HeartMAE: Advancing Cardiac MRI Analysis through Optical Flow Guided Masked Autoencoding

Vladislav Kim VLADISLAV.KIM@BAYER.COM
Lisa Schneider LISA.SCHNEIDER1@BAYER.COM
Machine Learning Research, Pharmaceuticals R&D, Bayer AG
Soodeh Kalaie S.KALAIE@IMPERIAL.AC.UK
Declan O'Regan DECLAN.OREGAN@LMS.MRC.AC.UK
MRC Laboratory of Medical Sciences, Imperial College London
Christian Bender CHRISTIAN.BENDER@BAYER.COM
Translational Sciences, Pharmaceuticals R&D, Bayer AG

Abstract

Cardiac MRI is a powerful diagnostic tool, but traditional analysis relies on complex segmentation-based workflows that may provide only a partial picture of cardiovascular health. To address these limitations, we introduce HeartMAE, a novel framework that uses masked autoencoding (MAE) to learn features directly from cardiac MRIs, without any labels. By incorporating optical flow during training, HeartMAE is guided towards cardiac regions, which significantly improves its downstream performance. A multitask model, built on a shared HeartMAE embedding layer, accurately predicts key cardiac health indicators, extracardiac features and major cardiovascular conditions. Moreover, HeartMAE features may be used as embeddings for clustering to enable patient stratification. Requiring only MRI data, HeartMAE is highly scalable and adaptable to larger datasets, paving the way for foundation models in cardiac imaging.

Keywords: Cardiac MRI, spatiotemporal masked autoencoders, vision transformers, optical flow, multiple instance learning, multitask classification and regression.

Data and Code Availability This research has been conducted using the UK Biobank Resource under application numbers 40616 and 28807, and the primary data source for this study is the cardiac MRI dataset, which is available to researchers upon successful application[1]. The complete source code for HeartMAE is available in a GitHub repository[2].

1. https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access
2. https://github.com/Bayer-Group/HeartMAE.git

Institutional Review Board (IRB) All participants provided written informed consent for participation in the study, which was approved by the UK National Research Ethics Service (11/NW/0382).

1. Introduction

Cardiac magnetic resonance (CMR) imaging is a non-invasive imaging technique that enables thorough assessment of the structure and function of the heart and surrounding blood vessels. Classical cardiac features such as myocardial thickness and ventricular ejection fractions are typically derived from segmentations of anatomical structures. To date, the evaluation based on these expert-derived features is considered a standard diagnostic procedure. Despite their effectiveness, expert-derived features are often limited to anatomical and functional characteristics. To uncover complex patterns that might not be perceived by human experts, we pretrained masked autoencoders on CMRs and used the embeddings for downstream tasks, including the prediction of cardiac health indicators, major cardiovascular diseases, extracardiac features as well as for patient stratification. Our end-to-end approach further streamlines CMR analysis by predicting cardiac features without the need for segmentation and postprocessing.

Historically, segmentation was performed manually (Suinesiaputra et al., 2018), but recent deep learning approaches have enabled fully automated analyses (Bai et al., 2018). While deep learning has been extensively applied to CMR segmentation (Jafari et al., 2023), end-to-end feature extraction remains relatively unexplored. Shad et al. (2023) used

contrastive learning to integrate patient CMRs with associated radiology reports. In contrast, our study focuses solely on CMR data to train masked autoencoders. We hypothesized that an unsupervised model, without textual supervision from radiology findings, might attend equally to cardiac and extracardiac tissues, diluting its focus on cardiac features. To address this, we exploited the fact that the heart is the largest moving object in CMRs and used optical flow to weight the reconstruction loss, encouraging the autoencoder to prioritize cardiac regions during training. The resulting model, HeartMAE, produced features that demonstrated superior performance in downstream evaluations. Our key contributions are

- We show that models trained exclusively on cardiac MRIs without text annotations can capture heart-related features, such as ejection fractions and cardiac chamber volumes.
- We developed a novel approach using optical flow densities as weights to focus the model on the cardiac region, leading to improved prediction of cardiac features and cardiovascular diseases.
- We found that all pretrained models capture extracardiac features such as demographic covariates (sex, age) or health-related metrics (e.g., BMI, blood pressure, lung capacity). These holistic CMR-derived phenotypes may lead to improved risk stratification strategies.
- We demonstrated that HeartMAE features can be used for patient stratification as shown for atrial fibrillation.

2. Related Work

2.1. Self-Supervised Learning for Videos

In self-supervised learning (SSL), models are pretrained on large-scale unlabelled data and subsequently fine-tuned for downstream prediction tasks. Mirroring advances in image and language domains, early video SSL methods used pretext-based learning such as temporal order inference (Misra et al., 2016) or future frame prediction (Vondrick et al., 2016). Contrastive learning (CL) has emerged as a dominant video SSL paradigm (Schiappa et al., 2023). CL generates positive pairs from the same video using augmentations, contrasting them against views from other videos (Han et al., 2020). Recent works have introduced multimodal contrastive learning (Miech et al., 2020; Morgado et al., 2021; Rouditchenko et al., 2021), integrating video, text and/or audio modalities to achieve state-of-the-art performance. While convolutional networks were traditionally used as video encoders, current approaches primarily use variants of the Vision Transformer (ViT) architecture (Dosovitskiy et al., 2021) adapted for video processing (Bertasius et al., 2021; Arnab et al., 2021; Neimark et al., 2021; Fan et al., 2021).

2.2. Masked Autoencoders

Masked autoencoders (MAEs), originally introduced for image-level tasks (He et al., 2022), learn features by masking a large portion (75%) of the tokenized input and reconstructing the original image. MAEs process only unmasked tokens, enabling efficient training with a lightweight decoder. This approach has been extended to videos (Feichtenhofer et al., 2022; Tong et al., 2022), incorporating frame sampling and higher masking ratios (90%) to address spatiotemporal characteristics. Recent works have combined masked autoencoding with contrastive learning objectives (Wang et al., 2022), achieving state-of-the-art performance on various video tasks.

2.3. Deep Learning for Cardiac MRI

The application of deep learning techniques to CMR imaging has seen significant growth in recent years. Beyond improving its acquisition and reconstruction, the analysis of CMR data has emerged as a major area of application (Leiner et al., 2019). Deep learning-based segmentation of anatomical structures has been extensively explored in previous works aiming to derive clinically-relevant cardiac health indicators (Chen et al., 2020; Jafari et al., 2023; Pirruccello et al., 2024).

Similarly, deep learning methods were used to segment pathological tissues such as myocardial scars to improve the assessment of myocardial infarctions (Jafari et al., 2023; Zhang et al., 2019). Another key focus of previous studies was cardiovascular disease prediction (Martini et al., 2020; Zhang et al., 2019) and risk stratification (Zhou et al., 2021).

