianming Zhang, Sidy Danioko, Hai Yao, Hangyuan Guo, and Cyril Rakovski. A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. *Scientific data*, 7(1):1–8, 2020b.

Guiping Zhu, Mingzhu Ma, Yuwen Huang, Kuikui Wang, and Gongping Yang. Dual-domain low-rank fusion deep metric learning for off-the-person ecg biometrics. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2914–2918. IEEE, 2022.

Appendix

In the appendix section of this work, we provide a more detailed and formalized explanation of various DML objectives, batch mining strategies, and loss functions introduced in the main paper. Specifically, we introduce the formalization for supervised DML joint-learning classification and regression objectives, as well as self-supervised downstream objectives (Sections A). Additionally, we introduce other loss objectives, including margin loss and angular loss (Section B) and discuss mining techniques incorporating slack parameters into hard mining objectives, such as semi-hard and soft-hard mining (Section C). We then introduce the results and discussions for the subgroup performance gap (Section D and explain clinical implications (Section E).

Throughout the Appendix, we used the same notation introduced in Section 3: x_i and y_i are input data instances and labels, respectively, in D_{sup} with the cardinal N. Instances in D_{sup} are sampled for triplet mining, where x_a is an anchor, x_p is positive, and x_n is negative. f_Φ is the metric embedding where we apply the DML objectives. For either joint training or finetuning hemodynamics inference, we use additional FC layer g.

Appendix A. Supervised DML and Self-supervised DML Learning Objectives

A.1. Supervised Deep Metric Learning for Classification

Below is the detailed formulation of the DML objective of SupDML joint classification, where α is the scaling coefficient of triplet loss and σ is the sigmoid function.

$$\begin{aligned}\mathcal{L}_{tot} &= \mathcal{L}_{CE} + \alpha\mathcal{L}_{triplet} \\ &= -\sum_{i=1}^{N} y_i log(\sigma(g(f_\Phi(x_i))) + \alpha\sum_{i=1}^{N}\{||f_\Phi(x_a) - f_\Phi(x_p)|| - ||f_\Phi(x_a) - f_\Phi(x_n)||\}\end{aligned}$$

A.2. Supervised Deep Metric Learning for Regression

Using the triplet (x_a, x_p, x_n), the model f_Φ is optimized by triplet loss ($\mathcal{L}_{Triplet}$) while at the same time the final classifier $g \cdot f_\Phi$ is jointly optimized with Root Mean Square Error (RMSE) with the triplet scaling factor of α: $\mathcal{L}_{tot} = \mathcal{L}_{RMSE} + \alpha\mathcal{L}_{triplet}$. RMSE for downstream classification loss is $\mathcal{L}_{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} ||y_i - g(f_\Phi(x_i))||^2}{n}}$ and the equation for the full loss term is:

$$\begin{aligned}\mathcal{L}_{tot} &= \mathcal{L}_{RMSE} + \alpha\mathcal{L}_{triplet} \\ &= \sqrt{\frac{\sum_{i=1}^{N} ||y_i - g(f_\Phi(x_i))||^2}{n}} + \alpha\sum_{i=1}^{N}\{||f_\Phi(x_a) - f_\Phi(x_p)|| - ||f_\Phi(x_a) - f_\Phi(x_n)||\}\end{aligned}$$

Table 4: Performance of downstream mPCWP classification task with various triplet loss weight coefficient α. We report the mean performance and its standard deviation of 1,0000 bootstraps. Also, the optimal performance for each dimension is highlighted in bold.

