University of Applied Sciences and Arts Northwestern Switzerland
School of Computer Science

Fine-Tuning Pre-Trained ECG and CMR Encoders with a Modular Pipeline and Comprehensive Evaluation for Cardiovascular Diagnostics

Bachelor Thesis

Windisch, 21. March 2025

Authors
Dominik Filliger
Noah Leuenberger

Supervisors
Prof. Dr. Sarah Brüningk
Institute of Data Science I4DS
University of Applied Sciences Northwestern Switzerland

Prof. Dr. Arzu Çöltekin
Institute of Interactive Technologies IIT
University of Applied Sciences Northwestern Switzerland

Clients
Federal University of Rio de Janeiro
Instituto Nacional de Cardiologia, Rio de Janeiro

Abstract

Cardiovascular Diseases (CVDs) remain a leading cause of mortality worldwide, underscoring the need for reliable diagnostic methods. This thesis investigates how fine-tuning pre-trained Electrocardiogram (ECG) and Cardiovascular Magnetic Resonance Imaging (CMR) encoders, originally developed on predominantly healthy populations, can improve the identification and representation of pathological cardiac conditions. Three publicly available datasets are used—two ECG (Chapman and PTB-XL) and one CMR (ACDC)—to assess whether fine-tuning yields clearer disease-related clusters, how demographic information (age, sex) influences the embedding space, and whether these refined representations remain robust when tested on new data.

Quantitative results indicate that fine-tuning on pathology-rich datasets substantially increases silhouette scores, suggesting more coherent clustering of arrhythmias and cardiomyopathies. Qualitative embedding analyses reveal the formation of subclusters for conduction abnormalities while showing limited imprint of demographic factors. Testing with an additional ECG dataset (PTB-XL) demonstrates satisfactory cross-domain alignment, although the small size of the ACDC dataset constrains CMR conclusions. These findings highlight that domain-adaptive fine-tuning can guide large-scale encoders toward clinically meaningful discriminations, without overtly amplifying demographic biases. A flexible, reproducible deep learning framework developed as part of this work facilitates further experimentation, including multimodal fusion and validation on larger, more diverse cohorts. This approach offers a step toward robust transfer learning in cardiovascular imaging, enabling better generalization across patient populations and clinical environments.

Keywords: Cardiovascular Diseases, Deep Learning, Electrocardiography, Cardiac Magnetic Resonance, Domain Shift, Transfer Learning

Acknowledgment

We would like to extend our heartfelt appreciation to everyone who has contributed, directly or indirectly, to the successful completion of this thesis.

First and foremost, our gratitude goes to Özgün Turgut from the Technical University of Munich for clarifying questions about his work and the direct consultations we had with him, which greatly informed and shaped our methodology.

We are deeply thankful to Dr. Adriana Bastos Carvalho (MD, PhD) from the Federal University of Rio de Janeiro for her invaluable medical consultation, particularly regarding electrocardiography. We also extend our thanks to Lucas Araújo, whose expertise provided essential medical background on cardiovascular magnetic resonance imaging.

We are grateful for the opportunity provided by the Integrated Project Oriented Learning Environment (IPOLE) at the University of Applied Sciences and Arts Northwestern Switzerland (FHNW), an international and interdisciplinary platform fostering innovation through collaboration. Additionally, we gratefully acknowledge the Amazon Web Services (AWS) resources provided through IPOLE, which were instrumental in supporting our computational requirements. Moreover, we had the valuable experience of traveling to Brazil, where we further benefited from discussions with Dr. Carvalho's team at the Federal University of Rio de Janeiro and collaborated closely with experts at the Instituto Nacional de Cardiologia (INC) in Rio de Janeiro. Special thanks to Dr. Helena Cramer Veiga Rey (MD, PhD) from INC, who coordinated our trip and provided extensive assistance throughout our visit.

Special thanks are due to our supervisors: Leticia Fernandez Moguel, PhD, who guided us with dedication at the beginning of the project, Prof. Dr. Arzu Çöltekin and Dr. Sarah C. Brüningk for their unwavering support, continuous encouragement, constructive feedback, and insightful discussions throughout the entire process. We also gratefully acknowledge the use of the High-Performance Computing resources provided by the FHNW, which was essential for conducting our computational experiments.

Lastly, our deepest gratitude is reserved for our families and friends, whose unwavering support and encouragement have been invaluable, not only during this bachelor thesis but throughout the journey of our lives.