Recent works shifted towards developing holistic cardiac health representations and foundation model approaches. Notably, Shad et al. (2023) integrated CMR videos with radiology reports using multimodal contrastive learning, demonstrating strong performance of this approach in predicting clinical indica-

tors and cardiovascular diseases. Additionally, cross-modal autoencoders (Radhakrishnan et al., 2023) were used to predict CMR from ECG data, enabling partial imputation of an expensive imaging modality from a more accessible diagnostic tool.

3. Data

We used imaging data from the UK Biobank (UKB, Bycroft et al. (2018)), a comprehensive biomedical data resource comprising 500,000 participants with extensive health records, genotype data and partially brain, cardiac, and full-body magnetic resonance (MR) imaging.

Specifically, we used CMR videos (Raisi-Estabragh et al., 2021), which capture temporal image sequences of the heart. Each CMR video consists of 50 frames, covering one complete cardiac cycle, acquired during breath hold to minimize motion artifacts. For each subject, CMR videos are obtained from multiple angles, producing short-axis (SAX) and long-axis (LAX) views, with multiple optical sections acquired for the SAX views.

The complete UKB CMR dataset consists of about 1 million CMR videos from 65,000 subjects. To optimize model training, we created a subset by randomly downsampling SAX optical sections by 50%. This additionally addressed the 10:3 data imbalance between SAX and LAX views. The resulting dataset of 370,000 CMR videos was partitioned into training, validation, and test sets using a 70%/15%/15% split.

For evaluation, we leveraged 99 expert-derived cardiac features (e.g., ejection fractions, volumes, wall thickness) from segmentation-based CMR analysis for a subset of 40,000 subjects. Data splits were derived from the CMR splits, which resulted in 31,000/4,500/4,500 subjects in train/validation/test sets. We further utilized ICD-10 codes to manually curate meaningful cardiovascular disease labels, which were available for a subset of 61,000 subjects. Given the class imbalance, we only included the diseases with at least 50 diagnoses in training, 20 in validation, and a minimum of 20 diagnoses for testing. The data was split according to CMR splits (50,000/5,500/5,500 subjects in train/validation/test splits). Lastly, we used physical measurements such as body composition, blood pressure, and lung function metrics to test the ability of models to capture extracardiac information.

4. Methods

4.1. Spatiotemporal Masked Autoencoder

We adapted spatiotemporal MAE (Feichtenhofer et al., 2022) for CMR data and trained an initial MAE-ViT-B[3] model with 90% mask ratio (MR) for 200 epochs. During training, we randomly sampled 24-frame clips[4] from each CMR video, with spatial and temporal augmentations (see Appendix A for full details).

To explore the impact of masking, we trained two additional models, initialized from the best MAE-ViT-B checkpoint: 1) *MAE-tube* with 90% MR and tube masking, i.e. randomly sampling patches across the entire temporal axis introduced in (Tong et al., 2022) and 2) *MAE-95* using random masking with a higher ratio (95%). These models were trained for 100 epochs and, as MAE-ViT-B, used the mean squared error (MSE) loss on the set M of masked token pixels:

$$\mathcal{L}_{MSE} = \frac{1}{|M|} \sum_{(i,j)\in M} (I_{ij} - \hat{I}_{ij})^2 \tag{1}$$

where I_{ij} is the original normalized pixel value and $\hat{I}_{ij}$ is the reconstructed pixel value.

4.2. HeartMAE: Optical Flow Guided MAE

HeartMAE (Figure 1) addresses the challenge of focusing unsupervised learning on cardiac regions in CMR videos. Leveraging the fact that the heart is the primary moving object, we incorporated cardiac motion information to guide MAE, prioritizing the heart area in the reconstruction task. We used optical flow (OF), a vector field representing point displacements between consecutive video frames, to estimate cardiac motion. Using the Farneback method (Farnebäck, 2003), we estimated OF between every 5th frame in each CMR video. The OF magnitude $\mathbf{\Phi}$ was calculated as the pointwise L_2-norm of the vertical and horizontal OF components $\mathbf{U}$ and $\mathbf{V}$:

$$\mathbf{\Phi}(i,j) = \sqrt{\mathbf{U}(i,j)^2 + \mathbf{V}(i,j)^2}$$

The mean OF magnitude was computed across the entire video range:

$$\overline{\mathbf{\Phi}} = \frac{1}{n} \sum_{k=1}^{n} \mathbf{\Phi}_k \text{ with } n = L_{video}/5$$

3. ViT-base with patch sizes $(p, p, t) = (16, 16, 4)$.
4. Clips consisted of consecutive frames to preserve temporal dynamics.

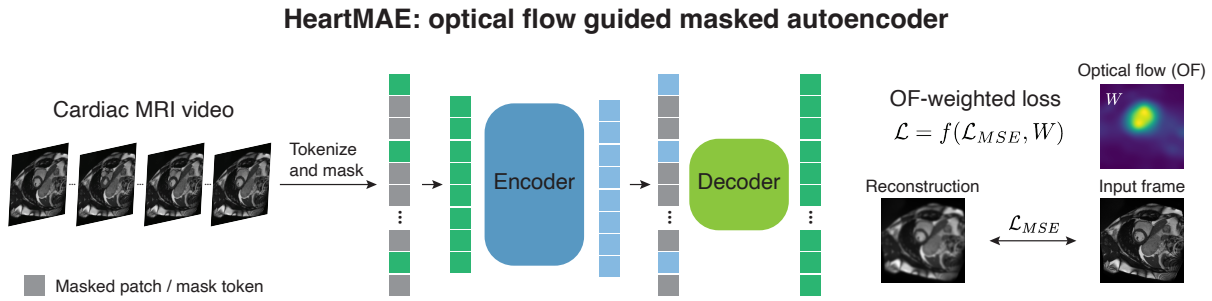

Figure 1: HeartMAE architecture and schematic representation of optical flow (OF) weighting during model training.

where L_{video} represents the number of frames in a CMR video. To produce the weight matrix $\mathbf{W}$, we smoothed $\overline{\mathbf{\Phi}}$ using a Gaussian kernel with $\sigma = 10$ and applied max-scaling. We incorporated cardiac motion weights $W_{ij} \in [0, 1]$ into the MSE loss of Equation (1), amplifying the contribution of moving masked pixels:

$$\mathcal{L} = \frac{1}{|M|} \sum_{(i,j)\in M} (I_{ij} - \hat{I}_{ij})^2 (1 + W_{ij}) \tag{2}$$

To reduce computational cost, we initialized HeartMAE with the weights of MAE-ViT-B[5] and trained using the loss in Equation (2) and random masking (MR = 90%) for 100 epochs. We compared HeartMAE features with those of MAE-ViT-B, MAE-tube, and MAE-95 on various downstream prediction tasks as described in Subsection 4.4.

4.3. Inference and Feature Aggregation

Features were extracted from each CMR video using pretrained MAE encoders. Following Feichtenhofer et al. (2022), we used multiview inference with $K = 3$ spatial crops (top/center/bottom or left/center/right) and $T = 2$ temporal clips to cover the full spatial extent and temporal duration of each video. Additionally, each subject had S cardiac views[6] (refer to Section 3), resulting in a set of $V = S \cdot K \cdot T$ features $\{\mathbf{h}_i : i \in 1, \ldots, V\}$ per subject.