		Classification		Regression
Dim	α	AUC	APR	RMSE
128	0.1	75.4 ± 1.8	**57.2 ± 3.3**	7.37 ± 0.25
	1.0	74.6 ± 1.8	**57.2 ± 3.2**	7.50 ± 0.22
	2.0	74.6 ± 1.8	56.2 ± 3.3	**7.15 ± 0.21**
	3.0	**76.7 ± 1.6**	56.1 ± 3.3	7.23 ± 0.21
	10.0	76.2 ± 1.7	55.0 ± 3.1	7.27 ± 0.20
256	0.1	73.1 ± 1.8	54.2 ± 3.2	7.53 ± 0.22
	1.0	75.3 ± 1.7	53.2 ± 3.1	7.55 ± 0.21
	2.0	**76.5 ± 1.7**	**57.2 ± 3.2**	7.27 ± 0.21
	3.0	75.5 ± 1.7	57.1 ± 3.2	**7.19 ± 0.20**
	10.0	74.1 ± 1.7	53.4 ± 3.1	7.37 ± 0.22
512	0.1	75.8 ± 1.8	55.2 ± 3.3	7.67 ± 0.24
	1.0	**77.1 ± 1.8**	**59.6 ± 3.2**	7.44 ± 0.23
	2.0	76.6 ± 1.6	55.8 ± 3.2	7.30 ± 0.23
	3.0	74.9 ± 1.7	56.9 ± 3.2	**7.21 ± 0.18**
	10.0	75.6 ± 1.7	55.5 ± 3.1	7.32 ± 0.22
1024	0.1	74.9 ± 1.7	55.8 ± 3.2	7.66 ± 0.23
	1.0	**77.5 ± 1.7**	**58.1 ± 2.9**	7.42 ± 0.18
	2.0	76.0 ± 1.7	55.7 ± 3.3	7.43 ± 0.20
	3.0	75.0 ± 1.8	54.9 ± 3.3	**7.20 ± 0.20**
	10.0	76.1 ± 1.7	54.8 ± 3.1	7.32 ± 0.21

A.3. Optimal Scaling Coefficient α for SupDML Classification and Regression

We present the results of our search for the optimal scaling factor α for the triplet loss and its effects on performance. We performed a greedy search for the best scaling factor α among the values 0.1, 1.0, 2.0, 3.0, and 10.0 for each embedding dimension of f_Φ (128, 256, 512, 1024) (Table 4).

We report the best performance among different α values for the main supDML results (Table 1): $\alpha = 3.0$ yielded the best performance for the classification of the supDML experiment with an embedding f_Φ dimension of 128, while $\alpha = 2.0$ was optimal for joint regression with a 128-dimensional embedding. For an embedding f_Φ dimension of 256, $\alpha = 2.0$ was best for joint classification, and $\alpha = 3.0$ was best for regression. With an embedding f_Φ dimension of 512, $\alpha = 1.0$ worked best for joint classification, and $\alpha = 3.0$ was optimal for regression. Lastly, for an embedding f_Φ dimension of 1024, $\alpha = 1.0$ was best for joint classification, and $\alpha = 3.0$ was optimal for regression.

A.4. Downstream Hemodynamics Inference for Self-Supervised DML

The pre-trained model f_Φ with the DML objective and finetune with the downstream hemodynamics classification and regression task with $\mathcal{D}_{sup}$. The objective uses the cross-entropy loss for binary classification ($\mathcal{L}_{CE}$) and RMSE for regression ($\mathcal{L}_{RMSE}$):

$$\mathcal{L}_{CE} = -\sum_{i=1}^{N} y_i log(\sigma(f_\Phi(x_i))$$

$$\mathcal{L}_{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} ||y_i - g(f_\Phi(x_i))||^2}{n}}$$

Appendix B. Deep Metric Learning Loss Functions

We used two different DML objectives to extend the use of triplet loss: margin loss and angular loss.

