# Multi-Modal Supervised Contrastive Learning for Myocardial Infarction Classification

Charlotte Richardson, Thalia Seale, Vicente Grau, Abhirup Banerjee

Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, OX3 7DQ, United Kingdom

## Abstract

*Accurate classification of myocardial infarction (MI) remains challenging reflecting its multifaceted pathology, which is not captured when electrocardiography (ECG) and cardiac magnetic resonance imaging (MRI) are considered in isolation. However, effectively integrating these modalities for MI classification remains an open challenge. In this work, we propose a multi-modal supervised contrastive learning framework for MI classification, combining 12-lead ECG signals with continuous three-dimensional (3D) cardiac motion derived from cine MRI. Modality-specific latent representations are learned independently and projected into a shared embedding space, where supervised contrastive learning is used to explicitly align ECG and MRI representations. We evaluate multiple fusion strategies, including latent concatenation, ensemble learning, and contrastive learning, on incident and prevalent MI (pMI and iMI, respectively) classification using UK Biobank data. Contrastive learning improved MI classification for both pMI and iMI, with F1 scores of 0.551 and 0.389 respectively. This is a 32.5% and 9.9% increase compared to ECG alone.*

## 1. Introduction

Cardiovascular disease (CVD) remains the leading cause of mortality worldwide, with ischaemic heart disease alone accounting for 108.7 deaths per 100,000 in 2021 [1]. Early and accurate identification of myocardial infarction (MI) is critical for effective treatment, yet remains challenging due to heterogeneity in both structural and electrophysiological manifestations.

Electrocardiography (ECG) is widely used for MI detection, with characteristic changes like ST-segment deviation and T-wave inversion. However, a substantial proportion of cases, particularly MI with non-obstructive coronary arteries (MINOCA), may present with normal ECG morphology, limiting diagnostic sensitivity [2]. Conversely, cardiac imaging provides structural and functional insight, but common metrics such as ejection fraction derive from discrete phases of the cardiac cycle and fail to capture the full spatiotemporal dynamics of cardiac motion [3].

Recent work has explored combining electrophysiological and anatomical information, for example by jointly modelling single-lead ECG signals with cardiac geometry at end-diastolic (ED) and end-systolic (ES) phases [4], or aligning 12-lead ECG and cardiac magnetic resonance imaging (MRI) using contrastive learning [5]. While these approaches demonstrate the potential of multi-modal learning, there remains a need for methods that learn task-specific, aligned representations across modalities for clinical classification.

In this work, we propose a novel multi-modal supervised contrastive learning framework for MI classification that integrates the 12-lead ECG with motion of the three-dimensional (3D) heart across a whole cardiac cycle. Specifically, we learn disentangled representations of 12-lead ECG signals using a $\beta$-total correlation variational autoencoder ($\beta$-TCVAE), and encode continuous 3D biventricular motion from cardiac cine MRI using geometric and temporal deep learning models. These modality-specific latent representations are projected into a shared embedding space using supervised contrastive learning, enabling explicit cross-modal alignment. A downstream classifier is then trained on the fused representation for both prevalent and incident MI classification (pMI and iMI, respectively). This approach leads to improvements of 6.9% and 12.3% in terms of AUC and 9.9% and 32.5% in terms of F1 for iMI and pMI, respectively, compared to ECG alone.

## 2. Methods

In this section, we present the multi-modal framework shown in Figure 1. We introduce the dataset, two modality-specific pipelines and subsequent latent fusion techniques. Most importantly, we introduce the contrastive learning approach, which drives the reported performance gains.

