Unsupervised Deep Metric Learning for the inference of hemodynamic value with Electrocardiogram signals

Hyewon Jeong
hyewonj@mit.edu
MIT

Marzyeh Ghassemi
mghassem@mit.edu
MIT

Collin Stultz
cmstultz@mit.edu
MIT

Abstract

An objective assessment of intrathoracic pressures remains an important diagnostic method for patients with heart failure. Although cardiac catheterization is the gold standard for estimating central hemodynamic pressures, it is an invasive procedure where a pressure transducer is inserted into a great vessel and threaded into the right heart chambers. 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 that were trained in a supervised fashion to estimate central pressures have shown good discriminatory ability over a heterogeneous cohort when the number of training examples is large. As obtaining central pressures (the labels) requires an invasive procedure that can only be performed in an inpatient setting, acquiring large labeled datasets for different patient cohorts is challenging. In this work, we leverage a dataset that contains over 5.4 million ECGs, without concomitant central pressure labels, to improve the performance of models trained with sparsely labeled datasets. Using a deep metric learning (DML) objective function, we develop a procedure for building latent 12-lead ECG representations and demonstrate that these latent representations can be used to improve the discriminatory performance of a model trained in a supervised fashion on a smaller labeled dataset. More generally, our results show that training with DML objectives with both labeled and unlabeled ECGs showed the downstream performance on par with the supervised baseline.

1 Introduction

Cardiac catheterization is a procedure for measuring hemodynamics, and is an important diagnostic tool. For instance, intracardiac pressures play an important role in the assessment of hemodynamic severity in patients with heart failure. The mean pulmonary capillary wedge pressure (mPCWP) - an estimate of the left atrial pressure - is one measurement that is used for treatment and prognostication in heart failure patients. The mPCWP, however, is acquired via right heart catherization (RHC), which is an invasive procedure that is only performed in an inpatient setting. To provide readily available diagnostic tools to detect and understand heart failure, we need safe routine screening methods that reliably predict patient hemodynamics.

One universally available piece of clinical data is the electrocardiogram, which has shown promise with respect to helping to guide the care of patients with suspected heart failure [1]. To reduce the risk of invasive procedures, a ECG-based model can be constructed to infer when the mean pulmonary capillary wedge pressure (mPCWP) is above normal [2].

Previous studies [3, 4, 5, 6, 7] have leveraged large unlabeled datasets to infer cardiac abnormalities using the similarity information between ECGs. In particular, several approaches have used contrastive learning to construct latent representations of ECG data for downstream inference tasks [3, 4, 5].

Submitted to Learning from Time Series for Health (TS4H) Workshop in NeurIPS 2022. Do not distribute.