Margin Loss (Wu et al., 2017) extends the standard triplet loss by incorporating a dynamic, learnable distance boundary β between the samples with the same labels and different labels (relative ordering, $\mathcal{P} = \{(x_i, x_j)|x_i \in \mathcal{B}, x_j \in \mathcal{B}\}$), which has transformed from the usual triplet ranking problem. The learning rate of the boundary, β is set to 0.0005, with an initial β value of 1.2 and a triplet margin γ of 0.2. The implementation and hyperparameters followed the code from Roth et al. (2020).

$$\mathcal{L}_{margin} = \sum_{(x_i,x_j)\in\mathcal{P}} \mathbb{1}_{y_i=y_j}(||f_\Phi(x_i) - f_\Phi(x_j)||_2^2 - \beta) - \mathbb{1}_{y_i\neq y_j}(||f_\Phi(x_i) - f_\Phi(x_j)||_2^2 - \beta) + \gamma$$

Angular Loss (Wang et al., 2017) The angular loss constrains the angle at the negative point of triplet triangles, resulting in higher convergence, achieving scale invariance and third-order geometric restrictions. We used the proposed parameter from the original paper, angular margin $\alpha = \frac{\pi}{4}$, and the parameter for the trade-off between npair and angular loss, $\lambda = 2$. The objective function below uses the triplets anchor x_a, positive x_p and negative x_n:

$$\mathcal{L}_{angular} = \mathcal{L}_{npair} + \frac{\lambda}{b}\sum[\log(1 + \sum \exp(4\tan^2(\alpha(f_\Phi(x_a) + f_\Phi(x_p))^T f_\Phi(x_n))) - 2(1 + \tan^2(\alpha))f_\Phi(x_a)^T f_\Phi(x_n)]$$

where $\mathcal{L}_{npair}$ is defined as below with the embedding regularization $\nu = 0.005$ following the original paper (Sohn, 2016):

$$\mathcal{L}_{npair} = \frac{1}{b}\sum \log(1 + \sum \exp(f_\Phi(x_a)^T f_\Phi(x_n) - f_\Phi(x_a)^T f_\Phi(x_p))) + \frac{\nu}{b}\sum_{x_i\in\mathcal{B}} ||f_\Phi(x_i)||_2^2$$

Table 5 summarizes the classification and regression performance with margin loss and angular loss, compared with triplet loss.

Table 5: Performance of the downstream Wedge pressure elevation classification task with various loss functions for supervised DML with an embedding dimension of 128.

	Classification		Regression
	AUC	APR	RMSE
Triplet Loss	**76.7 ± 1.6**	**56.1 ± 3.3**	**7.15** ± 0.21
Margin Loss	74.7 ± 1.7	53.4 ± 3.3	7.55 ± 0.22
Angular Loss	**76.8 ± 1.8**	**56.8 ± 3.1**	7.59 ± 0.22

Appendix C. Triplet Mining for Supervised DML

We formalize various mining techniques we used for supervised DML, including continuous label-based mining (*Label Based*) and hard negative mining techniques (*Semihard*, *Softhard*).

C.1. Continuous Label Based Mining

DML Triplet Selection: For a given anchor x_a, we define the positive sample x_p to be the input sample with the smallest absolute label difference with the anchor: $x_p = \{x_i | i = \arg\min_{y_i \in \mathcal{B}} |y_i - y_a|\}$ where $\mathcal{B}$ is the batch. Then the negative sample x_n is then randomly sampled from the set $\mathcal{N}_\mathcal{B} = \{x_j | j = \arg\max_{y_j \in \mathcal{B}} |y_j - y_a|\}$. We also performed greedy search hyperparameter tuning for the scaling factor α for label based mining loss and reported the best-performing model in Table 6. With an embedding dimension f_Φ of dimensions 128 and 256, α 10.0 worked the best, and for 512 and 1024 α of 0.1 worked the best.

C.2. Hard Negative Mining

Semihard Mining (Schroff et al., 2015) is a method to sample hard negatives based on the distance of negative samples with respect to anchors and positives. Semihard mining selects the fairly hard negatives with the embedding distance to the anchor embedding larger than that is from positive embedding, within a certain distance margin. If we let the embedding of negative samples that are to be selected for triplet pair $f_\Phi(x_n)$, then the selected negatives are:

$$x_n \in \{x_n | x_n \in \mathcal{B}, y_n \neq y_a, ||f_\Phi(x_a) - f_\Phi(x_p)||_2^2 < ||f_\Phi(x_a) - f_\Phi(x_n)||_2^2\}$$