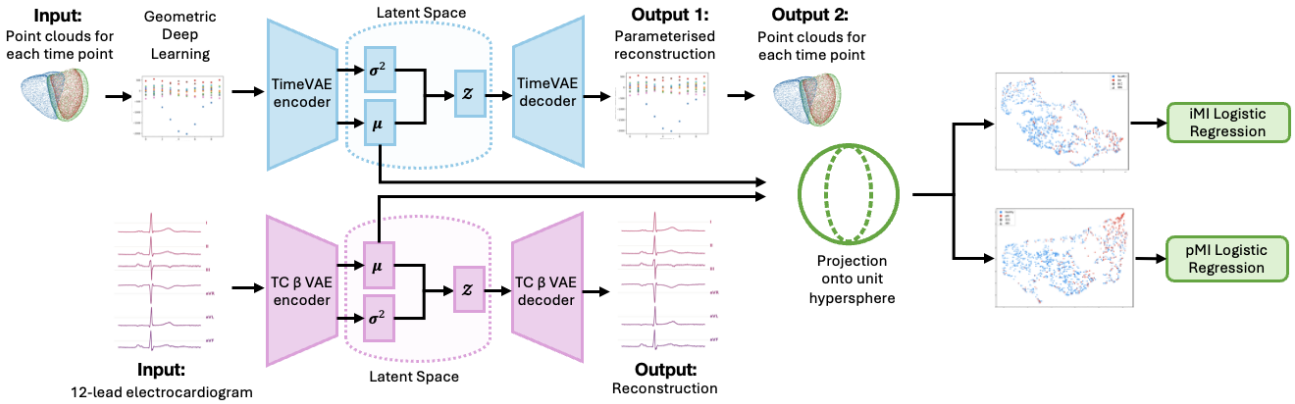


Figure 1. Architecture of the cross-modal supervised contrastive learning framework. Both $\mu$ representations are independently projected onto the unit hypersphere via a two-layer projection head (linear layer, layer normalisation, ReLU), producing normalised embeddings $e_i$ and $m_i$. The concatenated vector $\mathbf{h_i} = [\mathbf{e_i}|\mathbf{m_i}]$ is then passed to two separate logistic regression classifiers, trained independently for iMI vs healthy and pMI vs healthy classification.

## 2.1. Dataset

The dataset consisted of 719 subjects randomly chosen from the UK Biobank imaging study [6], excluding those with a history of CVD or obesity. Each subject had paired cardiac cine MRI and 12-lead resting ECG data.

3D biventricular point clouds (N=36,000) were generated from the cine MRI using [7,8]. For the 12-lead resting ECG recording, we used the median-beat representation of 1.2 seconds at 500Hz and was normalised per-lead. Signals with SNR $\leq$ 3dB were analysed individually, and 6 without recognisable P-QRS-T complexes were removed.

The data was split 64%, 16%, 20% for training, validation, and testing. 15.2% were labelled as iMI and have suffered from an MI since their initial screening visit. 21.7% are pMI and have had a myocardial infarction prior to their screening. In this work, we treat the two as separate binary classifications.

The ECG model was pretrained on the large Chapman-Shaoxing dataset [9] with over 10,646 patients.

## 2.2. Electrocardiogram Feature Extraction

All 12 ECG signals were concatenated into an input $x \in \mathbb{R}^{L\times T}$ where $L = 12$ denotes the number of leads and $T$ is the temporal dimension. Using a two-dimensional convolutional $\beta$ total correlation variational autoencoder ($\beta$-TCVAE) adapted from [10], all 12-leads were encoded jointly, allowing the model to capture both temporal and inter-lead patterns. Figure 2 demonstrates the architecture. The loss function is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{recon}} - \alpha \cdot \mathcal{L}_{\mathcal{MI}} + \beta \cdot \mathcal{L}_{\mathcal{TC}} + \gamma \cdot \mathcal{L}_{\mathcal{KL}},$$

where $\mathcal{L}_{\mathcal{MI}}$ represents mutual information, $\mathcal{L}_{\mathcal{TC}}$ enforces disentanglement via total correlation, and $\mathcal{L}_{\mathcal{KL}}$ is the dimension-wise Kullback–Leibler divergence term. The reconstruction loss $\mathcal{L}_{\text{recon}}$ was calculated using the mean squared error per lead.