Dominik Filliger and Noah Leuenberger

Windisch, 21. March 2025

Contents

Glossary

1AVB First-Degree Atrioventricular Block

ACDC Automated Cardiac Diagnosis Challenge

AFIB Atrial Fibrillation

AFLT Atrial Flutter

AI Artificial Intelligence

APB Atrial Premature Beat

BCE Binary Cross-Entropy

CAVB Complete AV Block

cCMR cine-CMR

CLBBB Complete Left Bundle Branch Block

CMR Cardiovascular Magnetic Resonance Imaging

CRBBB Complete Right Bundle Branch Block

CVD Cardiovascular Disease

DCM Dilated Cardiomyopathy

DL Deep Learning

ECG Electrocardiogram

FHNW University of Applied Sciences Northwestern Switzerland

GSVT Generalized Supraventricular Tachycardia

HCM Hypertrophic Cardiomyopathy

INC Instituto Nacional de Cardiologia

LAD Left Axis Deviation

LAF Left Anterior Fascicular Block

LQV Low QRS Voltage

LVH Left Ventricular Hypertrophy

MINF Myocardial Infarction

NOR Normal

PACE Pacemaker Rhythm

PSVT Paroxysmal Supraventricular Tachycardia

QWA Q Wave Abnormality

RAD Right Axis Deviation

RVH Right Ventricular Hypertrophy

RV Right Ventricular Abnormality

SB Sinus Bradycardia

SR Sinus Rhythm

SSL Self-Supervised Learning

STACH Sinus Tachycardia

STTA ST-T Wave Abnormalities

SVTAC Supraventricular Tachycardia

UFRJ Federal University of Rio de Janeiro

UKBB UK Biobank

UMAP Uniform Manifold Approximation and Projection

VPB Ventricular Premature Beat

WPW Wolff-Parkinson-White Pattern

List of Figures

List of Tables

1 Introduction

1.1 Background & Motivation

Cardiovascular Diseases (CVDs) remain the leading cause of death worldwide, with heart-related conditions topping global mortality rankings in 2019 [1]. In clinical practice, both Electrocardiogram (ECG) and Cardiovascular Magnetic Resonance Imaging (CMR) are crucial modalities for diagnosing a variety of cardiovascular conditions. ECG is a non-invasive, widely available method that records the electrical activity of the heart, making it extremely valuable for detecting arrhythmias, ischemic changes, and other abnormalities [2]. ECG interpretation typically begins by classifying the fundamental heart rhythm, such as Atrial Fibrillation (AFIB), Generalized Supraventricular Tachycardia (GSVT), Pacemaker Rhythm (PACE), Sinus Bradycardia (SB), or Sinus Rhythm (SR) [3], forming a baseline from which additional diagnostic findings are identified. A comprehensive evaluation of morphological features on the basis of ECG requires solving the inverse problem of reconstructing cardiac electrical activity, which is inherently ill-posed [4], [5]. Meanwhile, CMR offers high-resolution structural and functional information about the heart, aiding in diagnosing conditions like myocarditis, cardiomyopathies, and congenital heart diseases. When combined, ECG and CMR can yield a comprehensive clinical view of the patient's cardiovascular status.

Deep learning has increasingly been applied to cardiovascular disease detection by learning representations for ECG and CMR data [6]–[8]. However, these deep learning approaches are often trained on a single large dataset in a process known as pre-training. Pre-training on large datasets like the UK Biobank (UKBB) [9] can produce models that extract robust features for ECG or CMR. However, the UKBB specifically predominantly comprises healthy, middle-aged individuals of European descent [10], raising concerns about whether the extracted features will perform effectively in more diverse clinical populations [11]. This problem of domain shift leading to poor generalization is discussed in several studies [12]–[14]. Additional variability also arises in real-world hospital environments due to differences in imaging devices, ECG lead configurations, and protocol standards [15], compounding the issue of domain shifts.

Transfer learning is a promising approach to bridge this gap between domains. Generally, this involves taking a pre-trained model and adapting it to a related, more specific task. This adaptation step, known as fine-tuning, helps the model specialize in a new domain in regards to task and data [16]. This approach has been successfully applied in various deep learning tasks [6], [17], [18]. Nonetheless, it is often only evaluated on the downstream task that was used for fine-tuning, without looking into the learned representations or their robustness under domain shifts.

In light of these limitations, the present work aims to investigate how fine-tuning affects learned feature representations from pre-trained ECG and CMR encoders, with a special emphasis on pathological cases. By examining the embedding space before and after fine-tuning, we seek to identify whether distinct clusters emerge for specific conditions, how demographic factors (such as sex or age) manifest in these embeddings, and whether fine-tuned representations remain coherent when transferred to a third dataset—thus reflecting their robustness under domain shift. Together, these questions shed light on the capacity of fine-tuned encoders to capture clinically meaningful patterns that may not be adequately represented by baseline (pre-trained) models, while also highlighting any risks of amplifying or obscuring certain demographic signals.

1.2 Research Questions

This leads to the following key research questions:

1. How fine-tuning changes the embedding structure concerning pathological cases, specifically:
   (a) Are any cluster formations emerging compared to the baseline (pre-trained) encoder?
   (b) Do new or more distinct patterns of pathological cases appear in the embedding space post-fine-tuning?
2. How demographic and population metadata manifest in the embedding space for both the baseline and the fine-tuned encoders:
   (a) Does the baseline encoder already exhibit visible clustering or patterns tied to these attributes?
   (b) Does fine-tuning amplify or reduce the influence of these parameters?
3. Examining the robustness of fine-tuned representations under domain shifts:
   (a) Does the fine-tuned embedding structure maintain its qualitative patterns when applied to a separate ("third") dataset?
   (b) How does this performance compare to the original pre-trained encoder when facing this domain shift?

1.3 Contributions

This study makes three contributions. First, it explores how fine-tuning pre-trained encoders reshapes the embedding space of pathological cases, contrasting them with baseline representations to reveal emergent clusters and qualitative patterns tied to disease.

Second, it assesses whether demographic parameters, such as age and sex, become more or less pronounced following fine-tuning on pathological labels, thereby examining potential biases that might have manifested in the learned features.

Third, it evaluates the robustness of fine-tuned models when confronted with domain shifts, including a separate third dataset, to determine whether these refined embeddings maintain coherence or revert to more generic representations.

Finally, this work is part of a broader collaborative effort between Instituto Nacional de Cardiologia (INC)[1], Federal University of Rio de Janeiro (UFRJ)[2], and University of Applied Sciences Northwestern Switzerland (FHNW)[3], aiming to develop robust unimodal representations for a multimodal Deep Learning (DL) pipeline. Beyond this study, the framework developed serves as a foundation for future research, including potential iterations of a multimodal DL pipeline for clinical applications. The codebase is publicly available[4].

1.4 Structure

The remainder of this document is organized as follows: Section 2 provides an overview of prior research on transfer learning, self-supervised representation learning, and domain generalization in medical AI. Section 3 outlines the methodological framework, including dataset descriptions, preprocessing pipelines, and the fine-tuning strategy for adapting pre-trained encoders. Section 4 presents the experimental findings, highlighting improvements in classification performance and embedding structure. Section 5 discusses the implications of these results, focusing on the

[1] https://inc.saude.gov.br/
[2] https://ufrj.br/en/
[3] https://www.fhnw.ch/en/
[4] https://github.com/IPOLE-BAT-CMR-ECG/dl-framework

influence of demographic factors and the models' robustness under domain shifts. Section 6 proposes avenues for future research, and Section 7 concludes with a summary of the study's main contributions.