Softhard Mining (Roth et al., 2020) introduces a soft (probabilistic) selection of hard negatives. As observed by Roth et al. (2020), this decreases the likelihood of potential model collapses and undesirable local minima, which were the issues with semihard mining (Schroff et al., 2015). We select the hard negatives based on the two equations below:

$$x_n \in \{x_n | x_n \in \mathcal{B}, y_n \neq y_a, ||f_\Phi(x_a) - f_\Phi(x_n)||_2^2 < \arg\max_{x_p \in \mathcal{B}, y_a = y_p} ||f_\Phi(x_a) - f_\Phi(x_p)||_2^2\}$$

Table 6: Performance of the downstream mPCWP elevation classification and regression task with various mining methods for supervised DML (Random, Semihard, Softhard)

		Classification		Regression	Embedding
Mining	Dim	AUC	APR	RMSE	Recall@1
Random	128	76.7 ± 1.6	56.1 ± 3.3	**7.15** ± 0.21	**69.6**
	256	76.5 ± 1.7	57.2 ± 3.2	7.19 ± 0.20	66.5
	512	**77.1** ± **1.8**	**59.6** ± **3.2**	7.21 ± 0.18	67.5
	1024	**77.5** ± **1.7**	58.1 ± 2.9	7.20 ± 0.20	66.3
Label Based	128	76.1 ± 1.7	56.5 ± 1.3	7.29 ± 0.21	67.5
	256	75.8 ± 1.8	55.6 ± 3.2	7.39 ± 0.20	67.4
	512	74.9 ± 1.8	54.8 ± 3.3	7.40 ± 0.24	59.1
	1024	**77.4** ± **1.6**	57.6 ± 3.1	7.37 ± 0.23	65.4
Semihard	128	75.1 ± 1.8	56.4 ± 3.4	7.50 ± 0.21	66.2
	256	76.9 ± 1.7	56.5 ± 3.4	7.51 ± 0.20	66.2
	512	75.0 ± 1.8	57.5 ± 3.1	7.46 ± 0.22	65.3
	1024	74.4 ± 1.7	51.5 ± 3.2	**7.19** ± **0.21**	66.1
Softhard	128	**77.4** ± **1.7**	58.1 ± 3.1	7.53 ± 0.22	67.9
	256	75.9 ± 1.7	57.7 ± 3.2	7.38 ± 0.21	**69.3**
	512	74.3 ± 1.7	54.0 ± 3.1	7.50 ± 0.21	67.9
	1024	74.8 ± 1.8	52.2 ± 3.3	7.51 ± 0.21	65.4

and

$$x_n \in \{x_n | x_n \in \mathcal{B}, y_n \neq y_a, ||f_\Phi(x_a) - f_\Phi(x_n)||_2^2 > \underset{x_n \in \mathcal{B}, y_a \neq y_n}{\arg\min} ||f_\Phi(x_a) - f_\Phi(x_n)||_2^2\}$$

Table 6 summarizes the classification and regression performance with hard negative minings compared with random mining. Table 7 reports the performance of each gender subgroup and the performance gap.

Appendix D. Subgroup Performance Gap

In this paper, we showed the biases in data influencing the model performance, which can be improved using a supervised and self-supervised DML training scheme. Our dataset is over-represented by certain demographic groups (gender, age), and models could potentially perform better on these groups than on under-represented groups. As the female demographic group and age group who are between 18 to 35 are under-represented in our training data, baselines may be less accurate in predicting outcomes for these groups. This is why it is crucial to perform fairness audits and continue to evaluate the deployed performance of such models in other tasks.

D.1. Prevalence and Performance Metrics

Among our performance metrics, APR could be dependent on prevalence, as it influences the precision values. However, we found that the prevalence of the positive class (mPCWP

Table 7: Subgroup performance for the downstream Wedge pressure elevation classification task with various mining methods for supervised DML (Random, Semihard, Softhard). The embedding dimension used here is 128.