For feature aggregation, we used attention-based multiple instance learning (ABMIL) Ilse et al. (2018). The final aggregated feature $\mathbf{z}$ for each subject was computed by applying a weighted mean of the individual view features $\mathbf{h}_i$:

$$\mathbf{z} = \sum_{i=1}^{V} a_i \mathbf{h}_i \tag{3}$$

where the attention weights a_i represent the importance of each view:

$$a_i = \frac{\exp\{\mathbf{w}^\top (\tanh(\mathbf{V}\mathbf{h}_i{}^\top) \odot \mathrm{sigm}(\mathbf{U}\mathbf{h}_i{}^\top))\}}{\sum_{j=1}^{V} \exp\{\mathbf{w}^\top (\tanh(\mathbf{V}\mathbf{h}_j{}^\top) \odot \mathrm{sigm}(\mathbf{U}\mathbf{h}_j{}^\top))\}}$$

Attention weights a_i were learned simultaneously with supervised model weights for cardiac feature and disease prediction (see Subsection 4.4).

4.4. Multitask Classification and Regression

We used multitask learning to predict expert-derived cardiac features, cardiovascular diseases and extracardiac features using pretrained MAE features (see Subsection 4.1), with the encoder weights frozen. The multi-head architecture incorporated a shared ABMIL layer (see Subsection 4.3) for subject-level feature aggregation, followed by task-specific linear prediction heads. This approach effectively adapts linear probing, commonly used to evaluate SSL models (He et al., 2022), to the multiview cardiac imaging domain.

For cardiac feature regression, we jointly optimized the ABMIL layer and prediction head weights using the MSE loss:

$$\mathcal{L}_{MSE} = \frac{1}{N_{subj} \cdot N_{task}} \sum_{j=1}^{N_{subj}} \sum_{i=1}^{N_{task}} (y_{ij} - \hat{y}_{ij})^2 \tag{4}$$

5. As was done for MAE-tube and MAE-95.
6. The number of cardiac views S varied by subject, ranging from 1 to over 10, with the mean $\overline{S} = 5$.

Prior to regression modeling, we centered and scaled the target features y using training set statistics, which improved model performance. The same prediction scheme was used for extracardiac features.

For cardiovascular disease classification, we used the cardiovascular disease labels (see Section 3). We jointly learned the ABMIL and prediction head parameters using the binary cross-entropy (BCE) loss without class weights[7]:

$$\mathcal{L}_{BCE} = \frac{1}{N_{subj} \cdot N_{task}} \sum_{j=1}^{N_{subj}} \sum_{i=1}^{N_{task}} CE(p_{ij}, \hat{p}_{ij}) \quad (5)$$

with $CE(p, \hat{p}) = p \log(\hat{p}) + (1-p) \log(1-\hat{p})$.

For both regression and classification, we used a batch size of 1 to accommodate the variable number of input views (V) in the ABMIL layer (see Subsection 4.3).

4.5. Patient Stratification

To demonstrate the utility of HeartMAE features for patient stratification, we selected a subset of 1,700 UK Biobank subjects with atrial fibrillation[8] along with 1,700 healthy subjects without any history of cardiovascular disease.

We extracted subject-level HeartMAE features using the ABMIL layer (see Subsection 4.3), which was previously trained on the multitask cardiac feature regression (Subsection 4.4). We then used agglomerative clustering to group UKB subjects based on these HeartMAE features. For the second clustering, we adjusted HeartMAE features for sex to account for the difference in heart size between males and females. This adjustment was performed by independently regressing out the sex covariate from each feature. Specifically, we fit a linear regression model for each feature with sex as the predictor variable, and used the residuals from these models as the new, sex-adjusted features.

Hardware Details

All MAE model training was performed on 4 NVIDIA A100 GPUs (80GB memory each). The initial MAE-ViT-B model required approximately 24 days for 200 epochs. To save compute, all subsequent models (MAE-tube, MAE-95, HeartMAE) were initialized with the best MAE-ViT-B checkpoint weights and trained only for 100 epochs. This reduced training times by 50%.

Feature extraction using pretrained MAE models is computationally lightweight, processing 6 CMR videos per second on a single NVIDIA V100 GPU or 8 videos per second on an NVIDIA A100 GPU.

5. Results

To assess the predictive power of MAE features, we evaluated their ability to predict (1) key clinically-relevant cardiac metrics, (2) cardiovascular diseases, and (3) extracardiac features using multitask learning. We compared 4 MAE models trained on the UKB CMR dataset: MAE-ViT-B, MAE-tube, MAE-95, and our proposed HeartMAE model against a baseline using transfer learning features from natural videos (Feichtenhofer et al., 2022). For disease prediction, expert-derived cardiac features provided an additional baseline.

5.1. Prediction of Cardiac Health Indicators

Ejection fraction (EF) is a key cardiac health indicator that quantifies the percentage of blood volume ejected from a heart chamber during each contraction. Table 1 compares the performance of the 4 MAE models against two reference models in predicting EFs for various cardiac chambers.

Table 1: Mean absolute errors of ejection fraction (EF) prediction on the test set, with the best values highlighted in bold. LV = left ventricle, LA = left atrium, RV = right ventricle, RA = right atrium.

[§]MAE-ViT-L (Feichtenhofer et al., 2022) is a transfer learning baseline from natural videos. [†]mViT-B (Shad et al., 2023) was pretrained on a clinical CMR dataset.

Model	LVEF	LAEF	RVEF	RAEF
Reference models				
MAE-ViT-L[§]	4.14	5.93	4.07	6.33
mViT-B[†]	3.34	–	–	–
Models trained on the UKB CMR dataset				
MAE-ViT-B	3.14	4.32	3.03	5.39
MAE-95	3.19	4.38	3.09	5.42
MAE-tube	3.15	4.31	3.06	5.31
HeartMAE	**3.09**	**4.26**	**2.98**	**5.25**

7. Neither class weighting nor focal loss produced better results compared to BCE.
8. Diagnosis confirmed before the imaging visit.