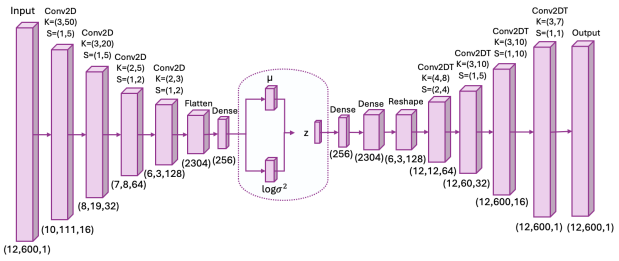


Figure 2. Architecture of the $\beta$-TCVAE, including kernel size ($K$), stride ($S$), and layer output dimensions.

## 2.3. 3D Cardiac Motion Feature Extraction

The architecture for the 3D cardiac motion is adopted from [11] and has two sequential steps. First, a $\beta$-VAE with a point completion network encoder is used to learn a compact latent representation of each cardiac point cloud frame independently. The second stage uses a TimeVAE that learns the dynamics of the shape representations over an entire cardiac cycle. The two models are trained separately, with the geometric model frozen during the temporal training.

## 2.4. Classification

We evaluate multiple fusion strategies for iMI and pMI classification, with supervised contrastive learning as the primary proposed approach. Single-modality results are

given as benchmarks, including the geometry of end diastolic and end systolic (ED+ES) phases of the cardiac cycle.

### 2.4.1. Contrastive Learning

To explicitly align representations across modalities, we propose a supervised contrastive learning (SCL) framework to merge the two latent spaces into a shared representation prior to classification, inspired by [12]. We extended the method to a multi-modal setting with positive pairs defined as subjects with the same label and *opposite* modality. Within-modality positive pairs would have allowed internal alignment, bypassing the cross-modal objective. The architecture is demonstrated in Fig. 1.

Both the ECG and MRI embeddings are independently projected via two linear feed-forward layers followed by layer normalisation, ReLU, and dropout, to a shared embedding space and $\ell_2$ normalised onto the unit hypersphere:

$$e_i = \frac{f_{\text{ecg}}(\mu_{\text{ecg},i})}{\|f_{\text{ecg}}(\mu_{\text{ecg},i})\|_2}, \qquad m_i = \frac{f_{\text{mri}}(\mu_{\text{mri},i})}{\|f_{\text{mri}}(\mu_{\text{mri},i})\|_2}.$$

The supervised contrastive loss iterates over the concatenated embedding set $\mathbf{z} = [\mathbf{e_1}, \ldots, \mathbf{e_N}, \mathbf{m_1}, \ldots, \mathbf{m_N}]$:

$$\mathcal{L}_{\text{con}} = \sum_{i \in \mathcal{I}} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\mathbf{z_i} \cdot \mathbf{z_p} \,/\, \tau)}{\sum_{a \in \mathcal{I}, a \neq i} \exp(\mathbf{z_i} \cdot \mathbf{z_a} \,/\, \tau)}$$

where $P(i)$ is the set of cross-modal positives for anchor $i$, $\mathcal{I} = \{1, \ldots, 2N\}$ is the index over all embeddings and $\tau$ is a temperature parameter, used to control the sharpness of the similarity distribution. $\mathbf{z_i}$ is the projected embedding of anchor $i$, $\mathbf{z_p}$ is the embedding of a positive pair to anchor, $i$ and $\mathbf{z_a}$ indexes all other embeddings in the set, $\mathbf{z}$.

Following training, the projection heads are frozen and the concatenated vector $\mathbf{h_i} = [\mathbf{e_i}|\mathbf{m_i}]$ is passed into a logistic regression classifier.

### 2.4.2. Concatenation of the Latent Spaces

A simple fusion approach is to concatenate the two raw latent embeddings to form an 88-dimensional representation, $\mu = [\mu_{\text{ecg}}|\mu_{\text{mri}}]$. Four classifiers were evaluated for both binary tasks: logistic regression, multi-layer perceptron, random forest, and XGBoost.

