MODALLY REDUCED REPRESENTATION LEARNING OF MULTI-LEAD ECG SIGNALS THROUGH SIMULTANEOUS ALIGNMENT AND RECONSTRUCTION

Nabil Ibtehaz *
Department of Computer Science
Purdue University
West Lafayette, IN, USA
nibtehaz@purdue.edu

Masood Mortazavi
IC Lab
Futurewei Technologies Inc.
Santa Clara, CA, USA
masood.mortazavi@futurewei.com

ABSTRACT

Electrocardiogram (ECG) signals, profiling the electrical activities of the heart, are used for a plethora of diagnostic applications. However, ECG systems require multiple leads or channels of signals to capture the complete view of the cardiac system, which limits their application in smartwatches and wearables. In this work, we propose a modally reduced representation learning method for ECG signals that is capable of generating channel-agnostic, unified representations for ECG signals. Through joint optimization of reconstruction and alignment, we ensure that the embeddings of the different channels contain an amalgamation of the overall information across channels while also retaining their specific information. On an independent test dataset, we generated highly correlated channel embeddings from different ECG channels, leading to a moderate approximation of the 12-lead signals from a single-channel embedding. Our generated embeddings can work as competent features for ECG signals for downstream tasks.

1 INTRODUCTION

Electrocardiogram (ECG) signals are used in the diagnosis of cardiovascular diseases—the leading cause of death worldwide, claiming a life every 33 seconds in the USA [1]. ECG signals represent the electrical activities of a heart during the cardiac cycle Hannun et al. (2019). Utilizing ECG signals to their fullest potential requires conventional 12-lead ECG recorder H. Bunce et al. (2020). This limits the use cases of near-clinical ECG in wearables and smartwatches. Hence, the growing interest in developing competent diagnosis algorithms with a limited set of channelsReyna et al. (2021), which introduces new levels of complexity and algorithmic challenges Reyna et al. (2022).

Oriented to specific applications, ECG investigations have used a variety of feature extraction methods, e.g., transformations Mohonta et al. (2022), decompositions Ibtehaz et al. (2019), fiducial point features Ullah et al. (2023), etc. However, features competent for one application seldom translates well to a different one Singh & Krishnan (2023). Observing the impact of representation learning in a multitude of domains Bengio et al. (2013), it becomes desirable to develop general representation learners for ECG Mehari & Strodthoff (2022), most notably using masked autoencoders He et al. (2021), which have proven to be quite competent Sawano et al. (2022); Yang et al. (2022); Zhang et al. (2023); Wang et al. (2023a). However, such models were assessed on a specific task Sarkar & Etemad (2022) and sometimes evaluated on only a portion of the pretraining dataset Sawano et al. (2022) while requiring all 12 channels Zhang et al. (2023) for satisfactory performance.

While these approaches partially address the issue of task-specific feature extraction, the requirement for multiple ECG leads persists. To explore mode (ECG lead) reduction, one can adopt the perspective of multi-modal learning. Meta-learning has been used to reconstruct missing modes from available ones Ma et al. (2021), and other solutions have relied on a shared embedding to

*Work done as a summer intern at Futurewei Technologies.

[1] National Center for Health Statistics. Multiple Cause of Death 2018–2021 on CDC WONDER Database. Accessed February 2, 2023.

achieve this Wang et al. (2023b). Semantic alignment of heterogeneous data modes in a shared vector space has also been used through joint learning of co-embeddings Mortazavi (2020). ECG signals represent a special case of homogeneous multi-modal data, i.e., cardiac electric activity of the same heart from different viewpoints (leads). As a result, some correlations already exist between the different modalities, i.e., channels of the ECG signal; for instance, channel II can be treated as the sum of channels I and III based on Einthoven's law Silverman & Willis Hurst (1992).

In this work, we propose a unified ECG representation learning method that is channel-independent based on simultaneous alignment and reconstruction. We achieve this by representing the different ECG channels in a joint embedding space, maintaining semantic alignment through contrastive learning while at the same time using an auto-encoder-based reconstruction objective so that the essential channel-specific information can survive in the alignment process.