2 Related Work

The following section provides an overview of prior research relevant to this work. First, existing studies on transfer learning in cardiovascular diagnostics are discussed, highlighting their relevance for model adaptation across different clinical settings. Then, Self-Supervised Learning (SSL) approaches for ECG and CMR representations are examined, as they serve as a foundation for robust feature extraction. The challenges posed by domain shifts in medical Artificial Intelligence (AI) models are then explored, focusing on both population-based and technical discrepancies. Finally, studies on fine-tuning for cross-population performance are reviewed, emphasizing its role in mitigating domain shift effects and improving model generalization.

2.1 Transfer Learning in Cardiovascular Diagnostics

Transfer learning has emerged as a key technique in medical AI, enabling models trained on large public datasets to be adapted to smaller, domain-specific clinical cohorts. Unlike traditional supervised learning, which requires vast amounts of labeled data, transfer learning allows pretrained models to retain and transfer knowledge across tasks, thereby improving performance in low-data regimes [19].

In the context of CVD diagnostics, transfer learning has been widely used to fine-tune models for ECG and CMR classification tasks. Studies show that pretraining on large-scale datasets provides feature extractors that can generalize well to different populations when fine-tuned [20], [21].

However, transfer learning in medical AI faces limitations, especially when models trained on highly homogeneous datasets are applied to diverse clinical settings. The UKBB [9], a widely used dataset for cardiovascular imaging research, predominantly consists of healthy, middle-aged individuals of European descent [10], leading to potential biases when applied to more heterogeneous populations [11]. Furthermore, the standardized data acquisition conditions in UKBB do not reflect the variability present in real-world hospital settings, where imaging and ECG data may exhibit greater noise, different lead configurations, or scanner inconsistencies [15]. These population-based and technical domain shifts must be addressed to ensure that transfer learning in medical AI remains effective across diverse settings.

2.2 Self-Supervised Learning for Feature Representation in Medical Imaging and Signal Processing

SSL has emerged as a promising alternative to fully supervised methods in medical AI, particularly for learning robust feature representations without requiring extensive labeled data [18], [22]. By leveraging pretext tasks, SSL models learn meaningful latent structures from unlabeled data, enabling more effective feature extraction for downstream tasks [18]. Commonly SSL uses loss functions that include contrastive losses that align different augmented views of the same data, reconstruction losses that aim to recover or reconstruct parts of the input from masked or corrupted data, and predictive losses that encourage models to predict representations based on contextual or temporal information [6], [23], [24]. Studies [25], [26] indicate that SSL-based encoders outperform traditional supervised models in low-data settings, making them particularly attractive for medical applications where labeled data is not always reliable or present[27].

In recent years, multimodal representation learning and contrastive learning techniques have advanced the state of the art by integrating information from heterogeneous data sources. For example, transformer-based architectures trained on large-scale datasets can learn shared latent representations for ECG and CMR, effectively enabling knowledge transfer across these modalities [7], [28].

A notable example of this approach is presented by [6], who proposed a two-stage process for multimodal self-supervised learning. In the first stage, ECG and CMR encoders are each pre-trained separately via unimodal SSL methods—masking and reconstructing ECG waveforms for the ECG encoder, and contrastive learning with image augmentations for the CMR encoder. In the second stage, the two encoders are jointly trained using a multimodal contrastive learning objective, aligning their representations in a shared latent space. This hierarchical pre-training strategy allows the ECG features to encode cardiovascular structural cues from CMR data, and vice versa, thereby enhancing cross-modal understanding. Their approach ultimately established a robust feature extraction paradigm that is relevant for a wide range of downstream classification tasks.

Despite these advances, generalization across large and diverse datasets remains an open challenge [29]. Even with SSL, embedding spaces may not fully capture pathological variability when shifting from well-curated datasets to more heterogeneous clinical data. However, recent work [21] on ECG-based SSL demonstrated that self-supervised encoders can achieve promising out-of-distribution performance, highlighting the potential of SSL-driven approaches for robust cross-population diagnostics.

Building on these findings, the two-stage multimodal SSL framework proposed by [6] forms a foundational element of the present study's methodology. The data preprocessing strategies, encoder architectures, and even aspects of the deep learning framework described in this study are heavily influenced by their open-source codebase [30] and training protocols.

2.3 Challenges of Domain Generalization in Medical AI

Domain generalization refers to enabling models to perform reliably on data distributions that differ from the training environment [29], [31]. In medical AI, such shifts stem largely from two sources: *population-based* variations and *technical-based* discrepancies [12], [13]. Population-based shifts include differences in demographic profiles (age, ethnicity, sex) and disease prevalence, which can cause models trained on narrowly sampled cohorts to misclassify underrepresented patient groups [12], [14], [21]. Technical-based shifts arise when imaging or signal acquisition protocols vary across hospitals or devices, leading to inconsistent input distributions that can confound model performance [12], [13].

The UKBB [9], which was used for pre-training of the encoders used in this work, exemplifies these challenges: although it provides extensive, high-quality cardiovascular data, its participants are primarily healthy, middle-aged, and of European descent, which limits representativeness for real-world clinical scenarios [11]. Additionally, UKBB's standardized ECG and CMR protocols contrast with the heterogeneity of acquisition methods found in routine clinical practice. As a result, models pre-trained on UKBB [6] may exhibit performance drops when faced with new populations or different scanning devices. These limitations underscore the importance of addressing domain shifts in medical AI, whether through multi-site training data, domain adaptation strategies, or fine-tuning on diverse cohorts to ensure robustness in real-world deployments [14], [21].

3 Methods

This section outlines the methodological framework adopted in this study. Section 3.1 introduces the three publicly available datasets used, highlighting their key characteristics and how they differ from the UKBB pre-training cohort. Section 3.2 details the modality-specific preprocessing pipelines for both ECG and CMR. Next, Section 3.3 explains the fine-tuning strategy employed to adapt the pre-trained encoders to downstream classification tasks. Section 3.4 specifies the performance metrics and qualitative embedding analyses used to assess model effectiveness. Finally, Section 3.5 describes the reproducible DL framework developed to standardize experimental configurations, streamline training, and support future extensions to additional modalities or datasets.