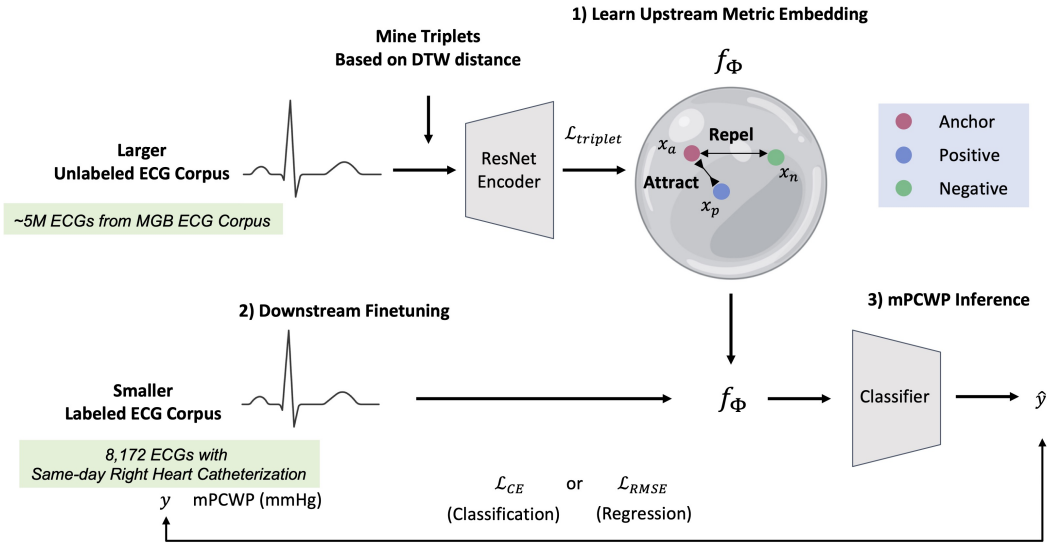

Figure 1: Schematic Diagram of Inferring cardiac pressure using ECGs without matched mPCWP. 1) Using larger unlabeld ECG corpus (ECGs without matched mPCWP), we learn upstream metric embedding f with deep metric learning objective. 2) Then we finetune this embedding f with smaller ECG dataset with matched mPCWP (labeled ECGs) to 3) infer mPCWP. Figure 2 includes the diagram for supervised deep metric learning with the joint learning of both triplet and finetuning task loss.

However, Deep Metric Learning (DML), a procedure that maps data to latent representations in a manner such that similar datapoints are mapped to similar points in the latent space, has not been used to develop learned representations of ECG data. Yet the use of DML for representation learning [8, 9, 10] has yielded frutiful results for other tasks such as few-shot image retrieval [11] or facial re-identification [12].

In this work, we apply DML to a real-world clinical task, intracardiac pressure classification and regression using the 12-lead ECGs. The goal is to arrive at learned representations that adequately represent the information within ECG signals, using a large resource of unlabeled ECGs (Figure 1). DML objective functions (e.g., triplet loss and angular loss) learn an embedding that preserves the inherent distance between ECGs by pulling similar or positive samples together and pushing dissimilar samples away in the embedding space. We define similar/positive labels by ranking the ECGs in a batch using the dynamic time warping distance between two ECGs which represents the inherent characteristics of ECG data. We finetune the model from upstream DML on downstream task to esimate intracardiac pressures. Our preliminary results showed improved performance compared to the supervised baseline, which shows promise that utilization of unlabeled dataset helps improve generation of embedding with useful information for downstream task.

2 Related Works

2.1 Deep Metric Learning

Deep Metric Learning (DML) is a similarity driven learning method for building a representation by learning a distance metric. DML works by optimizing the loss function defined on pairs or tuples generated from training examples, where the objectives are often distance metric on embedding space. One of the example distance metrics used for the objectives are ranking-based contrastive training, including a triplet loss function [13, 14] and a pairwise loss funciton [15]. DML has been demonstrated to improve generalization performance in representation learning [8, 9, 10]. However, DML has been applied mostly in the supervised setting, and only few recent works proposed ways for establishing DML in an unsupervised manner [16, 17, 18]. Furthermore, as DML generally requires class of labels to form the tuplets of data, only a few studies have applied DML to regression problems [19]. Here we apply unsupervised DML to a real-world regression problem to determine if similarity

information within ECG dataset helps create an embedding space that improves the performance of downstream hemodynamics inference.

2.2 Self-Supervised Contrastive Learning for Electrocardiogram

Self-supervised contrastive learning performs pre-training with pretext tasks to generate a global embedding over an unlabeled dataset, by modeling consistency while ensuring invariance to perturbations. There have been several attempts to leverage unlabeled ECG dataset using self-supervised representation learning for downstream ECG classification task [3, 4, 5, 6, 7, 20] and inferrence in biosignals [21]. However, to the best of our knowledge, unsupervised DML has not been applied to downstream tasks involving ECGs. In this work, we apply DML on unlabeled ECG for downstream hemodynamics inference task and compare the test performance with supervised learning.

3 Methods

At the core of our project is utilizing Deep Metric Learning to build a better representation by leveraging rich information from unlabeled ECG data. We build representation space by pulling similar samples into similar metric spaces and pushing different samples apart.

3.1 Supervised Deep Metric Learning for Classification

Given a n labeled ECG dataset $D = f(x_1; y_1); \quad ; (x_n; y_n)g$, where $x \; 2 \; \mathbf{R}^{d \quad T}$ and $y \; 2 \; f0; 1g$ with d being the number of ECG leads and T is the number of ECG timestep. Binary label y describes whether a patient has elevated mPCWP or not, which is defined by the cutoff mPCWP threshold 15 mmHg (1 if the patient has elevated mPCWP ($>$ 15 mmHg). Per each anchor ECG x_a, we define the positive sample x_p to be the ECGs having the same label with the anchor ECG and the negative sample x_n is defined to be the one that has different label. The triplet $(x_a; x_p; x_n)$ are then sampled from the batch B to calculate the triplet loss $L_{triplet}$. The model f that has D-dimensional output is optimized with triplet loss ($L_{Triplet}$) and f with the last fully-connected layer g. Thus, $g \quad f$ is optimized by the learning objective combining cross-entropy loss (L_{CE}) for a batch size of N with the triplet loss scaling factor of .