2 METHOD

Abstractions and High-level Assumptions. As activity snapshots, the 12-lead ECG signal $x(t) = \{x^{[1]}(t), x^{[2]}(t), x^{[3]}(t), \ldots, x^{[12]}(t)\} \in \mathbb{R}^{12}$ represent the state of a heart $\mathcal{H}$ at an instant t, where each channel $x^{[i]}(t)$ corresponds to a projection P_i, i.e., $x^{[i]}(t) = P_i(\mathcal{H}_{@t})$. We train encoder networks $\mathcal{E}_i, i = 1 \ldots 12$ to extract latent representations $h^{[i]}(t)$ of a heart from the input signals $x^{[i]}(t)$. The representations of the different channel signals of the *same* heart should be correlated, i.e., $h^{[i]}(t) \equiv \mathcal{E}_i(x^{[i]}(t)) \approx \mathcal{H}_{@t}, i = 1 \ldots 12$ and should also be "complete" in the sense that individual signals can be retrieved from them by a set of decoders, i.e., $\mathcal{D}_i, i = 1 \ldots 12$, i.e., $\hat{x}^{[i]}(t) \equiv \mathcal{D}_i(h^{[i]}(t)) \approx x^{[i]}(t)$.

Core Model Architecture. For our representation learner ($\mathcal{E}_i - \mathcal{D}_i$), we have developed a 1-D version of the original Masked AutoEncoder (MAE) architecture He et al. (2021). Our architecture follows the conventional style of having a strong encoder in conjunction with a weak decoder, and vanilla ViT layers Dosovitskiy et al. (2021) are used to as the building blocks. Our encoder and decoder have 12 and 8 blocks, with embedding dimension of 768 and 512, respectively. The model takes $T = 5$ sec, $f_s = 500$ Hz signals as input and breaks them into 100 samples or 0.2 sec long non-overlapping patches, for further processing.

Alignment and Reconstruction Losses. We distribute the task of 12-lead ECG representation learning to 12 1-D MAE models each assigned to a particular channel. Our models encode-decode to reconstruct their designated channels while producing correlated embeddings as approximation of the latent heart state by jointly optimizing on alignment and reconstruction losses. Expected mean-squared error (MSE) of the predictions $\hat{x}^{[i]}(t)$ over a minibatch B provides a reconstruction loss:

$$\text{reconstruction loss} = \mathbb{E}_B[\frac{1}{12 \times T} \sum_{i=1}^{12} \sum_{t=0}^{T} (x^{[i]}(t) - \hat{x}^{[i]}(t))^2] \tag{1}$$

Triplet loss helps correlate the $h^{[i]}(t), i = 1 \ldots 12$ of different channel signals of the same heart:

$$\text{alignment loss} = Triplet_B(A, P, N) \tag{2}$$

Here, anchor A corresponds to all the signals in the minibatch, positive P and negative N correspond to signals of different channels from the same vs. different hearts, respectively.

Combining Losses. Jointly optimizing these losses often led to trivial embeddings—highly correlated but unable to reconstruct the signals. To prevent this, we designed a curriculum which gradually shifts focus from reconstruction to alignment losses:

$$Loss_{@epoch-i} = sin(\frac{i}{N_{epochs}} \frac{\pi}{2}) \times \text{alignment loss} + cos(\frac{i}{N_{epochs}} \frac{\pi}{2}) \times \text{reconstruction loss} \tag{3}$$

Pretraining Dataset. As a large 12-lead ECG signal database for pretraining, we selected the PhysioNet Computing in Cardiology Challenge 2021 dataset Reyna et al. (2021); Perez Alday et al. (2020); Goldberger et al. (2000), which includes 88,000 public ECG recordings compiled from 6 sources. We only used ECG signals recorded at 500 Hz sampling frequency (f_s) to train our models. The recordings from PTB and INCART databases were used as test dataset. For training and

inference, a random $T_s = 5$ s long window was cropped from the signals, and mean normalization was performed without any additional filtering or preprocessing steps.