3.1 Datasets

This study selects three publicly available datasets to assess the generalizability and adaptability of pre-trained encoders beyond the UKBB cohort [9] used for pre-training them in [6]. Given that these datasets originate from different institutions, patient populations, and clinical settings, they introduce a natural domain shift, allowing for a structured evaluation of model robustness across diverse settings.

The selected ECG datasets consist of 12-lead recordings sampled at $500Hz$ with a minimum duration of 10 seconds. Rhythm conditions are especially well represented in the chosen ECG datasets, which is fitting since rhythm is not only one of the most commonly assessed features but also a fundamental characteristic of cardiac function that forms the diagnostic foundation upon which additional findings are built [3].

In contrast, the CMR dataset provides a structural and functional view of the heart, focusing on anatomical characteristics rather than electrical activity. Specifically, it comprises short-axis cine-CMR (cCMR) capturing dynamic cardiac function.

These standardized input formats ensure compatibility with the pre-trained models [6] while enabling an assessment of their ability to adapt to variations in acquisition settings and cohort demographics. A summary of the datasets used in this study along with their characteristics is provided for ECG and CMR in Tables 3.1 and 3.2 respectively.

Dataset	# Samples	Age (Mean±SD)	Sex (M/F)	Country	Reference
Chapman	44,817	58±20	25,272/19,523 (56%/44%)	USA, China	[32], [33]
PTB-XL	18,869	60±17	9,229/9,640 (51%/49%)	Germany	[34], [35]

Table 3.1: Summary of ECG datasets used in this study, including modality, sample count, demographic information, country of origin, and references.

Dataset	# Samples	BMI (Mean±SD)	Country	Reference
ACDC	150	26.3±5.7	France	[36]

Table 3.2: Summary of the CMR dataset used in this study, including modality, sample count, demographic information, country of origin, and references.

All three datasets differ in their geographic origins, with Chapman sourced from the USA and China, PTB-XL from Germany, and Automated Cardiac Diagnosis Challenge (ACDC) from France. This diversity in origin introduces potential domain shifts due to variations in healthcare systems, clinical practices, and population characteristics. In terms of demographic comparability, the ECG datasets (Chapman and PTB-XL) report similar mean ages (58 ± 20 and

60 ± 17 years, respectively) and a relatively balanced sex distribution (56%/44% and 51%/49%, respectively), making them overall comparable.

A detailed description of each dataset, including its characteristics and data collection protocols, is provided in the following subsections.

3.1.1 Chapman

The Chapman dataset [32], [33] consists of 10-second ECG recordings from $45,152$ patients, sampled at $500Hz$. It was collected by Chapman University, located in Orange, California, USA; Shaoxing People's Hospital, situated in Shaoxing, Zhejiang, China; and Ningbo First Hospital, based in Ningbo, Zhejiang, China. It includes a broad range of arrhythmias and cardiovascular conditions.

To enhance clinical relevance, raw diagnostic labels were mapped following the taxonomy used by [8]. The original 63 unique codes were consolidated into 20 broader diagnostic groups based on the Minnesota Code Manual [37]. These groups were further categorized into higher-level pathological expressions, enabling a more structured and interpretable classification system. Instances without a clear mapping within this taxonomy (0.74% of all records) were excluded, resulting in a final dataset size of $44,817$ records.

Table 3.3 presents an overview of the consolidated label structure, grouping conditions into four primary categories: Amplitude, Duration, Morphology, and Rhythm. The Amplitude category includes conditions related to electrical axis deviations and hypertrophy, such as Right Axis Deviation (RAD), Left Axis Deviation (LAD), and Left Ventricular Hypertrophy (LVH). The Duration category accounts for abnormalities in conduction times, including First-Degree Atrioventricular Block (1AVB) and Atrial Premature Beat (APB). Morphology-based classifications cover structural waveform changes, such as Q Wave Abnormality (QWA) and ST-T Wave Abnormalities (STTA). Finally, the Rhythm category captures clinically significant arrhythmias, including AFIB, Complete AV Block (CAVB), and SB. Notably, over 58% of all cases include either AFIB or SB labels, underscoring the dataset's strong representation of rhythm disorders. In contrast to datasets that predominantly feature healthy rhythms, only 23.8% of records in this dataset exhibit a normal SR, while the remaining 76.2% presents various pathological cardiac conditions.

Most records contain one rhythm label (AFIB, CAVB, GSVT, SB, SR) and may contain additional diagnostic labels (as seen in Table 3.3). While Chapman considers PACE as part of the morphology group, the discussion with medical experts clarified that while PACE has morphological impacts it is acknowledged to make sense to be considered as a rhythm label.

The dataset comprises $25,272$ (56%) male and $19,523$ (44%) female participants, with 22 (0.05%) records missing this information. The age range spans from 4 to 98 years with a median age of 62 years (IQR=24 years). Plots detailing the sex and age distributions can be found in Appendix A.1.1.

Despite the dataset's size and diversity, it exhibits significant class imbalance across diagnostic groups. Certain conditions, such as STTA and SB, are highly prevalent, collectively representing a large portion of the dataset. In contrast, rare conditions such as Wolff-Parkinson-White Pattern (WPW) and CAVB are significantly underrepresented, comprising only 0.16% and 0.17% of cases, respectively. Plots detailing the label distribution can be found in Appendix A.2.

This imbalance is considered and counteracted during fine-tuning and detailed in Section 3.3. The dataset was stratified by labels and split into 70% training ($n = 31,371$), 15% validation ($n = 6,723$), and 15% test ($n = 6,723$) partitions, ensuring that each subset maintains a representative distribution of the labels.