$$
\begin{aligned}
L_{tot} &= L_{CE} + \quad L_{triplet} \\
&= \sum_{i=1}^{N} y_i log(\quad (g(f \quad (x_i))) + \sum_{i=1}^{N} fjjf \quad (x_a) \quad f \quad (x_p)jj \quad jjf \quad (x_a) \quad f \quad (x_n)jjg
\end{aligned}
$$

where is the scaling coefficient of triplet loss and is the sigmoid function.

3.2 Supervised Deep Metric Learning for Regression

Similar to the case of supervised DML classification task, where we have N labeled ECG dataset $D = f(x_1; y_1); \quad ; (x_N; y_N)g$, where $x \; 2 \; \mathbf{R}^{d \quad T}$ and $y \; 2 \; \mathbf{R}$ with d ECG leads. We define the positive sample x_p to be the input sample with the smallest absolute label difference with the anchor: $x_p = fx_i ji = \arg\min_{y_i 2 B} jy_i \quad y_a jg$ where B is the batch. Then the negative sample x_n is then randomly sampled from the set of ECGs in a batch except the positive sample: $N_B = fx_j \; 2 \; Bjx_j \; 6= \; x_p g$. Again, we construct the triplet $(x_a; x_p; x_n)$ and let the model f be optimized by triplet loss ($L_{Triplet}$) while at the same time the final classifier $g \quad f$ is jointly optimized with Root Mean Square Error (RMSE) with the triplet scaling factor of : $L_{tot} = L_{RMSE} + \quad L_{triplet}$. See Appendix Section A.1 for the full loss function.

3.3 Unsupervised Deep Metric Learning

Now we expand this setting to unsupervised metric learning where we nave M ECG data $D = fx_1; \quad ; x_M g$, where $x \; 2 \; \mathbf{R}^d$ without the matched mPCWP labels. 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. Here we measure the distance between two inputs using distance measures suitable for signal input and define the similarity between inputs, dynamic

Table 1: Performance of downstream mPCWP classification task with unsupervised DML pretraining according to each embedding dimensions (Dim).

		Classification			Regression
	Dim	AUC	APR	ACC	RMSE
Supervised	-	0.75 ± 0.00	0.54 ± 0.00	$\mathbf{0.72 \pm 0.00}$	7.78 ± 0.1
Supervised DML	128	$\mathbf{0.78 \pm 0.00}$	$\mathbf{0.69 \pm 0.01}$	0.69 ± 0.00	7.65 ± 0.02
	256	0.75 ± 0.00	0.67 ± 0.00	0.70 ± 0.00	6.97 ± 0.01
	512	0.77 ± 0.00	0.69 ± 0.01	0.71 ± 0.00	7.42 ± 0.00
	1024	$\mathbf{0.78 \pm 0.00}$	0.68 ± 0.00	0.71 ± 0.01	7.5 ± 0.00
Unsupervised DML (DTW)	128	0.75 ± 0.00	0.53 ± 0.00	$\mathbf{0.72 \pm 0.01}$	7.28 ± 0.00
	256	0.74 ± 0.00	0.52 ± 0.00	$\mathbf{0.72 \pm 0.00}$	7.16 ± 0.02
	512	0.76 ± 0.00	0.53 ± 0.00	$\mathbf{0.72 \pm 0.00}$	$\mathbf{7.98 \pm 0.01}$
	1024	0.74 ± 0.00	0.53 ± 0.01	0.71 ± 0.00	7.48 ± 0.00

time warping (DTW) [22]: d_{DTW}. As the complexity of DTW calculation is quadratic ($O(n^2)$), we use pretrained DTW model to get the surrogate DTW ($\hat{d}_{DTW}$, See Appendix B for more detail). We can then define the positive sample to be $x_p = \{x_i | i = \arg\min_{x_i \in B} \hat{d}_{DTW}(x_i, x_a)\}$. Then the negative sample x_n is randomly sampled from the set containing dataset from the batch excluding the positive sample: $N_B = \{x_j \in B | x_j \neq x_p\}$. The model f is optimized with the upstream deep metric learning with $L_{triplet}$. The input triplets (x_a, x_p, x_n) are mapped on D-dimensional representation space , where we finetune the model with the downstream finetuing task (See Appendix A.2, Figure 1).