Distributed Training. We adopted a distributed training protocol where MAE models are trained separately on their individual channels, with a coordinator exchanging embeddings required for computing alignment losses. We used AdamW Loshchilov & Hutter (2019) with default parameters to jointly optimize reconstruction and alignment. The models were trained with a base leaning rate of $1e^{-3}$ accompanied by a cosine learning rate scheduler, for 200 epochs with a batch size of 256.

3 RESULTS

We assessed different aspects of our method and some downstream tasks on several independent test datasets varying in terms of acquisition protocol, demographics, signal quality, sampling rate, etc.

Correlating Embeddings of Dissimilar Signal Channels. Our primary objective is to develop a unified representation for ECG signals irrespective of the signal channel. In order to validate our joint optimization strategy, we took 590 signals from PTB R. Bousseljot et al. (1995) and INCART Tihonenko et al. (2008) databases, with sampling rates of 1000 and 257 Hz, respectively. For randomly cropped 5-second-long signals, we computed the similarity between signals of two channels and the corresponding embeddings using cross-correlation coefficient and cosine similarity, respectively. We performed this random experiment 10 times for each channel pair. The overall correlation, presented in Fig. 1B, reveals that our pipeline managed to generate strongly correlated embeddings, even when the signals are weakly correlated (e.g. V5-aVR). Conversely, for already correlated channel pairs, our embeddings either retained (e.g. V1-V2) or improved (e.g. V4-V5) the correlations.

Reconstructing Disparate Signals from Unified Embeddings. In order to assess our reconstruction performance, we have used 827 ECG recordings used in Ribeiro et al. (2020) ($f_s = 400$Hz). In Fig. 1C, we have presented an example of 2 sec reconstruction of a random sample. Since, our generated embeddings are correlated, we have tried to reconstruct all the channel signals both from the native channel and channel 1 signals individually, after masking 75% of the signals. Overall, the reconstructed signals follow the morphology of the original signal, the model failed to reach the full amplitude of the R-peaks, despite predicting a sharp peak. This is a limitation of MAE, which usually reconstructs blurry images He et al. (2021), and for ECG reconstruction, reduces the signal magnitude to a fraction Sawano et al. (2022). Additionally, we computed mean absolute error values of reconstructing each of the channel signals from the different input channels (Fig. 1D), and we can observe that due to the correlation of our generated embeddings, the reconstruction losses stay within $\pm 5\%$ of the loss for native channel as input. Interestingly, for V1, the native embeddings seem to perform worse, which is also apparent from Fig. 1C. This is likely due to the V1 channel signals of this dataset being substantially different from our pretraining dataset.

Improving Single-Channel Myocardial Infraction Prediction. The diagnosis of myocardial infraction, one important application of ECG signals, substantially benefits from multi-channel information Gupta et al. (2021). Therefore, we aim to investigate whether the proposed unification of the embeddings across channels can help improve single-channel predictions. In order to assess this, we trained baseline MAE models for the individual channels without any embedding alignment and performed a 5-fold cross-validation on the PTB dataset R. Bousseljot et al. (1995). Next, for each of the different channels, embeddings were extracted from the correlated and baseline MAE models and logistic regression classifiers were used. From the F1 scores presented in Fig. 1F, it can be observed that the correlated models performed better in all the cases. They not only managed to improve the less informative channels (I and aVL) but also retained the superior performance of the highly informative channels (V5 and V6) Gupta et al. (2021). Therefore, our proposed embedding correlation injects more information to less informative channels, without hampering the representations of the highly informative channels.

Implicitly Learning ECG Biometric Authentication. Since the alignment objective distinguishes and aligns ECG embeddings from an individual's heart, we used the ECG-ID benchmark dataset Lugovaya (2005); Nemirko & Lugovaya (2005) to assess model performance in biometric authentication. In order to capture the entire ECG signature, we avoided any masking during this inference and the class token embedding was obtained from the model. A t-SNE van der Maaten & Hinton (2008) plot presented in Fig 1E shows that the ECG embeddings of different individuals are clus-