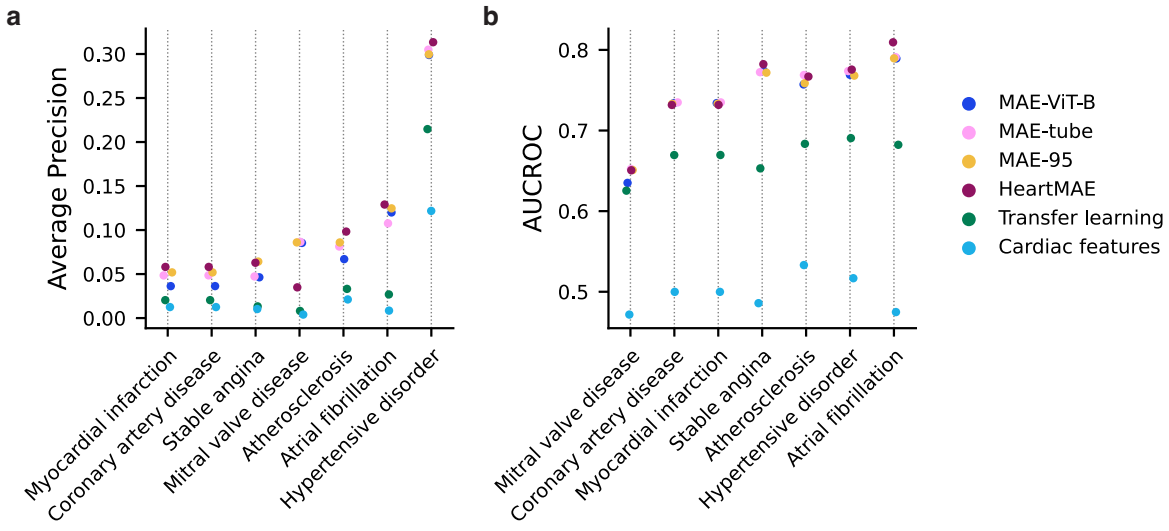

Figure 2: a) Average precision and b) area under receiver operating characteristic curve (AUCROC) of cardiovascular disease (x-axis) prediction on the test set. The diseases were sorted by their mean a) AP and b) AUCROC values, respectively. Transfer learning is a MAE-ViT-L pretrained on the Kinetics-700 dataset (Feichtenhofer et al., 2022).

All models pretrained on CMR data outperformed the transfer learning baseline from natural videos (Table 1). Neither tube masking (MAE-tube) nor a higher random masking ratio (MAE-95) produced superior results compared to the original MAE-ViT-B model. Notably, HeartMAE outperformed all other models across all 4 EF prediction tasks (Table 1).

Left ventricular ejection fraction (LVEF) is an important measure quantifying the severity of cardiac dysfunction. We included LVEF prediction results from Shad et al. (2023) on the UKB dataset in Table 1 for comparison[9]. HeartMAE outperformed all models, including mViT-B of Shad et al. (2023). Given that Shad et al. (2023) reported their LVEF estimation to be within clinicians' error limits, our results indicate that HeartMAE features can predict LVEF with high accuracy.

Additionally, we extended our comparison to the prediction of all 99 expert-derived cardiac features (Supplementary Fig. S1). HeartMAE demonstrated superior performance in 86 out of 99 tasks, including left ventricular mass (LVM) and left ventricular end-diastolic volume (LVEDV) – important cardiac health indicators requiring 3D volume estimation. The improvement of HeartMAE over MAE-ViT-B was highly significant ($p < 0.001$, Wilcoxon signed-rank test comparing errors across 99 tasks). This suggests that focusing the model on the cardiac region during training improves the extraction of clinically relevant features from CMR data.

5.2. Cardiovascular Disease Prediction

Next, we used MAE features to predict cardiovascular disease labels of UKB subjects (see Section 3). The results for 4 MAE models and 2 baselines are reported in Figure 2. To account for the class imbalance caused by the predominance of healthy subjects, we adopted average precision (AP) as our primary metric.

All CMR-pretrained MAE models consistently outperformed transfer learning and expert-derived cardiac feature baselines across all diseases (Figure 2). This demonstrates their efficacy in predicting cardiovascular diseases, even on the unbalanced UK Biobank dataset. HeartMAE demonstrated superior performance (Figure 2a), surpassing all models on 5 out of 7 disease prediction tasks, albeit with smaller margins than in the cardiac feature prediction task. Interestingly, MAE-95 ranked second despite its weak

9. Shad et al. (2023) reported results only for LVEF prediction on the UKB dataset, limiting our comparison to this specific metric.

Table 2: Prediction of extracardiac features. Mean absolute errors ("Err.") and Pearson correlations ("Cor.") between predictions and ground-truth values are provided for predictions of age, body mass index (BMI), systolic and diastolic blood pressure (SBP and DBP), forced expiratory volume (FEV), and forced vital capacity (FVC). Transfer learning is a MAE-ViT-L pretrained on Kinetics-700 (Feichtenhofer et al., 2022). Best values are highlighted in bold.

Model	Age		BMI		DBP		SBP		FEV		FVC	
	Err.	Cor.	Err.	Cor.	Err.	Cor.	Err.	Cor.	Err.	Cor.	Err.	Cor.
MAE-ViT-B	2.90	0.88	0.95	0.95	6.58	0.51	10.91	0.63	**0.31**	0.81	0.32	0.88
MAE-95	2.89	0.88	0.96	0.95	6.57	0.51	11.01	0.61	0.32	0.82	0.32	0.88
MAE-tube	2.81	**0.89**	**0.94**	**0.96**	**6.53**	**0.52**	10.95	0.62	**0.31**	**0.83**	**0.31**	**0.89**
HeartMAE	**2.79**	**0.89**	0.95	0.95	6.57	**0.52**	**10.71**	**0.64**	**0.31**	0.82	0.32	0.88
Transfer learning	4.70	0.66	1.48	0.89	7.16	0.36	13.10	0.28	0.39	0.72	0.44	0.78

performance in predicting expert-derived cardiac features (Subsection 5.1).

AUCROC[10] analysis (Figure 2b) revealed a wider performance gap between MAE models and transfer learning. HeartMAE outperformed all other models on 3 out of 7 disease prediction tasks and ranked second on another, underscoring the benefits of focusing on cardiac regions during pretraining of unsupervised models for CMR data. While HeartMAE achieved higher AUCROC values, the improvement over MAE-ViT-B was only marginally significant ($p < 0.05$, Wilcoxon signed-rank test). The smaller performance gap between HeartMAE and other MAE models suggests that disease predictions may be influenced by extra-cardiac health indicators and risk factors (see Subsection 5.3 and Section 6).

5.3. Prediction of Extracardiac Features

Unsegmented CMR videos capture not only the heart but also surrounding tissues, potentially containing valuable extracardiac information. To explore this, we evaluated all MAE models (Subsection 4.1) and transfer learning features in their ability to predict various extracardiac features such as age, body mass index (BMI), blood pressure, and lung function metrics.

All MAE models (Table 2) demonstrated comparable performance in predicting extracardiac features. BMI and lung capacity measures (FEV and FVC) were easily inferred by all models, including the transfer learning baseline, indicating that CMR data contains body composition and pulmonary function information, extractable even by models pretrained on natural videos. In contrast, blood pressure predictions (SBP and DBP) proved more challenging.

Interestingly, subject age was predicted with high correlations by all MAE models (Table 2), with HeartMAE achieving the lowest error, which paves the way for developing more accurate cardiac age (Lima et al., 2021) models for improved risk stratification.