### 2.4.3. Ensemble Model

We explored an ensemble approach as an alternative to feature-level concatenation. Two logistic regression classifiers were trained independently on the ECG and MRI latent representations with probability outputs, $[\hat{p}_{\text{ecg}}, \hat{p}_{\text{mri}}]$, being passed to a further logistic regression classifier.

## 2.5. Experimental Setup

During the pretraining of the ECG, we evaluated five latent dimensions, $d \in \{16, 32, 64, 128, 256\}$, choosing $d = 64$ based on validation reconstruction loss and the proportion of active dimensions. Fine-tuning on the UK Biobank dataset used a batch size of 32 for 150 epochs with a grid search over 12 configurations of $\beta$ and $\gamma$. The optimal model used ($\alpha$, $\beta$, $\gamma$) = (1.0, 0.10, 0.10).

In SCL, the two projection heads are trained for 200 epochs. Logistic regression hyperparameters were selected by three-fold stratified cross-validation on the training set (penalty $\in \{\ell_1, \ell_2, \text{elasticnet}\}$, $C \in \{0.01, 0.1, 1, 10\}$).

# 3. Results and Discussion

All classifiers used 5-fold stratified cross-validation on the entire dataset with an inner 3-fold used to determine optimal hyperparameters. Final metrics were derived from the global confusion matrix accumulated across all folds.

Table 1. Performance metrics for individual and fused modalities. ECG, ED+ES, MRI, and ECG+MRI (which denotes the concatenated latent vector) are all classified using XGBoost.

| Feature | Acc. | F1 | Prec. | Rec. | AUROC | AUPRC |
|---|---|---|---|---|---|---|
| *iMI vs healthy* | | | | | | |
| ECG | 0.709 | 0.354 | 0.310 | 0.413 | 0.607 | 0.272 |
| ED+ES | 0.661 | 0.339 | 0.272 | 0.450 | 0.620 | 0.266 |
| MRI | 0.636 | 0.341 | 0.262 | 0.486 | 0.589 | 0.268 |
| ECG+MRI | 0.727 | 0.394 | 0.345 | 0.459 | 0.646 | 0.324 |
| Ensemble | 0.627 | 0.386 | 0.283 | 0.606 | 0.683 | 0.339 |
| SCL | 0.638 | 0.389 | 0.289 | 0.596 | 0.649 | 0.307 |
| *pMI vs healthy* | | | | | | |
| ECG | 0.682 | 0.416 | 0.392 | 0.442 | 0.642 | 0.391 |
| ED+ES | 0.693 | 0.464 | 0.420 | 0.519 | 0.689 | 0.420 |
| MRI | 0.653 | 0.485 | 0.391 | 0.641 | 0.683 | 0.422 |
| ECG+MRI | 0.680 | 0.488 | 0.413 | 0.596 | 0.676 | 0.433 |
| Ensemble | 0.690 | 0.531 | 0.433 | 0.686 | 0.725 | 0.451 |
| SCL | 0.728 | 0.551 | 0.477 | 0.654 | 0.721 | 0.461 |

Simple concatenation improves F1 scores over any of the single modality benchmarks for both iMI (0.394) and pMI (0.488), confirming complementary discriminative information across modalities. This is further supported by the latent traversals of the third most influential ECG dimension for pMI classification. Figure 3 reveals that the dimension encodes the ST-segment morphology, which is a well established indicator of MI in the literature [2].

The ensemble approach allowed each modality to form independent probability estimates before fusion, which prevents one modality's latent space impacting the other when there is no alignment.