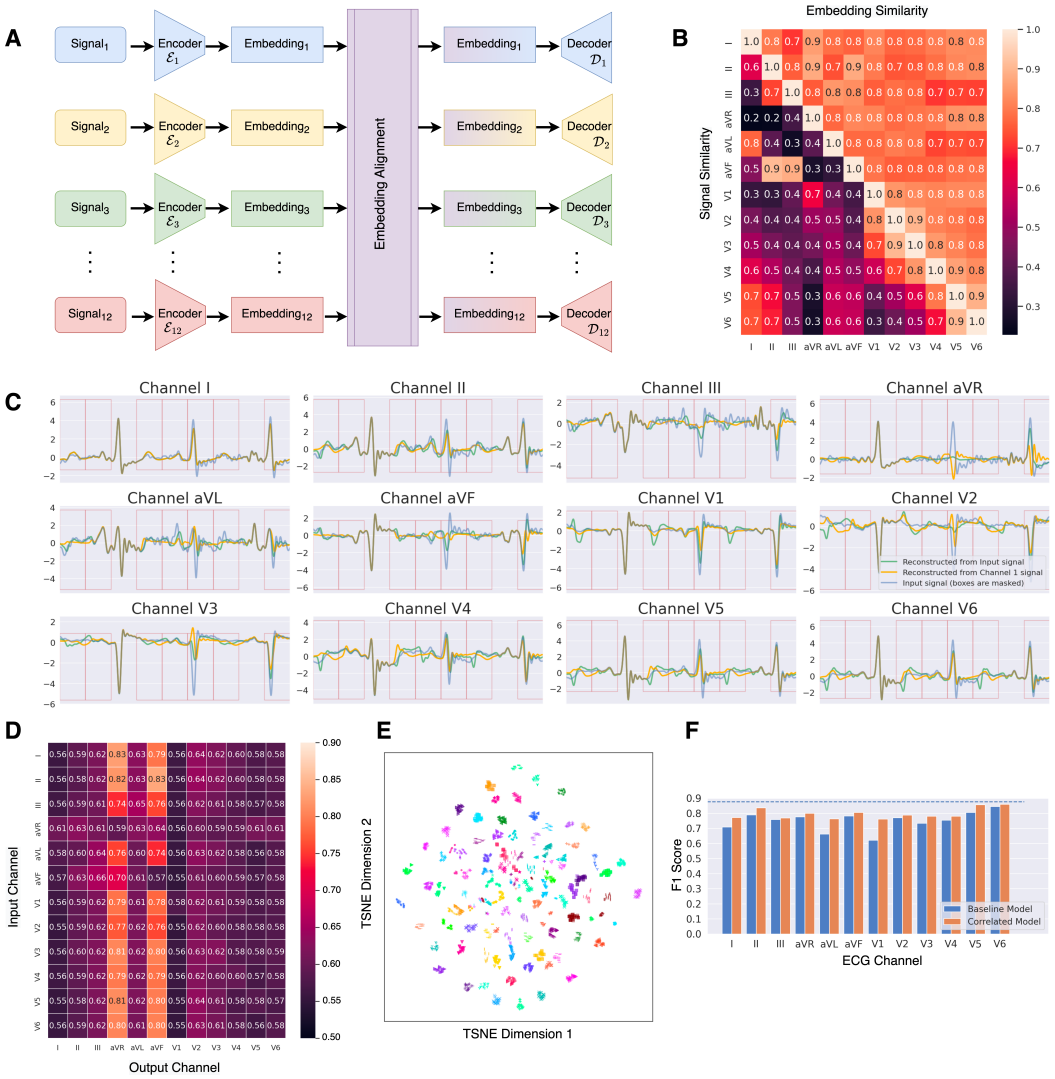

Figure 1: A) Our proposed architecture for correlating Masked AutoEncoders. (B) Signal (lower triangle) versus embedding similarity (upper triangle) for pairs of ECG channels on INCART and PTB datasets. (C) 2s ECG reconstruction of a sample from the test set: blue represents the original signal, while green and yellow correspond to reconstruction from the native versus channel 1 embedding, respectively. The red boxes denotes the masked windows. (D) Mean absolute error values of reconstructing each of the channel signals from different channels, computed on the test set. (E) t-SNE plot of the embeddings of the signals of different individuals from the ECG-ID dataset. (F) Improvement in the single-channel myocardial infarction diagnosis task on the PTB dataset.

tered together. Moreover, as a true assessment of the quality of our embeddings, we used a 1-KNN classifier and achieved an accuracy of 99.68% from 10-fold cross-validation. This is indeed impressive as without any finetuning, on an independent test dataset, we are on par with state-of-the-art methods Ibtehaz et al. (2022) which are explicitly trained on that dataset for this task.

