

Enhancing Cardiovascular Disease Prediction through Multi-Modal Self-Supervised Learning

Francesco Girlanda¹

fgirlanda@student.ethz.ch

Olga Demler^{1,2}

odemler@bwh.harvard.edu

Bjoern Menze³

bjoern.menze@uzh.ch

Neda Davoudi^{1,3,4}

neda.davoudi@ai.ethz.ch

¹ Department of Computer Science

ETH Zürich

Zürich, Switzerland

² Brigham and Women's Hospital

Harvard Medical School

Boston, Massachusetts, USA

³ Department of Quantitative Biomedicine

University of Zürich

Zürich, Switzerland

⁴ ETH AI Center

ETH Zürich

Zürich, Switzerland

Abstract

Accurate prediction of cardiovascular diseases remains imperative for early diagnosis and intervention, necessitating robust and precise predictive models. Recently, there has been a growing interest in multi-modal learning for uncovering novel insights not available through uni-modal datasets alone. By combining cardiac magnetic resonance images, electrocardiogram signals, and available medical information, our approach enables the capture of holistic status about individuals' cardiovascular health by leveraging shared information across modalities. Integrating information from multiple modalities and benefiting from self-supervised learning techniques, our model provides a comprehensive framework for enhancing cardiovascular disease prediction with limited annotated datasets.

We employ a masked autoencoder to pre-train the electrocardiogram ECG encoder, enabling it to extract relevant features from raw electrocardiogram data, and an image encoder to extract relevant features from cardiac magnetic resonance images. Subsequently, we utilize a multi-modal contrastive learning objective to transfer knowledge from expensive and complex modality, cardiac magnetic resonance image, to cheap and simple modalities such as electrocardiograms and medical information. Finally, we fine-tune the pre-trained encoders on specific predictive tasks, such as myocardial infarction. Our proposed method enhanced the image information by leveraging different available modalities and outperformed the supervised approach by 7.6% in balanced accuracy.

1 Appendix

1.1 Experimental setup

1.1.1 MAE

The specific ViT is equipped with 3 layers, and 6 projection heads that results in an embedding of size 384. We used a patch size of (1,100). Mean Squared Error (MSE) and the Normalized Correlation Coefficient (NCC) are used to evaluate the reconstruction performance where λ_{MAE} is a parameter that weights the two components of the loss:

$$\mathcal{L}_{MAE} = (1 - \lambda_{MAE})\mathcal{L}_{MSE} + \lambda_{MAE}\mathcal{L}_{NCC} \quad (1)$$

Signal augmentations include random cropping at a ratio of 0.5, Fourier transform surrogate augmentation [24] with a phase noise magnitude of 0.1, Gaussian noise with a sigma of 0.25, and a rescaling factor of 0.5. We trained the masked autoencoder using AdamW optimizer [25] with weight decay of 0.15, batch size of 128, base learning rate of 1e-5, and cosine annealing scheduler over 400 epochs with 10% of warm-up epochs and a value of $\lambda_{MAE} = 0.1$.

1.1.2 Image Encoder

We used different augmentation techniques such as random horizontal flips with a probability of 50%, random rotations up to 45 degrees in the image, adding color noise to brightness, contrast, saturation, and a random resized crop of the image.

We trained the image encoder using the AdamW optimizer [25] with weight decay of 1e-4, batch size of 512, base learning rate of 1e-4, and cosine annealing scheduler over 500 epochs with 10 warm-up epochs and the temperature parameter τ to 0.1.

1.1.3 Multimodal SSL

We trained the multimodal step using the AdamW optimizer [25] with weight decay of 1e-4, batch size of 256, base learning rate of 1e-4, cosine annealing scheduler over 200 epochs with 10% of warm-up epochs saving the checkpoint with the best loss. We set λ to 0.5 to balance the components of the loss function and the temperature parameter τ to 0.1. We used both the global pooling layer and the attention pooling to help with the next step. We found similar results with both of them during the grid-search to find the right hyperparameters.


1.1.4 Finetune

We trained the fine-tune step using the AdamW optimizer [25] with weight decay of 1e-4, batch size of 64, base learning rate of 1e-5. We tried both cosine annealing scheduler over 400 epochs with 5% of warm-up epochs, and a reduce LR on plateau scheduler, in the end we saw that the cosine annealing was the most consistent. We saved the best model for both schedulers according to the evaluation metric.

1.2 Extension to Stroke Disease

We tried our method for predicting another cardiovascular disease (stroke). Our approach capitalizes on the available labeled stroke data to effectively leverage the learned representations' discriminative power. We find that the predictive performance, while promising, is

Table 1: Comparison of different diagnostic modalities and training strategy for stroke prediction. Columns indicate which pre-train (pretr.)/training strategy is used. Best scores are in **BOLD** font. The second best is underlined. Our approach outperforms all baseline models with regard to AUC and balanced accuracy (Bal. Acc) metrics.