5.4. Patient Stratification

Given the scarcity of comprehensive disease labels in the biomedical domain, patient stratification through clustering is an important use case of deep learning embeddings. We evaluated HeartMAE's potential for embedding-based patient stratification using a subset of healthy subjects and those with atrial fibrillation (see Subsection 4.5). Initial clustering using "raw" HeartMAE features (Figure 3a) revealed distinct atrial fibrillation subgroups, but was dominated by biological sex, an effect also observed globally in the UMAP visualization (Figure S2). After adjustment for sex (Subsection 4.5), HeartMAE features produced two primary clusters (Figure 3b): one with predominantly healthy and another with mostly atrial fibrillation subjects. Left atrial ejection fraction (LAEF) was significantly lower in the diseased cluster (Figure S3), demonstrating HeartMAE's ability to stratify subjects based on both disease status and clinically relevant cardiac features.

10. While direct comparison with mViT-B of Shad et al. (2023) is not feasible due to their use of a proprietary CMR dataset, our test set AUCROC values are comparable in magnitude to the results reported in their study.

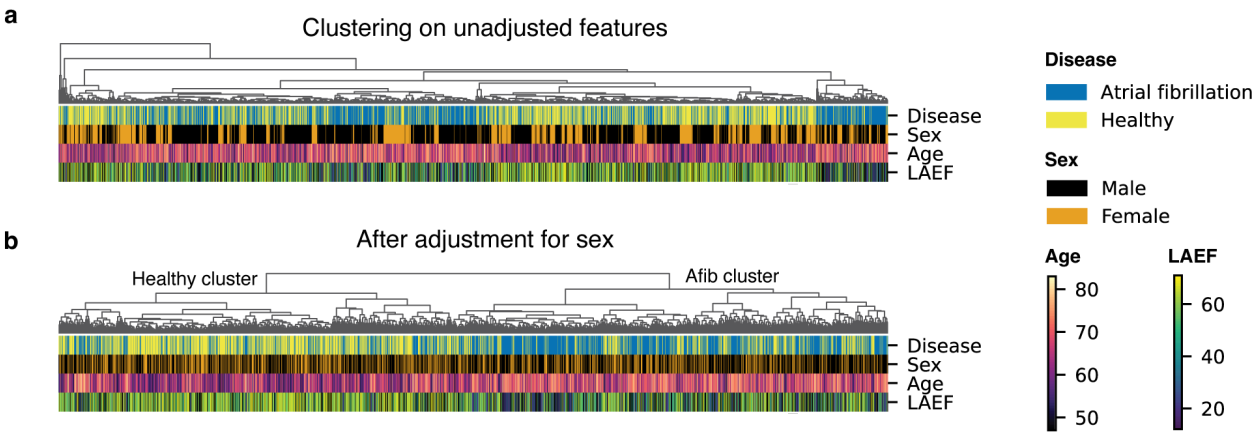

Figure 3: Agglomerative clustering of a subset with healthy UKB subjects and those diagnosed with atrial fibrillation based on a) unadjusted (raw) HeartMAE features and b) HeartMAE features adjusted for biological sex. Four colored annotation tracks were added to indicate subject disease status, sex, age, and left atrial ejection fraction (LAEF).

6. Discussion

In this work, we introduced HeartMAE, a novel deep learning framework for cardiac MRI analysis. By focusing an unsupervised masked autoencoder on cardiac regions during training, we improved the prediction of 86 cardiac features, including LVEF, directly from CMR videos. HeartMAE processes CMR data without segmentation, offering an efficient approach to cardiac assessment.

Our results show that all MAE models effectively predict key cardiac health indicators (Subsection 5.1) and, to some extent, cardiovascular disease diagnoses (Subsection 5.2). The superior performance of MAE models over classical cardiac features in disease prediction suggests that our approach captures additional information potentially missed by expert-derived cardiac features. Furthermore, all MAE models outperformed the transfer learning baseline, underscoring the importance of training models directly on domain-specific datasets.

We note that cardiac disease prediction using the UKB dataset presented challenges due to its predominantly healthy cohort, resulting in significant class imbalance. Furthermore, the potential time lag between imaging and disease onset introduced further complexity. For simplicity and to assess our models' potential for early disease detection, we did not distinguish between subjects diagnosed before or after the imaging visit. Given the limitations of UKB disease labels, we additionally demonstrated HeartMAE features' ability to stratify subjects based on their disease status without curated labels, successfully separating healthy subjects from those with atrial fibrillation through clustering (Subsection 5.4).

Our transfer learning results suggest that HeartMAE will generalize effectively to other cardiac MRI datasets with minimal adaptations. Although the original MAE (Feichtenhofer et al., 2022) performed below CMR-pretrained MAE models, it showed remarkable predictive capabilities in absolute terms despite being pretrained only on natural videos. HeartMAE, trained on all four standard CMR views, should demonstrate even better in-domain transferability. We plan to validate this hypothesis on the dataset from Shad et al. (2023), which comprises MRI measurements from multiple US hospitals and includes a population with higher disease prevalence.

Despite focusing HeartMAE on the heart during training, all MAE models, including HeartMAE and a MAE pretrained on natural videos, could predict certain extracardiac features with high accuracy. This suggests that the ability to capture extracardiac information is inherent to such end-to-end approaches. However, we observed lower performance for some extracardiac learning tasks, such as systolic and diastolic blood pressure (SBP and DBP). This variability in performance is expected, as not all extracar-

diac features are equally represented in cardiac imaging. Furthermore, we observed a performance gap between CMR-pretrained MAE models and transfer learning in age prediction, potentially due to CMR-predicted age being more closely related to cardiac health, as previously shown for ECG-predicted age (Lima et al., 2021).

While HeartMAE's training is computationally demanding, inference with the pretrained model is efficient, processing 8 CMR videos per second – making it suitable for real-time analysis. However, clinical adoption of HeartMAE may be hindered by its "black-box" nature. Although off-the-shelf interpretability can be achieved through visualization of self-attention or multiple instance learning (MIL) weights, these may not provide adequate clinical insights. Self-attention weights merely indicate which regions of the CMR sequence were important during pretraining, while MIL weights only show which CMR views contributed most to predictions.

Our study has several limitations that present opportunities for future research. First, the implications of capturing extracardiac features remain uncertain. Future studies should investigate whether these features provide complementary information for cardiac health assessment or potentially introduce confounding factors. Second, the current HeartMAE features lack interpretability in disease prediction. To address this, future research should explore explainable AI methods (Selvaraju et al., 2017; Achtibat et al., 2024), which could provide insights into the model's decision-making process and facilitate clinical adoption. Another promising direction is combining masked image modelling with the contrastive objective as in (Wang et al., 2022), which could improve model performance. Lastly, our study focused solely on CMR data, not utilizing the full range of modalities available in the UK Biobank. Integrating additional data such as ECG or genetic information could provide a more holistic view of cardiac health, as demonstrated by recent multimodal studies (Shad et al., 2023; Radhakrishnan et al., 2023).