4 Conclusion

In this work, we have explored the possibility of correlating the representations of the different ECG channels to pave the path to a competent unified embedding and evaluated it on several independent test datasets. Our current system fails to reconstruct of sharp changes, which can be improved by refinement through a generative adversarial network Fei et al. (2023). We believe our mode-reduction approach has the potential to deliver generalized, channel-agnostic ECG feature extractors. The broader impact of this approach is to improve the versatility of wearable and smartwatch-based ECG sensors towards medical-grade ECG sensors.

REFERENCES

Y. Bengio, A. Courville, and P. Vincent. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 8 2013. ISSN 0162-8828. doi: 10.1109/TPAMI.2013.50.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

Zhengcong Fei, Mingyuan Fan, Li Zhu, Junshi Huang, Xiaoming Wei, and Xiaolin Wei. Masked Auto-Encoders Meet Generative Adversarial Networks and Beyond. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24449–24459. IEEE, 6 2023. ISBN 979-8-3503-0129-8. doi: 10.1109/CVPR52729.2023.02342.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet. *Circulation*, 101(23), 6 2000. ISSN 0009-7322. doi: 10.1161/01.CIR.101.23.e215.

Arjun Gupta, Eliu Huerta, Zhizhen Zhao, and Issam Moussa. Deep Learning for Cardiologist-Level Myocardial Infarction Detection in Electrocardiograms. pp. 341–355. 2021. doi: 10.1007/978-3-030-64610-3{_}40.

Nicholas H. Bunce, Robin Ray, and Hitesh Patel. 30. Cardiology. In Adam Feather, David Randall, and Mona Waterhouse (eds.), *Kumar and Clark's Clinical Medicine*, pp. 1019–1133. Elsevier, 10th edition edition, 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, 1 2019. ISSN 1078-8956. doi: 10.1038/s41591-018-0268-3.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. 11 2021.

Nabil Ibtehaz, M. Saifur Rahman, and M. Sohel Rahman. VFPred: A fusion of signal processing and machine learning techniques in detecting ventricular fibrillation from ECG signals. *Biomedical Signal Processing and Control*, 49:349–359, 3 2019. ISSN 17468094. doi: 10.1016/j.bspc.2018.12.016.

Nabil Ibtehaz, Muhammad E. H. Chowdhury, Amith Khandakar, Serkan Kiranyaz, M. Sohel Rahman, Anas Tahir, Yazan Qiblawey, and Tawsifur Rahman. EDITH : ECG Biometrics Aided by Deep Learning for Reliable Individual Authentication. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(4):928–940, 8 2022. ISSN 2471-285X. doi: 10.1109/TETCI.2021.3131374.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=Bkg6RiCqY7`.

T S Lugovaya. Biometric human identification based on electrocardiogram. *Master's thesis, Faculty of Computing Technologies and Informatics, Electrotechnical University 'LETI', Saint-Petersburg, Russian Federation*, 2005.

Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. SMIL: Multimodal Learning with Severely Missing Modality. 3 2021.

Temesgen Mehari and Nils Strodthoff. Self-supervised representation learning from 12-lead ECG data. *Computers in Biology and Medicine*, 141:105114, 2 2022. ISSN 00104825. doi: 10.1016/j.compbiomed.2021.105114.

Shadhon Chandra Mohonta, Mohammod Abdul Motin, and Dinesh Kant Kumar. Electrocardiogram based arrhythmia classification using wavelet transform with deep learning model. *Sensing and Bio-Sensing Research*, 37:100502, 8 2022. ISSN 22141804. doi: 10.1016/j.sbsr.2022.100502.