3.4 Datasets, Tasks, and Model

Dataset with mPCWP Labels (labeled dataset) We construct the 12 lead ECG dataset with corresponding right heart catheterization procedure from the data warehouse of Massachusetts General Brigham where we have total of $8,172$ right heart catheterization procedure data and ECGs from $4,051$ patients. Our data includes the patients who had ECG recordings the same date they had right heart catheterization from 2010 to 2020. The ECG data contains 10 seconds of signal data which was sampled with 250 Hz sampling rate, and the mPCWP label was obtained from the right heart catheterization procedure. We divided the dataset into train, valid, and test set which contains 4,680, 1,667, 1,825 data points and 2,430, 810, and 811 patients, respectively.

Dataset without mPCWP Labels The larger 12 lead ECG dataset without mPCWP mapping has provided by the agreement from Massachusetts General Hospital and Brigham and Women's Hospital where we have total of $5,426,614$ ECGs with $1,195,268$ patients who had ECG recordings while their daytime visit or admission. We removed patient cohort 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.

Task We train a model to predict the mPCWP of patient in either classification and regression task. Binary classification task is to infer the elevation of mPCWP with the threshold of 15 mmHg, and the regression task would be to predict the value of mPCWP itself.

Model For the basic model architecture, we used ResNet18 [23] backbone as encoder for generating D-dimensional space (metric learning experiments), classifier in combination with the last fully connected layer with sigmoid function (supervised learning, downstream finetuning). See Appendix C for more detail.

4 Results and Future Works

In this work, we have demonstrated with our baseline experiment that unsupervised DML improves the performance of downstream wedge pressure prediction. We summarize the test performance (Supervised baseline, Supervised DML) and downstream performance on hemodynamics inference (Unsupervised DML) in Table 1. The results show that supervised and unsupervised DML helps

improve the performance of the supervised baseline, specifically more in regression task. This implies that the underlying similarity information in the ECG signal pattern captured by DTW metric helps inferring the downstream wedge pressure in both classification and regression settings. Although we obtained the RMSE value of 6.9 mmHg in supervised DML and 7.16 mmHg in unsupervised DML, the regressinon error is not sufficiently small to be applied in clinical ward given that the mean mPCWP measure of our dataset is 15.6 mmHg and standard deviation is 8.07 mmHg.

In the future, we will generalize our experimental setting to the full 5 million ECG dataset we have in the dataset warehouse, and expect the deep metric embedding achieves better performance as it builds embedding with richer similarity based information. Furthermore, we will add more baselines onto our dataset to compare the result with contrastive learning settings. One additional baseline and direction we can adopt in our future work would be to apply self-training with noisy student [24], where we generate pseudolabels from unlabeled input ECG with learned model and keep validate with supervised test set. We will release our final code base and model weights trained on our ECG datasets upon release of our final paper.

5 Acknowledgements

We would like to thank members of the Computational Cardiovascular Research Group and HealthyML Lab for their invaluable feedback.

References

[1] AP Davie, CM Francis, MP Love, L Caruana, IR Starkey, TRD Shaw, GR Sutherland, and JJV McMurray. Value of the electrocardiogram in identifying heart failure due to left ventricular systolic dysfunction. *British Medical Journal*, 312(7025):222–223, 1996.

[2] 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.

[3] Nathaniel Diamant, Erik Reinertsen, Steven Song, Aaron Aguirre, Collin Stultz, and Puneet Batra. Patient contrastive learning: a performant, expressive, and practical approach to ecg modeling. *arXiv preprint arXiv:2104.04569*, 2021.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

[11] Fatih Cakir, Kun He, Xide Xia, Brian Kulis, and Stan Sclaroff. Deep metric learning to rank. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1861–1870, 2019.

[12] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017.

[13] 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.

[14] 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.

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

[16] 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, 2020.

[17] 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.

[18] 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, 2020.