Acknowledgments

We thank members of the Machine Learning Research and Translational Sciences teams at Bayer for their valuable feedback throughout this research. We are grateful to the UK Biobank staff for providing and maintaining the data resource. We also thank our collaborators at Imperial College for their support and insights.

References

Reduan Achtibat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. AttnLRP: Attention-aware layer-wise relevance propagation for transformers. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 135–168. PMLR, 21–27 Jul 2024.

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. ViViT: A Video Vision Transformer, November 2021. URL http://arxiv.org/abs/2103.15691. arXiv:2103.15691 [cs].

Wenjia Bai, Matthew Sinclair, Giacomo Tarroni, Ozan Oktay, Martin Rajchl, Ghislain Vaillant, Aaron M. Lee, Nay Aung, Elena Lukaschuk, Mihir M. Sanghvi, Filip Zemrak, Kenneth Fung, Jose Miguel Paiva, Valentina Carapella, Young Jin Kim, Hideaki Suzuki, Bernhard Kainz, Paul M. Matthews, Steffen E. Petersen, Stefan K. Piechnik, Stefan Neubauer, Ben Glocker, and Daniel Rueckert. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. *Journal of Cardiovascular Magnetic Resonance*, 20 (1):65, February 2018. ISSN 1097-6647.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is Space-Time Attention All You Need for Video Understanding?, June 2021. URL http://arxiv.org/abs/2102.05095. arXiv:2102.05095 [cs].

Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, Adrian Cortes, Samantha Welsh, Alan Young, Mark Effingham, Gil McVean, Stephen Leslie, Naomi Allen, Peter Donnelly, and Jonathan Marchini. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726): 203–209, October 2018. ISSN 1476-4687. doi: 10.1038/s41586-018-0579-z. URL https://doi.org/10.1038/s41586-018-0579-z.

Chen Chen, Chen Qin, Huaqi Qiu, Giacomo Tarroni, Jinming Duan, Wenjia Bai, and Daniel Rueckert. Deep learning for cardiac image segmentation: a review. *Frontiers in cardiovascular medicine*, 7:25, 2020.

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, June 2021. URL http://arxiv.org/abs/2010.11929. arXiv:2010.11929 [cs].

Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale Vision Transformers, April 2021. URL http://arxiv.org/abs/2104.11227. arXiv:2104.11227 [cs].

Gunnar Farnebäck. Two-Frame Motion Estimation Based on Polynomial Expansion. In Josef Bigun and Tomas Gustavsson, editors, *Image Analysis*, pages 363–370, Berlin, Heidelberg, 2003. Springer. ISBN 978-3-540-45103-7. doi: 10.1007/3-540-45103-X_50.

Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked Autoencoders As Spatiotemporal Learners, October 2022. URL http://arxiv.org/abs/2205.09113. arXiv:2205.09113 [cs].

Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, pages 5679–5690, Red Hook, NY, USA, December 2020. Curran Associates Inc. ISBN 978-1-71382-954-6.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. pages 16000–16009, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/He_Masked_Autoencoders_Are_Scalable_Vision_Learners_CVPR_2022_paper.html.

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based Deep Multiple Instance Learning, June 2018. URL http://arxiv.org/abs/1802.04712. arXiv:1802.04712 [cs, stat].

Mahboobeh Jafari, Afshin Shoeibi, Marjane Khodatars, Navid Ghassemi, Parisa Moridian, Roohallah Alizadehsani, Abbas Khosravi, Sai Ho Ling, Niloufar Delfan, Yu-Dong Zhang, Shui-Hua Wang, Juan M. Gorriz, Hamid Alinejad-Rokny, and U. Rajendra Acharya. Automated diagnosis of cardiovascular diseases from cardiac magnetic resonance imaging using deep learning models: A review. *Computers in Biology and Medicine*, 160: 106998, June 2023. ISSN 0010-4825.

Tim Leiner, Daniel Rueckert, Avan Suinesiaputra, Bettina Baeßler, Reza Nezafat, Ivana Išgum, and Alistair A Young. Machine learning in cardiovascular magnetic resonance: basic concepts and applications. *Journal of Cardiovascular Magnetic Resonance*, 21(1):61, 2019.

Emilly M. Lima, Antônio H. Ribeiro, Gabriela M. M. Paixão, Manoel Horta Ribeiro, Marcelo M. Pinto-Filho, Paulo R. Gomes, Derick M. Oliveira, Ester C. Sabino, Bruce B. Duncan, Luana Giatti, Sandhi M. Barreto, Wagner Meira Jr, Thomas B. Schön, and Antonio Luiz P. Ribeiro. Deep neural network-estimated electrocardiographic age as a mortality predictor. *Nature Communications*, 12(1):5117, August 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-25351-7. Number: 1 Publisher: Nature Publishing Group.

Nicola Martini, Alberto Aimo, Andrea Barison, Daniele Della Latta, Giuseppe Vergaro, Giovanni Donato Aquaro, Andrea Ripoli, Michele Emdin, and Dante Chiappino. Deep learning to diagnose cardiac amyloidosis from cardiovascular magnetic resonance. *Journal of Cardiovascular Magnetic Resonance*, 22(1):84, 2020. ISSN 1097-6647. doi: https://doi.org/10.1186/s12968-020-00690-4. URL https://www.sciencedirect.com/science/article/pii/S1097664723003368.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, September 2020. URL http://arxiv.org/abs/1802.03426. arXiv:1802.03426 [cs, stat].

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zis-

serman. End-to-End Learning of Visual Representations From Uncurated Instructional Videos. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9876–9886, Seattle, WA, USA, June 2020. IEEE. ISBN 978-1-72817-168-5. doi: 10.1109/CVPR42600.2020.00990. URL https://ieeexplore.ieee.org/document/9157128/.

Ishan Misra, C. Lawrence Zitnick, and Martial Hebert. Shuffle and Learn: Unsupervised Learning using Temporal Order Verification, July 2016. URL http://arxiv.org/abs/1603.08561. arXiv:1603.08561 [cs].

Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-Visual Instance Discrimination with Cross-Modal Agreement. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12470–12481, Nashville, TN, USA, June 2021. IEEE. ISBN 978-1-66544-509-2. doi: 10.1109/CVPR46437.2021.01229. URL https://ieeexplore.ieee.org/document/9578129/.

Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video Transformer Network, August 2021. URL http://arxiv.org/abs/2102.00719. arXiv:2102.00719 [cs].

James P. Pirruccello, Paolo Di Achille, Seung Hoan Choi, Joel T. Rämö, Shaan Khurshid, Mahan Nekoui, Sean J. Jurgens, Victor Nauffal, Shinwan Kany, Kenney Ng, Samuel F. Friedman, Puneet Batra, Kathryn L. Lunetta, Aarno Palotie, Anthony A. Philippakis, Jennifer E. Ho, Steven A. Lubitz, and Patrick T. Ellinor. Deep learning of left atrial structure and function provides link to atrial fibrillation risk. *Nature Communications*, 15 (1):4304, May 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-48229-w. URL https://www.nature.com/articles/s41467-024-48229-w. Publisher: Nature Publishing Group.