Group	Abbr.	Full Name	Count (%)
Amplitude (6)	LAD	Left Axis Deviation	1545 (3.45%)
	LAF	Left Anterior Fascicular Block	380 (0.85%)
	LQV	Low QRS Voltage	1043 (2.33%)
	LVH	Left Ventricular Hypertrophy	6048 (13.49%)
	RAD	Right Axis Deviation	853 (1.90%)
	RVH	Right Ventricular Hypertrophy	110 (0.25%)
Duration (5)	1AVB	First-Degree Atrioventricular Block	1192 (2.66%)
	APB	Atrial Premature Beat	1386 (3.09%)
	CLBBB	Complete Left Bundle Branch Block	453 (1.01%)
	CRBBB	Complete Right Bundle Branch Block	1096 (2.45%)
	VPB	Ventricular Premature Beat	1504 (3.36%)
Morphology (4)	PACE	Pacemaker Rhythm	1182 (2.64%)
	QWA	Q Wave Abnormality	1063 (2.37%)
	STTA	ST-T Wave Abnormalities	17955 (40.06%)
	WPW	Wolff-Parkinson-White Pattern	72 (0.16%)
Rhythm (5)	AFIB	Atrial Fibrillation	9840 (21.96%)
	CAVB	Complete AV Block	76 (0.17%)
	GSVT	Generalized Supraventricular Tachycardia	8304 (18.53%)
	SB	Sinus Bradycardia	16559 (36.95%)
	SR	Sinus Rhythm	10675 (23.82%)

Table 3.3: Hierarchical overview of mapped Chapman dataset labels, with counts and percentages relative to the total dataset size after mapping ($n = 44,817$). As this is a multi-label dataset, instances may have multiple labels, so percentages exceed 100%. Counts adopted from [8].

3.1.2 PTB-XL

The PTB-XL dataset [34], [35] consists of 21,799 clinical 12-lead ECG recordings from 18,869 patients. It was collected by the Physikalisch-Technische Bundesanstalt (PTB), Germany, across 51 clinical sites using 11 different recording devices. Each recording is available in both $100Hz$ and $500Hz$ versions, but only the $500Hz$ subset is retained to ensure methodological consistency with [6]. All recordings have a minimum duration of 10 seconds and may have multiple diagnostic labels assigned by up to two cardiologists.

To prevent patient-level data leakage on evaluation, only one randomly selected recording per patient is included in the dataset. Since 2,127 (11.3%) of patients have multiple ECG recordings, this results in 2,930 fewer recordings, reducing the total number of recordings from 21,799 to 18,869 [35]. This ensures that the dataset reflects the prevalence of diagnoses at the patient level rather than being skewed by patients with multiple ECGs. It is acknowledged that due to the removal of repeated recordings from the same patient, label distributions are impacted.

There is a total of 71 raw diagnostic labels, which are further categorized into 24 diagnostic subgroups and organized into five higher-level categories [34]: Conduction Disturbances, Hypertrophy, Myocardial Infarction (MINF), ST-T Changes, and Normal. A detailed overview of the PTB-XL label structure is provided in Appendix A.2.2, where these subgroups and categories are detailed.

PTB-XL contains a broad spectrum of cardiac conditions, grouped into Conduction Disturbances, Hypertrophy, MINF, ST-T Changes, and Normal ECGs. These include abnormalities affecting the heart's electrical conduction, structural changes in the heart muscle, and signs of past or ongoing ischemia. Unlike Chapman, which focuses more on rhythm abnormalities and

axis deviations, PTB-XL includes a higher proportion of normal ECGs (46.03%) and more cases related to MINF and conduction disorders.

To enable meaningful comparisons between Chapman and PTB-XL, a standardized mapping of rhythm-related conditions is applied. Most PTB-XL rhythm labels (Appendix A.2) have a direct counterpart in Chapman, including SR, SB, PACE, and AFIB. PTB-XL's Atrial Flutter (AFLT) is merged with AFIB in Chapman, reflecting their shared electrophysiological mechanisms, while its multiple supraventricular tachycardia subtypes (Sinus Tachycardia (STACH), Supraventricular Tachycardia (SVTAC), Paroxysmal Supraventricular Tachycardia (PSVT)) are consolidated into Chapman's GSVT. Only labels that could be matched to one of the original 63 Chapman labels were considered and mapped in the same way as [8] following [37]. This results in excluding Trigeminal Pattern (TRIGU) and Bigeminal Pattern (BIGU). This approach ensures consistency across datasets and preserves clinical relevance. A breakdown of these mappings is provided in Appendix A.3.

The dataset includes $9,640$ (51%) female and $9,229$ (49%) male participants (no missing information), spanning an age range from 0 to 90 years and a median of 61 years (IQR=23 years). Age values greater than 89 are clipped to 90 per HIPAA guidelines[5]. Plots detailing the sex and age distributions can be found in Appendix A.2.1.

PTB-XL serves solely as a cross-domain evaluation set complimentary to the Chapman dataset used for fine-tuning. To facilitate visualization, a subset comprising 10% ($n = 1,887$) of the total recordings was selected, with stratified sampling applied to preserve the overall label distribution. This pragmatic choice ensures clearer and more interpretable visual representations, similar to the implicit approach taken in the Chapman partitioning. Quantitative metrics for PTB-XL are also calculated on this subset, to be comparable to their qualitative counterpart.

3.1.3 Automated Cardiac Diagnosis Challenge

The Automated Cardiac Diagnosis Challenge (ACDC) dataset [36] consists of 150 clinical CMR exams in the form of cine-CMRs (cCMRs), acquired from real-world clinical exams at the University Hospital of Dijon, France. Scans were acquired using 1.5 T (Siemens Area) and 3.0 T (Siemens Trio Tim) scanners, with 100 (67%) of exams performed at 1.5 T and 50 (33%) at 3.0 T. Differences in field strength impact image contrast and signal-to-noise ratio, which is acknowledged as a potential source of domain shift and variability in data quality [38], [39].

Each exam includes expert manual annotations and patient metadata, such as weight and height. Labels are assigned based on predefined clinical criteria, categorizing cases into Normal (NOR), MINF, Dilated Cardiomyopathy (DCM), Hypertrophic Cardiomyopathy (HCM), and Right Ventricular Abnormality (RV). These groups represent a mix of healthy subjects and various structural heart diseases, capturing conditions that affect ventricular size, wall motion, and overall cardiac function. A detailed breakdown of the diagnostic categories is provided in Table 3.4, and information on classification criteria is available in the dataset description[6].