[19] 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.

[20] 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.

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

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

[23] 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.

[24] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020.

A Deep Metric Learning

A.1 Supervised Deep Metric Learning for Regression

The full loss function for supervised DML regression task uses the Root Mean Square Loss (RMSE) for downstream classification loss, $L_{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} ||y_i - f(x_i)||^2}{n}}$ where the equation for the full loss

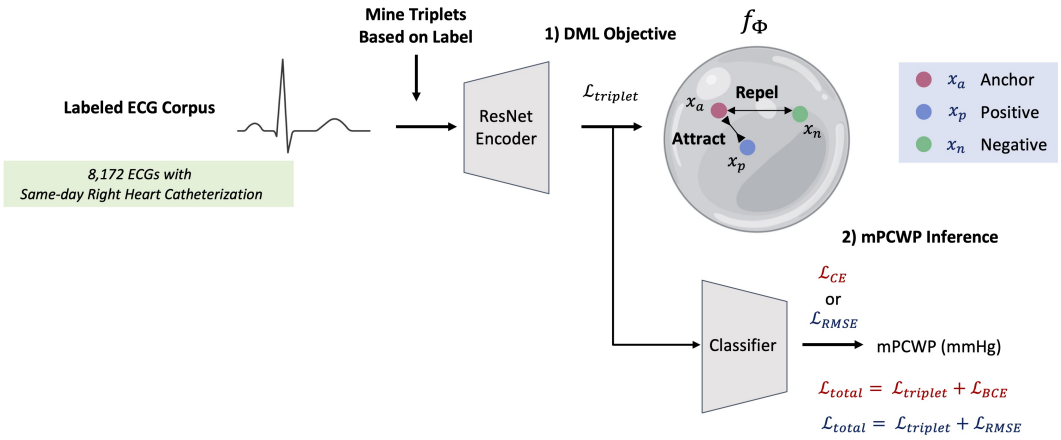

Figure 2: Supervised Deep Metric Learning with joint learning of deep metric learning objective, triplet loss and the loss function for mPCWP inference task. The triplets are mined based on the label for classification task, and based on the absolute label difference for regression task (see Section 3.2)

term is:

$$L_{tot} = L_{RMSE} + L_{triplet} = \sqrt{\frac{\sum_{i=1}^{n} \|y_i - f(x_i)\|^2}{n}} + \sum_{i=1}^{N} \max\left(\|f(x_a) - f(x_p)\| - \|f(x_a) - f(x_p)\| + \alpha, 0\right)$$

A.2 Downstream Finetuning

We finetune the model f with the downstream finetuning task with either supervised classification (L_{CE}) or regression task (L_{RMSE}) on the ECG dataset $D = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ introduced in the Sections 3.1 and 3.2: $L_{CE} = -\sum_{i=1}^{N} y_i \log(\sigma(f(x_i)))$, $L_{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} \|y_i - g(f(x_i))\|^2}{n}}$

B Dynamic Time Warping for similarity-based ECG Ranking

DTW calculates the amount of dissimilarity between two waveforms in terms of distance (e.g. Euclidean) under admissible temporal alignments of those waveforms. The algorithm attempts to minimize the distance for finding the optimal alignment (warping). Using discrete matching between existing elements of the waveform, the admissible warping paths are defined; DTW finds the best warping path that minimizes the overall distance using dynamic programming. This recursive approach is computationally expensive as it costs near $O(n^2)$.

To use DTW as our ranking metric for the ECGs, we calculate the DTW (d_{DTW}) between respective leads of two ECGs and take the average metric over the 12 leads. This approach adds a large computational burden for each batch processing, given the complexity of the DTW calculation. Hence, we design a surrogate model (Figure 3) to calculate the d_{DTW} without using the dynamic programming-based recursive algorithm, as described below.

B.1 Surrogate Model to infer DTW between two ECGs

To bypass the cost of calculating the DTW which is up to $O(n^2)$, we build a surrogate model to get the surrogate DTW value $\hat{d}_{DTW}$, which is the prediction output of the model. We build the surrogate model with ResNet18 encoders and two fully connected layers (FC layer) (Figure 3). We first embed two ECGs using the ResNet18 encoder to get 50 dimension output, respectively. Then we concat the two embeddings to generate 100-dimensional input to two FC layers, getting the final output. The