Adityanarayanan Radhakrishnan, Sam F Friedman, Shaan Khurshid, Kenney Ng, Puneet Batra, Steven A Lubitz, Anthony A Philippakis, and Caroline Uhler. Cross-modal autoencoder framework learns holistic representations of cardiovascular state. *Nature Communications*, 14(1):2436, 2023.

Zahra Raisi-Estabragh, Nicholas C Harvey, Stefan Neubauer, and Steffen E Petersen. Cardiovascular magnetic resonance imaging in the uk biobank: a major international health research resource. *European Heart Journal-Cardiovascular Imaging*, 22 (3):251–258, 2021.

Andrew Rouditchenko, Angie Boggust, David Harwath, Brian Chen, Dhiraj Joshi, Samuel Thomas, Kartik Audhkhasi, Hilde Kuehne, Rameswar Panda, Rogerio Feris, Brian Kingsbury, Michael Picheny, Antonio Torralba, and James Glass. AVLnet: Learning Audio-Visual Language Representations from Instructional Videos. In *Proc. Interspeech 2021*, pages 1584–1588, 2021. doi: 10.21437/Interspeech.2021-1312.

Madeline C. Schiappa, Yogesh S. Rawat, and Mubarak Shah. Self-Supervised Learning for Videos: A Survey. *ACM Comput. Surv.*, 55(13s): 288:1–288:37, July 2023. ISSN 0360-0300. doi: 10.1145/3577925. URL https://doi.org/10.1145/3577925.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

Rohan Shad, Cyril Zakka, Dhamanpreet Kaur, Robyn Fong, Ross Warren Filice, John Mongan, Kimberly Kalianos, Nishith Khandwala, David Eng, Matthew Leipzig, Walter Witschey, Alejandro de Feria, Victor Ferrari, Euan Ashley, Michael A. Acker, Curtis Langlotz, and William Hiesinger. A Generalizable Deep Learning System for Cardiac MRI, December 2023. URL http://arxiv.org/abs/2312.00357. arXiv:2312.00357 [cs, eess].

Avan Suinesiaputra, Mihir M. Sanghvi, Nay Aung, Jose Miguel Paiva, Filip Zemrak, Kenneth Fung, Elena Lukaschuk, Aaron M. Lee, Valentina Carapella, Young Jin Kim, Jane Francis, Stefan K. Piechnik, Stefan Neubauer, Andreas Greiser, Marie-Pierre Jolly, Carmel Hayes, Alistair A. Young, and Steffen E. Petersen. Fully-automated left ventricular mass and volume MRI analysis in the UK Biobank population cohort: evaluation of initial results. *The International Journal of Cardiovascular Imaging*, 34(2):281–291, February 2018. ISSN 1573-0743.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked Autoencoders are

Data-Efficient Learners for Self-Supervised Video Pre-Training, October 2022. URL http://arxiv.org/abs/2203.12602. arXiv:2203.12602 [cs].

Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating Visual Representations from Unlabeled Video, November 2016. URL http://arxiv.org/abs/1504.08023. arXiv:1504.08023 [cs].

Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. InternVideo: General Video Foundation Models via Generative and Discriminative Learning, December 2022. URL http://arxiv.org/abs/2212.03191. arXiv:2212.03191 [cs].

Nan Zhang, Guang Yang, Zhifan Gao, Chenchu Xu, Yanping Zhang, Rui Shi, Jennifer Keegan, Lei Xu, Heye Zhang, Zhanming Fan, et al. Deep learning for diagnosis of chronic myocardial infarction on nonenhanced cardiac cine mri. *Radiology*, 291(3): 606–617, 2019.

Hongyu Zhou, Lu Li, Zhenyu Liu, Kankan Zhao, Xiuyu Chen, Minjie Lu, Gang Yin, Lei Song, Shihua Zhao, Hairong Zheng, et al. Deep learning algorithm to improve hypertrophic cardiomyopathy mutation prediction using cardiac cine images. *European Radiology*, 31:3931–3940, 2021.

Supplementary Figures

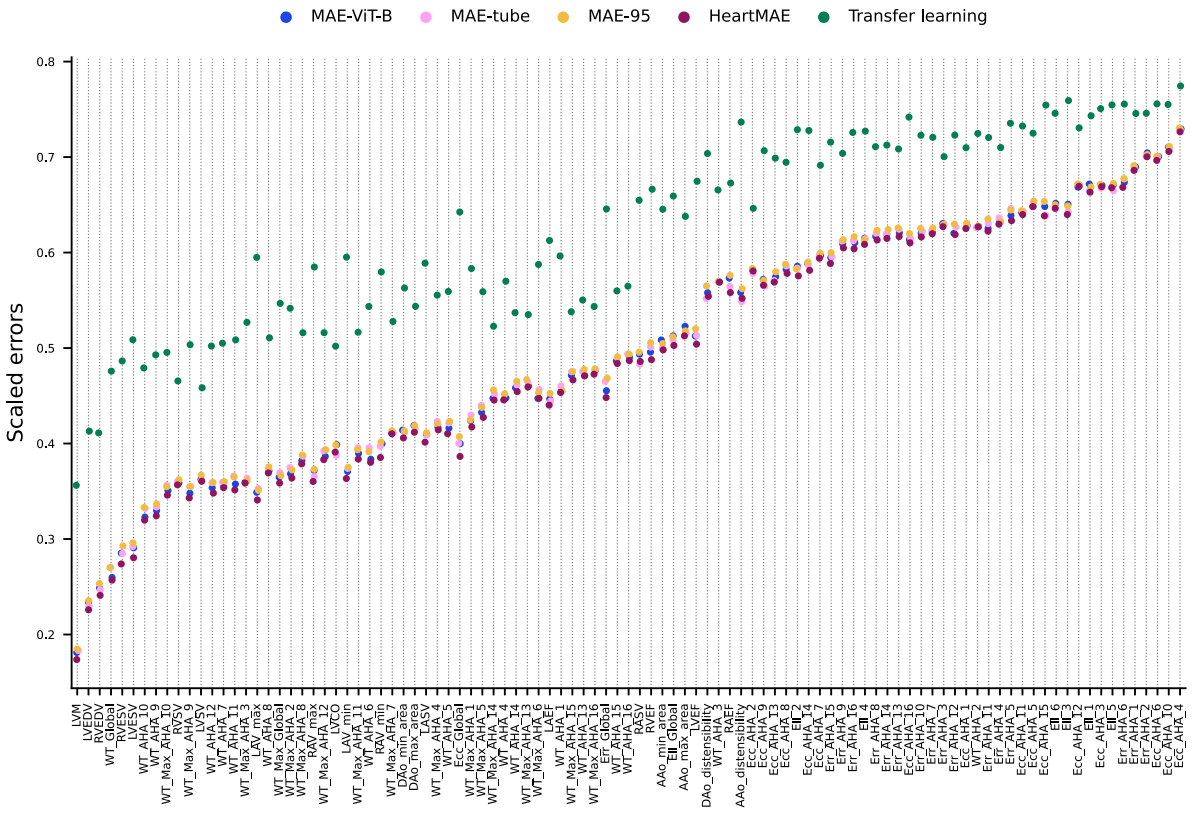

Figure S1: Prediction errors on 99 cardiac feature prediction tasks. For each feature, mean absolute errors were scaled by the standard deviation. HeartMAE achieves the lowest errors in 86 tasks.