Cross-modal supervised contrastive learning achieves the strongest pMI classification performance of any method with an F1 score of 0.551. By enforcing cross-modal alignment during training, the contrastive objective

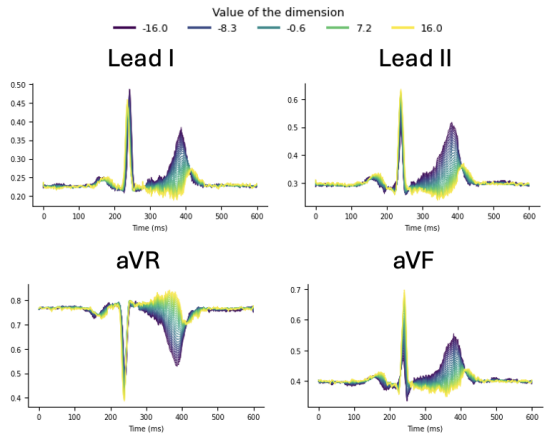


Figure 3. Latent traversals of the third most discriminative ECG latent dimension for pMI classification.

encourages the model to identify shared discriminative patterns between the electrical and mechanical manifestations of MI, producing a more physiologically coherent joint representation.

The results consistently show across all methods and metrics that iMI is more difficult to identify with the best iMI F1 score being 28.5% lower than the best pMI F1 score. Subjects with iMI were imaged prior to their infarction meaning structural cardiac remodelling and established electrophysiological changes may not be present at the time of screening.

## 4. Conclusion

We proposed a multi-modal framework for MI classification combining 12-lead ECG representations with continuous-time biventricular motion derived from cardiac MRI. Among the evaluated fusion strategies, supervised contrastive learning achieved the best overall performance, particularly for pMI classification.

The results highlight the complementary nature of electrophysiological and motion-based features, while also emphasising the challenge of predicting incident MI, where discriminative signals may be subtle or absent at the time of imaging. These findings suggest that explicitly aligning ECG and MRI representations may improve cardiovascular risk stratification."

## Acknowledgments

This research has been conducted using the UK Biobank Resource under Application Number '135228'. The authors express no conflict of interest. TS is supported by the EPSRC Centre for Doctoral Training in Health Data Science (EP/S02428X/1). AB is supported by the Royal Society University Research Fellowship (Grant No. URF\R1\221314).

## References

[1] Institute for Health Metrics and Evaluation (IHME). Global burden of disease 2021: Findings from the GBD study 2024;.

[2] Raul A, et al. A narrative review of electrocardiogram manifestation of myocardial infarction with non-obstructive coronary arteries (MINOCA). J Thorac Dis 2025;17.

[3] Borlaug BA, Sharma K, Shah SJ, Ho JE. Heart failure with preserved ejection fraction: JACC scientific statement. J Am Coll Cardiol 2023;.

[4] Beetz M, Banerjee A, Grau V. Multi-domain variational autoencoders for combined modeling of mri-based biventricular anatomy and ecg-based cardiac electrophysiology. Front Physiol Jun. 2022;13:886723.

[5] Li F, et al. CARDIAC-FM: A multimodal foundation model for cardiovascular risk prediction using ECG and cardiac MRI. medRxiv 2026;.

[6] Petersen SE, et al. UK biobank's cardiovascular magnetic resonance protocol. J Cardiovasc Magn Reson 2015; 18(1):1–7.

[7] Banerjee A, et al. A completely automated pipeline for 3D reconstruction of human heart from 2D cine magnetic resonance slices. Phil Trans R Soc A May 2021; 379(2212):20200257.

[8] Beetz M, et al. Multi-class point cloud completion networks for 3D cardiac anatomy reconstruction from cine magnetic resonance images. Med Image Anal 2023;90:102975.

[9] Zheng J, et al. A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. Scientific Data feb 2020;7(1):48.

[10] Kapsecker M, Moller MC, Jonas SM. Disentangled representational learning for anomaly detection in single-lead electrocardiogram signals using variational autoencoder. Comput Biol Med 2025;184:109422.

[11] Seale T, Rodriguez B, Grau V, Banerjee A. Modelling multi-phase cardiac anatomy with generative deep learning. Comput Cardiol 2024;51:461.

[12] Khosla P, et al. Supervised contrastive learning. Adv Neural Inf Process Syst Mar. 2020;33.

Address for correspondence:

Charlotte Grace Richardson
Old Road Campus Research Building, Oxford, OX3 7DQ, UK
charlotte.richardson@wadham.ox.ac.uk