While the full motion of the heart is captured, only the end-diastole and end-systole phase frames are annotated. These phases correspond to the moments when the ventricles are most filled and most contracted, respectively, and are used to assess ejection fraction, ventricular volume, and myocardial wall motion.

The dataset includes patient demographic metadata, with a median weight of 76 kg (IQR = 24.5 kg), a median height of 170 cm (IQR = 12 cm), and a median BMI of 25.77 (IQR = 6.51).

[5] `https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html`
[6] `https://www.creatis.insa-lyon.fr/Challenge/acdc/databasesClassification.html`

Group	Abbr.	Count (%)
Normal	NOR	30 (20%)
Myocardial Infarction	MINF	30 (20%)
Dilated Cardiomyopathy	DCM	30 (20%)
Hypertrophic Cardiomyopathy	HCM	30 (20%)
Right Ventricular Abnormality	RV	30 (20%)

Table 3.4: Summary of the five diagnostic groups in the ACDC dataset.

BMI values range from 15.43 to 59.52, reflecting a diverse patient population. Additional details on weight, height, and BMI distributions can be found in Appendix A.3.1.

The dataset was stratified by group, using them as the target labels for fine-tuning. A five-fold cross-validation approach ($k = 5$) was applied, with each fold being stratified to preserve the label distribution. Given the overall small sample size, k-fold cross-validation was used to maximize data utilization and improve model robustness. Since $k = 5$, each fold consisted of an 80% training ($n = 120$) and 20% validation ($n = 30$) split.

3.2 Preprocessing

The preprocessing of ECG and CMR follows the pipeline outlined by [6] to maintain consistency with their pre-trained encoder. Most steps and implementations are directly adopted from their open-source code [30]. The following sections summarize the adopted preprocessing steps and detail any modifications made. A visual summary of the preprocessing pipelines for both modalities is in Figure 3.1.

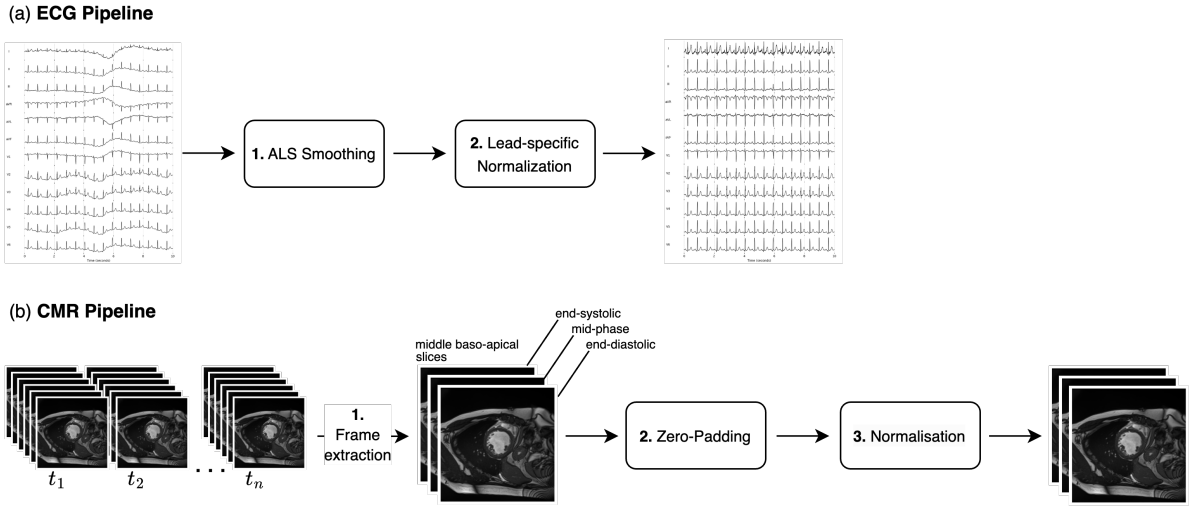

Figure 3.1: Overview of the preprocessing pipelines for (a) ECG and (b) CMR data, adapted from [6]. ECG preprocessing includes ASL smoothing [40] and lead-specific normalization [41], while CMR preprocessing consists of frame extraction, zero-padding, and normalization.

3.2.1 ECG

Baseline drift, caused by respiration or electrode instability, is removed using asymmetric least square smoothing [6], [40]. Lead-specific normalization is applied separately to the Einthoven

(I, II, III), Goldberger (aVR, aVL, aVF), and Wilson chest leads (V1-V6) following [6], [41], ensuring comparability across leads by standardizing amplitude variations.

As the encoder is pre-trained on fixed-length 10-second recordings, input signals must conform to this duration. While the Chapman dataset recordings are already of this length, the PTB-XL dataset recordings are of variable-length, some exceeding 10 seconds. To maintain consistency with the encoder's input configuration, these longer recordings are truncated. All signals are expected to be sampled at $500Hz$ and are neither downsampled nor upsampled as part of the preprocessing pipeline.

The final preprocessed data is represented as a tensor $\mathbf{X} \in \mathbb{R}^{12\times5000}$, where 12 corresponds to the number of leads and 5000 to the time steps (10 seconds at $500Hz$).

3.2.2 CMR

Each cCMR scan is processed into a three-channel, two-dimensional representation of the left and right ventricles, where the channels correspond to the end-systolic, end-diastolic, and mid-phase frames extracted from middle baso-apical slice [6].

The ACDC dataset, provides annotations only for the end-systolic and end-diastolic frames. This requires the mid-phase frame to be approximated to maintain a consistent three-frame input. The mid-phase is estimated as the frame equidistant time point between the other two frames. Variations in acquisition protocols mean that the true mid-phase may not always correspond precisely to this midpoint, introducing a degree of approximation. It is acknowledged that this methodological limitation affects the accuracy of the extracted mid-phase frame.