	Test	Classification		Regression
Mining	Subgroup	AUC	APR	RMSE
Random	Full	76.7 ± 1.6	56.1 ± 3.3	7.15 ± 0.21
	Male	77.4 ± 1.5	54.2 ± 2.8	7.07 ± 0.19
	Female	75.6 ± 1.2	56.4 ± 2.4	7.34 ± 0.17
	Gap (M-F)	1.8	2.2	0.27
Label Based	Full	76.1 ± 1.7	56.5 ± 1.3	7.29 ± 0.21
	Male	75.3 ± 1.0	54.7 ± 1.9	7.11 ± 0.17
	Female	75.4 ± 1.3	57.3 ± 2.5	7.40 ± 0.17
	Gap (M-F)	**0.1**	2.5	0.29
Semihard	Full	75.1 ± 1.8	56.4 ± 3.4	7.50 ± 0.21
	Male	77.5 ± 1.6	56.3 ± 3.0	7.39 ± 0.18
	Female	73.9 ± 1.4	56.9 ± 2.3	7.49 ± 0.26
	Gap (M-F)	3.6	**0.5**	**0.10**
Softhard	Full	77.4 ± 1.7	58.1 ± 3.1	7.53 ± 0.22
	Male	78.5 ± 1.5	54.5 ± 3.0	7.71 ± 0.18
	Female	74.9 ± 1.3	57.0 ± 2.4	7.92 ± 0.17
	Gap (M-F)	3.6	2.5	0.21

$>$ 18) across gender groups in our dataset is similar: Female/Male prevalence ratio of 0.461/0.441 for the Train set, 0.395/0.410 for the Validation set, and 0.466/0.385 for the Test set. The APR is the correct metric in this case because it provides a measure of model performance across various threshold levels and is less sensitive to class imbalances than other metrics such as accuracy.

D.2. Age Subgroup Performance Gap

We present a summary of performance gaps in mPCWP inference across age groups in Table 8. The gaps in performance are the averaged relative pairwise performance differences between all combinations of age groups. Both DML models achieved fairer downstream task performance across different age subgroups against the baselines. Compared to the supervised baseline, supervised DML not only achieved narrower gaps in both AUC and APR but also surpassed overall regression performance, while maintaining equivalent classification performance. Self-supervised DML showed lower APR and RMSE gaps compared to random initialization and self-supervised contrastive learning baselines (SimCLR, CLOCS, PCLR).

Appendix E. Clinical Implication and Potential Future Works

Our proposed supervised DML and self-supervised DML non-invasively estimate central hemodynamics using the ECG signal. Current non-invasive methods for assessing a number

Table 8: Age subgroup performance gap of downstream mPCWP classification task has decreased on either Supervised or Self-Supervised DML compared to supervised and contrastive baselines (with an embedding dimension of 128 for DML and contrastive learning encoder). We report the mean with a standard deviation based on 1,000 bootstraps, and models with the lowest performance gaps are boldfaced. We included four age subgroups: 18-35, 35-50, 50-75, 75-. The average gap indicates the average pairwise group performance gap