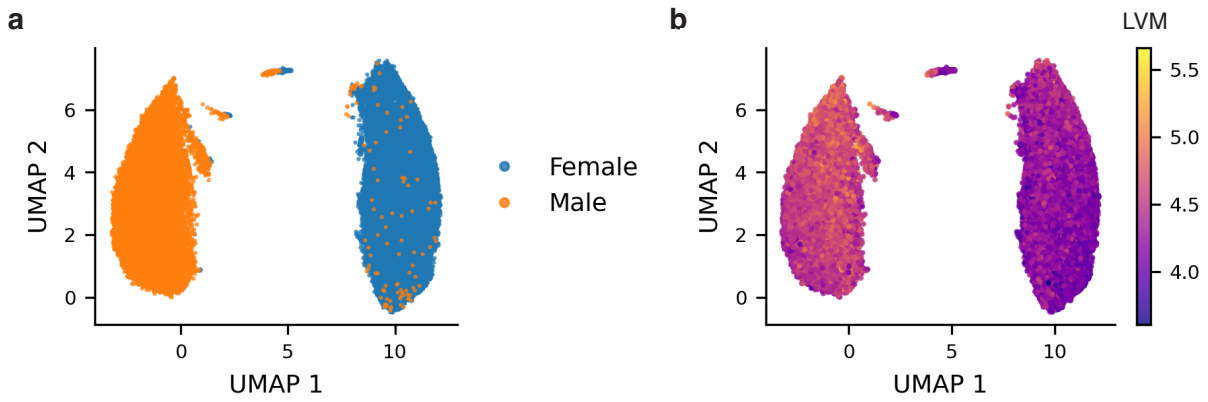

Figure S2: UMAP (McInnes et al., 2020) visualization of UKB subjects based on HeartMAE features. Each point represents a subject colored by a) biological sex b) left ventricular mass (LVM).

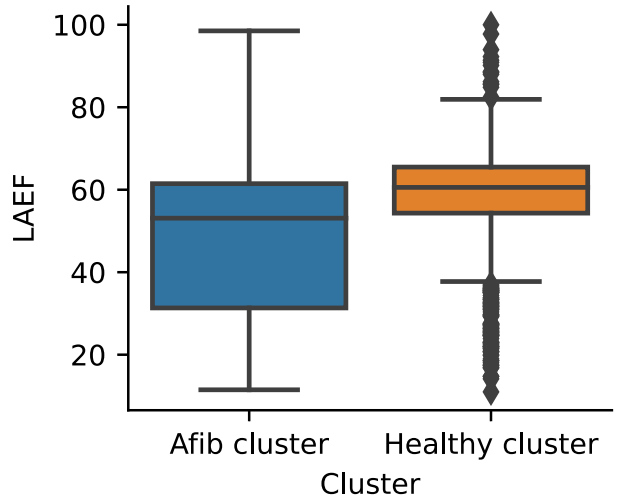

Figure S3: Left atrial ejection fraction (LAEF) distribution across the two clusters identified in Figure 3b. The comparison reveals a significant difference (Mann-Whitney U test $p = 9 \cdot 10^{-54}$).

Appendix A. MAE Training Details

This section provides full training details for all masked autoencoder (MAE) models used in our study. We present the information in 4 subsections: common configurations, model-specific hyperparameters, data preprocessing and hardware details.

Common Configurations

For all MAE models, we used the base vision transformer (ViT-B) architecture, adapted for spatiotemporal data. All models shared several common configurations. They used a patch size of $p \times p \times t = 16 \times 16 \times 4$ and a base learning rate of 1.5×10^{-4}. A learning rate scaling rule was applied: $lr = blr \times \text{batch size}/256$. We used a 30-epoch linear warm-up period for all models. A constant weight decay value of 0.05 was maintained during training. The batch size was set to 56, with repeated sampling of 4 clips from each CMR video, resulting in an effective batch size of 224. All MAE models received input clips of 24 consecutive frames extracted from each CMR video. Following Feichtenhofer et al. (2022), we used separable positional embeddings for space and time dimensions and reconstructed only 12 frames instead of the full video.

Model-Specific Hyperparameters

Table A1 summarizes the hyperparameters for MAE-ViT-B, MAE-tube, MAE-95, and HeartMAE. MAE-ViT-B was randomly initialized and trained for 200 epochs. MAE-tube used tube masking introduced by Tong et al. (2022). MAE-95 used random masking with a 95% mask ratio. HeartMAE implemented optical flow guided loss and random masking with a 90% mask ratio. MAE-tube, MAE-95, and HeartMAE were initialized with the best MAE-ViT-B checkpoint (epoch 150) and trained for 100 epochs.

The loss functions varied by model: MAE-ViT-B, MAE-tube, and MAE-95 used the MSE loss (Equation 1) in normalized pixel space, while HeartMAE employed a custom loss function (Equation 2).

Data Preprocessing and Augmentations

Prior to training, we preprocessed all CMR videos. Intensities were max-scaled to the range [0, 1], then centered and scaled using the training set mean and standard deviation. We implemented both temporal and spatial augmentations. Temporal augmentation consisted of selecting a random starting point from which a 24-frame clip was sampled. Spatial augmentations included horizontal and vertical flips along with random crop with resizing. Specifically, spatial crops with a minimum size of 224x224 and maximum size of 256x256 were sampled, with the short dimension (height or width) rescaled if needed to match the crop scale. This crop was then rescaled to 224x224. For detailed hyperparameter settings and model-specific configurations, refer to Table A1 on the next page.

Table A1: Overview of hyperparameters for MAE models. §For MAE-ViT-B initialization, we used the weights of the best checkpoint (epoch 150) of MAE-ViT-B evaluated on the task of ejection fraction prediction.

Hyperparameter	**MAE-ViT-B**	**MAE-tube**	**MAE-95**	**HeartMAE**
Common configurations				
Encoder architecture	ViT-B			
Spatial patch size	16×16			
Temporal patch size	4			
Base learning rate	$1.5 \cdot 10^{-4}$			
Batch size	224			
Number of sampled input frames	24			
Frames to reconstruct	12			
Model-specific configurations				
Initialization	random	MAE-ViT-B§	MAE-ViT-B§	MAE-ViT-B§
Training epochs	200	100	100	100
Masking strategy	random	tube	random	random
Mask ratio	90%	90%	95%	90%
Loss	MSE (1)	MSE (1)	MSE (1)	OF-MSE (2)