Moreover, the ACDC dataset does not provide explicit annotations for the middle baso-apical slice, it is approximated as the central slice along the basal-to-apical axis. This selection ensures that the extracted cross-section remains consistent across samples, but as the number of slices varies between acquisitions, this approximation introduces uncertainty regarding the exact anatomical positioning. It is acknowledged that this methodological limitation affects the accuracy of the extracted slice.

After all frames of the selected slice are extracted, images are zero-padded to 210×210 pixels and min-max normalized between 0 and 1 to ensure uniform input dimensions and intensity ranges. The original method described in [6] included a manual cropping step where images were cropped around the visible heart and resized back to the original size. This step was found to be applied in the pre-training setup of the accompanying codebase [30]; however, it was not used in their supervised training and evaluation-related code of the CMR approaches. To ensure consistency with the supervised setup, which is the focus of this study, the cropping step was omitted.

The resulting preprocessed data is structured as a tensor $\mathbf{X} \in \mathbb{R}^{3\times210\times210}$, where 3 corresponds to the selected temporal frames (end-systolic, end-diastolic, and mid-phase), and 210×210 represents the spatial resolution of the extracted middle baso-apical slice.

3.3 Fine-Tuning

Building upon the multimodal SSL framework by [6], this study leverages their pre-trained ECG and CMR encoder weights, originally trained via a two-stage SSL process, and fine-tunes them on downstream classification tasks.

The following sections describe the encoder architectures, modifications for downstream classification, and the training setup, including optimization strategies, data augmentation, and

hyperparameter tuning. Figure 3.2 provides an overview of the classifiers, highlighting the classification components introduced in this study.

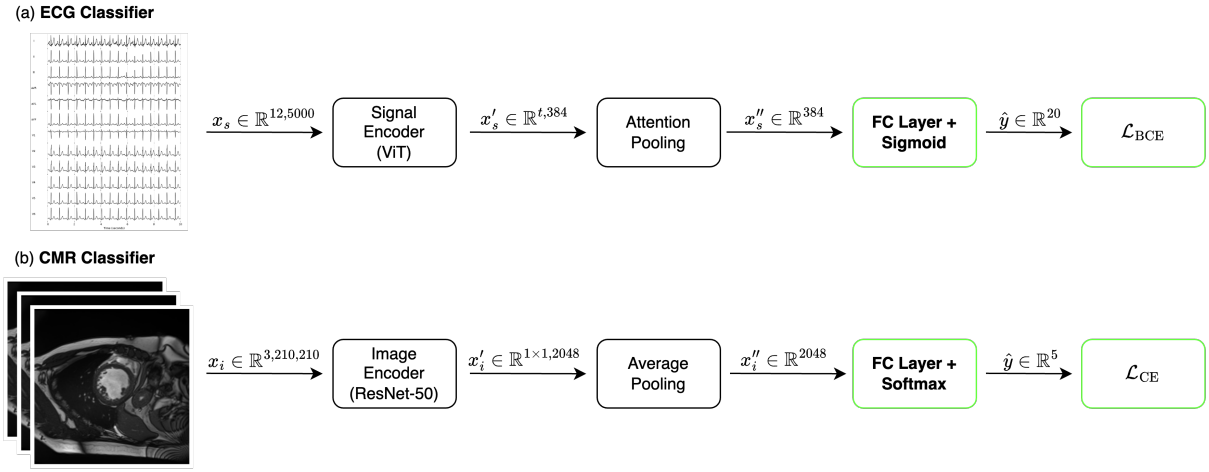

Figure 3.2: Overview of the classification components introduced in this study (highlighted in green). (a) The ECG classifier applies attention pooling followed by a fully connected layer with sigmoid activation, optimized with binary cross-entropy loss. (b) The CMR classifier uses average pooling and a fully connected layer with Softmax activation, trained with categorical cross-entropy loss.

3.3.1 ECG Classifier Model

The pre-trained ECG encoder is based on the Vision Transformer architecture [42] architecture. Initially, raw ECG signals are tokenized via a convolutional layer producing embeddings of dimension $\mathbb{R}^{384}$. These embeddings pass through stacked transformer encoder blocks, maintaining the embedding dimensionality. Attention pooling, as employed in the supervised setup of [6], aggregates these embeddings into a fixed-sized vector of dimension $\mathbb{R}^{384}$.

This study then adds a fully connected classification head consisting of a dense layer reducing dimensionality to $\mathbb{R}^{20}$, followed by sigmoid activation for multi-label classification on the Chapman dataset.

The model is trained using Binary Cross-Entropy (BCE) loss using a standard train-validation-test split, as detailed in Section 3.1.1. Furthermore, the training process is optimized using the AdamW optimizer [43]. All model weights are kept trainable throughout the fine-tuning process, enabling the encoder representation to adjust to the downstream classification task. The learning rate is set to 3×10^{-6} with a weight decay of 0.05 and a layer-wise learning rate decay of 0.75. The model undergoes a 20-epoch warm-up phase before training continues for a total of 200 epochs.

Given the class imbalance present in the Chapman dataset, the BCE loss function is extended by incorporating positive class weights computed as the inverse of the positive sample ratio, scaled logarithmically. Specifically, for each class c, if p_c is the fraction of positive samples, the scaled positive sample ratio p_{cs} for that class is defined as seen in Equation 3.1. p_{cs} is then integrated into the final loss per sample and class as found in Equation 3.2 where $y_{n,c}$ is the binary ground-truth label for sample n and class c, with $y_{n,c} \in \{0, 1\}$. The function $\sigma(x_{n,c})$ represents the sigmoid activation applied to the model's output $x_{n,c}$. This way p_{cs} amplifies the gradient contribution from minority labels (underrepresented positive classes).

$$p_{cs} = \log\Big(1 + \frac{1 - p_c}{p_c}\Big) \quad (3.1) \qquad \mathcal{L}_{\mathrm{BCE}_{n,c}} = -\Big[p_{cs} y_{n,c} \log \sigma(x_{n,c}) + (1 - y_{n,c}) \log(1 - \sigma(x_{n,c}))\Big] \quad (3.2)$$

Each sample undergoes a sequence of data augmentations following the methodology outlined in [6]. This doesn't increase the number of samples but merely transforms them individually. During training, the augmentations are applied sequentially and include Fourier Transform-based surrogate generation [44] to modify phase information while retaining overall frequency characteristics, signal jittering to introduce small temporal shifts, and amplitude rescaling to adjust signal intensity. By introducing carefully controlled variations, these steps expose the model to a broader range of input scenarios. During validation, no augmentations are applied so that the input remains unchanged, providing an evaluation of the original, unaltered signals.