SSL Modality	MAE pretr.	Image pretr.	ECG Train	Tab. Train	CMRI Train	AUC [%]	Bal. Acc [%]
Tabular	✗	✗	✗	✓	✗	0.59	0.5
ECG	✗	✗	✓	✗	✗	0.57	0.52
ECG	✓	✗	✓	✗	✗	<u>0.63</u>	0.59
CMRI	✗	✗	✗	✗	✓	0.62	<u>0.60</u>
ECG (MMCL )	✓	✓	✓	✗	✓	0.61	0.57
ECG + Tabular (Ours)	✓	✓	✓	✓	✓	0.67	0.62

not as robust as observed for myocardial infarction (MI). This discrepancy can be attributed to the limited size of the stroke dataset, which comprises only about half the number of instances compared to MI. By exploiting the informative labels associated with stroke instances, our method surpasses alternative approaches, and supervised methods achieving AUC improvement of 9.8% and 21.8% respectively (table 1 and 2). This highlights the efficacy of our model in harnessing labeled data efficiently and underscores the quality of the trained embeddings, which proves useful even in scenarios with limited datasets, enhancing predictive performance for stroke disease prediction.

Table 2: Comparison between self-supervised and supervised techniques. Best scores are in **BOLD** font. The second best is underlined.

Modality	ECG	Tabular	AUC [%]	Balanced Acc [%]
Supervised NN	✗	✓	0.55	0.52
Supervised NN	✓	✗	0.55	0.53
Supervised NN	✓	✓	0.55	0.53
Ours	✓	✓	0.67	0.62

1.3 Tabular Data

In table 3, we present a comprehensive overview of the tabular data utilized in our study. This dataset encompasses a wide array of information, including demographics, comorbidities, and lifestyle factors. To handle missing tabular data, we imputed the categorical features with the most frequent ones and we used an iterative multivariate imputer for numerical features as a function of existing features over multiple imputation rounds. The last section also gives information about the cardiovascular diseases as labels that we want to predict. This tabular data will then be paired together with ECG and CMRI data so, as mentioned in Section 3, some tabular data could be missing.

Table 3: Tabular features for 45257 individuals

Variable	Units	Descriptor	Missing
Demographics			
Age	Years (SD)	64.6 (7.8)	0
Waist circumference	cm (SD)	88.7 (12.8)	0
Height	cm (SD)	170.1 (9.4)	0
Weight	Kg (SD)	76.1 (15.2)	0
BMI	Kg/m ² M (SD)	26.5 (4.4)	0
Sex	Female (%)	23375 (51.7)	0
Comorbidities			
Diabetes	Positive (%)	2542 (5.6)	134
Health rating	Good (%)	28532 (63.2)	86
Vascular heart problem	Positive (%)	2672 (5.9)	14
Stroke of father	Positive (%)	6345 (14)	0
Stroke of mother	Positive (%)	6389 (14.1)	0
Stroke of siblings	Positive (%)	1642 (3.6)	0
Breathe shortness	Yes (%)	3099 (6.9)	645
Anxiety visit	Yes (%)	13453 (29.9)	256
Chest pain	Yes (%)	4687 (10.4)	365
Stenosis	Positive (%)	131 (0.3)	0
Hypertension	Positive (%)	9798 (21.7)	0
Kidney disease	Positive (%)	1221 (2.7)	0
Dementia	Positive (%)	18 (0)	0
Thyrotoxicosis	Positive (%)	611 (1.4)	0
Migraine	Positive (%)	3344 (7.4)	0
Atrial fibrillation	Positive (%)	1427 (3.2)	0
Heart failure	Positive (%)	313 (0.7)	0
Embolism	Positive (%)	411 (0.9)	0
Deep-vein thrombosis	Years (SD)	720 (1.6)	35
Lifestyle			
Smoke	Smoker (%)	923 (2.0)	9
Alcohol intake	Three or four times a week. (%)	12715 (28.1)	19
Diet salt	Never/rarely (%)	25820 (57.1)	9
TV Time	hour/day (SD)	2.8 (1.6)	125
PC Time	hour/day (SD)	1.5 (1.5)	142
Physical activity	number of days/week (SD)	4.1 (2.2)	1053
Sleep duration	hours/day (SD)	7.2 (1.1)	128
Coffee intake	cups/day (SD)	2.0 (1.9)	28
Vascular Disease			
Stroke	Positive (%)	738 (1.6)	0
Myocardial infarction	Positive (%)	1399 (3.1)	0

References

- [1] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.
- [2] Justus TC Schwabedal, John C Snyder, Ayse Cakmak, Shamim Nemati, and Gari D Clifford. Addressing class imbalance in classification problems of noisy signals by using fourier transform surrogates. *arXiv preprint arXiv:1806.08675*, 2018.
- [3] Özgün Turgut, Philip Müller, Paul Hager, Suprosanna Shit, Sophie Starck, Martin J Menten, Eimo Martens, and Daniel Rueckert. Unlocking the diagnostic potential of ecg through knowledge transfer from cardiac mri. *arXiv preprint arXiv:2308.05764*, 2023.