Masood S. Mortazavi. Speech-Image Semantic Alignment Does Not Depend on Any Prior Classification Tasks. 10 2020. doi: 10.21437/Interspeech.2020.

A P Nemirko and T S Lugovaya. Biometric human identification based on electrocardiogram. In *Proceedings of the XIIIth Russian Conference on Mathematical Methods of Pattern Recognition, Moscow, Russian*, pp. 20–26, 2005.

Erick A Perez Alday, Annie Gu, Amit J Shah, Chad Robichaux, An-Kwok Ian Wong, Chengyu Liu, Feifei Liu, Ali Bahrami Rad, Andoni Elola, Salman Seyedi, Qiao Li, Ashish Sharma, Gari D Clifford, and Matthew A Reyna. Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020. *Physiological Measurement*, 41(12):124003, 12 2020. ISSN 0967-3334. doi: 10.1088/1361-6579/abc960.

R. Bousseljot, D. Kreiseler, and A. Schnabel. Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet. *Biomedizinische Technik / Biomedical Engineering*, 40(s1):317–318, 1995.

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, Hamid Ghanbari, Qiao Li, Ashish Sharma, and Gari D Clifford. Will Two Do? Varying Dimensions in Electrocardiography: The PhysioNet/Computing in Cardiology Challenge 2021. In *2021 Computing in Cardiology (CinC)*, pp. 1–4. IEEE, 9 2021. ISBN 978-1-6654-7916-5. doi: 10.23919/CinC53138.2021.9662687.

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, Hamid Ghanbari, Qiao Li, Ashish Sharma, and Gari D Clifford. Issues in the automated classification of multilead ecgs using heterogeneous labels and populations. *Physiological Measurement*, 43(8):084001, 8 2022. ISSN 0967-3334. doi: 10.1088/1361-6579/ac79fd.

Antônio H. Ribeiro, Manoel Horta Ribeiro, Gabriela M. M. Paixão, Derick M. Oliveira, Paulo R. Gomes, Jéssica A. Canazart, Milton P. S. Ferreira, Carl R. Andersson, Peter W. Macfarlane, Wagner Meira, Thomas B. Schön, and Antonio Luiz P. Ribeiro. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nature Communications*, 11(1):1760, 4 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-15432-4.

Pritam Sarkar and Ali Etemad. Self-Supervised ECG Representation Learning for Emotion Recognition. *IEEE Transactions on Affective Computing*, 13(3):1541–1554, 7 2022. ISSN 1949-3045. doi: 10.1109/TAFFC.2020.3014842.

Shinnosuke Sawano, Satoshi Kodera, Hirotoshi Takeuchi, Issei Sukeda, Susumu Katsushika, and Issei Komuro. Masked Autoencoder-Based Self-Supervised Learning for Electrocardiograms to Detect Left Ventricular Systolic Dysfunction. In *NeurIPS 2022 Workshop on Learning from Time Series for Health*, 2022. URL `https://openreview.net/forum?id=gz7c2HIcub7`.

Mark E. Silverman and J. Willis Hurst. Willem einthoven—the father of electrocardiography. *Clinical Cardiology*, 15(10):785–787, 10 1992. ISSN 0160-9289. doi: 10.1002/clc.4960151020.

Anupreet Kaur Singh and Sridhar Krishnan. ECG signal feature extraction trends in methods and applications. *BioMedical Engineering OnLine*, 22(1):22, 3 2023. ISSN 1475-925X. doi: 10.1186/s12938-023-01075-1.

Vikto Tihonenko, A Khaustov, S Ivanov, A Rivin, and E Yakushenko. St Petersburg INCART 12-lead arrhythmia database. *PhysioBank PhysioToolkit and PhysioNet*, 2008.

Nadeem Ullah, Tahir Mahmood, Seung Gu Kim, Se Hyun Nam, Haseeb Sultan, and Kang Ryoung Park. DCDA-Net: Dual-convolutional dual-attention network for obstructive sleep apnea diagnosis from single-lead electrocardiograms. *Engineering Applications of Artificial Intelligence*, 123: 106451, 8 2023. ISSN 09521976. doi: 10.1016/j.engappai.2023.106451.

Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

Guoxin Wang, Qingyuan Wang, Ganesh Neelakanta Iyer, Avishek Nag, and Deepu John. Unsupervised Pre-Training Using Masked Autoencoders for ECG Analysis. In *2023 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pp. 1–5, 2023a.

Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal Learning with Missing Modality via Shared-Specific Feature Modelling. 7 2023b.

Shunxiang Yang, Cheng Lian, and Zhigang Zeng. Masked Autoencoder for ECG Representation Learning. In *2022 12th International Conference on Information Science and Technology (ICIST)*, pp. 95–98. IEEE, 10 2022. ISBN 978-1-6654-8582-1. doi: 10.1109/ICIST55546.2022.9926900.

Huaicheng Zhang, Wenhan Liu, Jiguang Shi, Sheng Chang, Hao Wang, Jin He, and Qijun Huang. MaeFE: Masked Autoencoders Family of Electrocardiogram for Self-Supervised Pretraining and Transfer Learning. *IEEE Transactions on Instrumentation and Measurement*, 72:1–15, 2023. ISSN 0018-9456. doi: 10.1109/TIM.2022.3228267.

A METHOD OVERVIEW

Our proposed method aims at correlating multiple representation learners, each handling a different modality.

Algorithm 1 Modally Reduced ECG Representation Learning Pipeline

Encoder : $\mathcal{E}_i, i = 1, 2, 3, \ldots, 12$
Decoder : $\mathcal{D}_i, i = 1, 2, 3, \ldots, 12$
Input signal : $x[i], i = 1, 2, 3, \ldots, 12$

for i=1 to 12 **do**
  $h[i] = \mathcal{E}_i(x[i])$
  $\hat{x}[i] = \mathcal{D}_i(h[i])$
end for
for i=1 to 12 **do**
  $reconstruction_loss+ = MSE(x[i], \hat{x}[i])$
end for
$alignment_loss = Triplet(x[i], x[j], x'[k])$

B ARCHITECTURAL DETAILS

Our developed Masked AutoEncoder architecture is a 1-D variant of the original Masked AutoEncoder. Since we are working with 1D signals instead of 2D images, we have modified the patch encoder layer using 1D convolutions, discarding the overlapping nature, i.e., $stride = window_len$. After the patches or windows are embedded, we use conventional sine-cosine position embedding to preserve the temporal information.

The patches equipped with temporal identity are then passed through the encoder network, which is a sequence of ViT layers. The encoder can be summarized as:

$$\begin{aligned} Encoder = [ViT_Layer(&embed_dim = 768, \\ &num_heads = 8, \\ &mlp_ratio = 4, \\ &qkv_bias = True, \\ &norm_layer = LayerNorm)]_{\times 12} \end{aligned}$$

The decoder network is similar to the encoder network, only difference being reduced complexity in terms of depth and embedding dimension.

$$\begin{aligned} Decoder = [ViT_Layer(&embed_dim = 512, \\ &num_heads = 16, \\ &mlp_ratio = 4, \\ &qkv_bias = True, \\ &norm_layer = LayerNorm)]_{\times 8} \end{aligned}$$

C DATASET DESCRIPTION

The pretraining data contains twelve-lead ECGs from various sources. The twelve leads are the standard ones, i.e., I, II, III, aVR, aVL, aVF, V1, V2, V3, V4, V5, V6. These signals also contain information related to cardiac abnormalities, but they were ignored in the pretraining.

The sources of the pretraining database are:

- CPSC Database and CPSC-Extra Database
- INCART Database
- PTB and PTB-XL Database
- The Georgia 12-lead ECG Challenge (G12EC) Database
- Chapman-Shaoxing and Ningbo Database
- The University of Michigan (UMich) Database
- Augmented Undisclosed Database

These databases contain more than 100,000 signals, and out of them 88,000 signals are considered as the public training data, which is the split we also used for pretraining. These signals comprise individuals from four countries across three continents.

We used the signals with sampling frequency of 500 Hz for our pretraining. INCART and PTB consists of signals of different sampling frequency which were used for testing.