			Classification		Regression
	Models	Subgroup	AUC	APR	RMSE
Random Init		Full	74.3 ± 1.9	55.7 ± 3.3	7.88 ± 0.26
		18-35	80.6 ± 12.8	53.0 ± 21.9	3.63 ± 0.52
		35-50	74.7 ± 1.8	54.4 ± 1.8	7.99 ± 0.76
		50-75	74.6 ± 0.9	53.8 ± 1.6	8.51 ± 0.55
		75-	74.6 ± 0.8	54.8 ± 1.6	8.64 ± 0.82
		Average Gap	3.0	1.0	2.6
Supervised	Supervised Baseline	Full	78.4 ± 1.1	59.5 ± 2.2	7.45 ± 0.15
		18-35	83.5 ± 5.4	59.4 ± 11.7	5.07 ± 0.51
		35-50	73.2 ± 4.2	59.5 ± 6.3	7.41 ± 0.46
		50-75	79.8 ± 1.5	55.6 ± 3.1	7.30 ± 0.18
		75-	66.7 ± 2.8	69.1 ± 3.6	8.28 ± 0.37
		Average Gap	9.5	6.8	**1.62**
	Supervised DML (Random Mining)	Full	77.3 ± 3.6	57.8 ± 3.3	7.33 ± 0.23
		18-35	77.4 ± 1.6	57.5 ± 3.3	3.07 ± 0.39
		35-50	77.3 ± 1.6	56.6 ± 3.1	7.31 ± 0.23
		50-75	77.8 ± 2.2	60.0 ± 3.9	6.94 ± 0.21
		75-	77.4 ± 1.2	57.6 ± 2.4	7.33 ± 0.16
		Average Gap	**0.3**	**1.7**	2.19
Self-Supervised	SimCLR	Full	75.0 ± 1.9	58.2 ± 3.2	7.58 ± 0.22
		18-35	74.8 ± 1.7	57.1 ± 3.1	3.93 ± 0.64
		35-50	74.2 ± 1.7	55.8 ± 3.0	7.32 ± 0.84
		50-75	74.8 ± 2.3	52.3 ± 4.1	7.32 ± 0.22
		75-	74.7 ± 1.3	56.7 ± 2.2	7.41 ± 0.15
		Average Gap	**0.3**	2.5	1.74
	CLOCS	Full	75.9 ± 1.7	56.2 ± 3.4	7.45 ± 0.22
		18-35	73.5 ± 1.9	52.9 ± 3.2	4.47 ± 0.89
		35-50	72.7 ± 1.7	52.5 ± 2.9	7.50 ± 0.70
		50-75	74.5 ± 2.3	51.5 ± 3.9	7.50 ± 0.26
		75-	73.4 ± 1.3	53.6 ± 2.2	7.66 ± 0.17
		Average Gap	0.9	1.1	1.60
	PCLR	Full	74.2 ± 1.6	54.2 ± 2.5	10.04 ± 0.27
		18-35	65.6 ± 16.1	65.9 ± 15.3	8.34 ± 1.31
		35-50	73.1 ± 4.6	57.2 ± 7.6	11.29 ± 0.84
		50-75	74.9 ± 1.9	51.7 ± 3.3	10.58 ± 0.30
		75-	63.8 ± 3.8	61.1 ± 4.7	11.72 ± 0.69
		Average Gap	6.8	15.7	1.81
	Self-supervised DML (DTW)	Full	77.2 ± 1.7	60.5 ± 3.1	7.65 ± 0.21
		18-35	74.9 ± 1.0	54.7 ± 1.8	6.97 ± 1.22
		35-50	74.8 ± 0.9	54.6 ± 1.8	7.68 ± 0.74
		50-75	76.2 ± 2.1	52.8 ± 4.1	7.99 ± 0.24
		75-	75.2 ± 0.8	54.8 ± 1.5	8.19 ± 0.17
		Average Gap	0.7	**0.2**	**0.67**

of hemodynamic measurements include clinical analysis using echocardiography (Nagueh et al., 1997, 2009; Tolia et al., 2018). However, this method requires trained personnel to obtain ultrasound images and a specialist to interpret those images, hence it is challenging to use this information in routine practice. Indeed, such information is not readily available in many healthcare settings. On the contrary, our methods will be easily incorporated into clinical workflows to aid in early diagnosis and continuous patient monitoring, and thus contribute to personalized patient management. For instance, the model can be modified to get the ECG signal from one lead to get the limb lead signal from a smartwatch or

portable device, and connected to a mobile app for heart failure patient early monitoring. Or, it can be connected to implantable CardioMEMS^{TM} (Abraham et al., 2011, 2016; Tolia et al., 2018) to get the measured hemodynamics to monitor patient status. Our approach would help screen out the patients with abnormal mPCWP and help physicians assess treatment response.