Detailed parameters used for fine-tuning can be found in Table C.1.

3.3.2 CMR Classifier Model

The pre-trained encoder is based on a ResNet-50 [45] architecture, which consists of a convolutional stem followed by four sequential residual stages. The subsequent stages employ bottleneck residual blocks that progressively expand the feature dimensions while reducing spatial resolution. The final stage outputs a representation in $\mathbb{R}^{2048}$, obtained via an average pooling layer.

To enable classification, this study adds a fully connected classification head. The head consists of a single linear layer that maps the $\mathbb{R}^{2048}$ feature vector directly to $\mathbb{R}^5$, followed by a softmax activation for multi-class classification on the ACDC dataset.

The model is fine-tuned using the cross-entropy loss function (CE). The training process is optimized using the AdamW optimizer [43]. All model weights remain trainable during the fine-tuning process, allowing the encoder representation to adapt to the downstream classification task. A learning rate of 3×10^{-3} and a weight decay of 10^{-4} are used. Training spans 150 epochs, preceded by a 15-epoch warm-up phase to facilitate stable convergence.

Given the dataset size and structure outlined in Section 3.1.3, model training follows a cross-validation strategy, with the specific number of folds and resulting fold sizes detailed in that section. This approach ensures that all available data contributes to both training and evaluation while improving generalization and reducing sensitivity to individual splits.

Data augmentations are applied to each sample following the methodology outlined in [6]. This doesn't increase the number of samples but merely transforms them individually. During training, augmentations are applied sequentially with a probability of 0.95, including random horizontal flipping to introduce spatial variability, random rotations to improve orientation invariance, and random resized cropping to simulate scale variations while preserving key structures. Since the input consists of three-channel grayscale CMR frames, color jittering modifies intensity distributions by adjusting brightness, contrast, and saturation across frames. These transformations together enhance robustness in training the model. During validation, no augmentations are applied so that the input remains unchanged, allowing for a direct assessment of model performance on unmodified images.

Detailed parameters used for fine-tuning can be found in Table C.2.

3.3.3 Hyperparameter Optimization

Hyperparameter optimization was conducted using Tree-structured Parzen Estimator sampling, an adaptive search algorithm that efficiently explores the parameter space while prioritizing

promising configurations [46]. Different optimization objectives were used for the two classifiers: Macro-averaged F1-score was used for the ECG classifier due to its suitability for class imbalance in multi-label classification, whereas accuracy was selected for the CMR classifier given the balanced multi-class task.

For the ECG classifier, hyperparameter tuning was performed using a fixed validation split, while for the CMR classifier, evaluation during optimization was based on cross-validation, with metrics averaged across folds to ensure robust selection. Early stopping with a patience of 15 epochs was applied to terminate unpromising configurations early, reducing computational overhead.

3.3.4 Model Selection

Once hyperparameter optimization identified the best configurations, the final models were trained for a fixed number of epochs (200 for the ECG, 150 for the CMR classifier). Checkpointing was employed to periodically save model states throughout training, with checkpoints stored every 10 epochs. This allowed for a detailed analysis of how the encoder representations evolved, providing multiple points from which the final model could be selected. This approach also enabled an investigation of the encoder space beyond the apparent optimal point. Visualizations illustrating the progression of encoder representations throughout training are provided in Figure D.3.

In the case of the ECG classifier, the selection was based on validation performance at predefined checkpoint intervals, using an early stopping criterion with a patience of 1 epoch, evaluated every 10 epochs. Similarly, for the CMR classifier, validation performance was assessed at the same checkpoint intervals, with selection determined based on aggregated performance across the validation sets. In both cases, the selection criteria remained consistent with those used during hyperparameter optimization—macro-averaged F1-score for the ECG classifier and accuracy for the CMR classifier—ensuring coherence between optimization and final model selection.

The final selected models for evaluation correspond to the checkpoints at epoch 50 for the ECG classifier and epoch 80 for the CMR classifier.

3.4 Evaluation

This section outlines the procedure for evaluating the fine-tuned encoders. The approach encompasses both classification metrics and a qualitative review of the learned representations.

3.4.1 Label Structure

Given the background of the Chapman ECG dataset, in which records contain at least one rhythm label, the evaluation follows this natural organization. This is achieved by splitting the datasets into two groups: single-labeled rhythm and multi-labeled records. Single-labeled rhythm records are defined as records with exactly one rhythm label (AFIB, GSVT, PACE, SB, SR) and no additional diagnostic label for the remaining 15 labels of Chapman (Table 3.3). Multi-labeled is defined as records containing one rhythm label and at least one other diagnostic label. This is exclusively for both of the ECG datasets Chapman and PTB-XL, since the CMR dataset ACDC only contains records with one label where all labels (Table 3.4) are considered.

3.4.2 Metrics

Following fine-tuning, classification performance is assessed using task-appropriate evaluation metrics. For the multi-label ECG classification task (Chapman dataset), macro-averaged F1-

score is used as the primary metric. Due to class imbalances in the dataset, F1-score provides a balanced assessment by considering both precision and recall. Per-class F1-scores are also reported to provide a detailed breakdown of classification performance across labels. Metrics for the ECG model are computed on the holdout test partition.

For the multi-class CMR classification task (ACDC dataset), accuracy is used as the primary evaluation metric. Given the balanced nature of the dataset, accuracy serves as an interpretable and reliable measure of overall classification performance. Performance for the CMR classifier is evaluated by averaging accuracy across the five validation folds, with standard error reported to quantify variability across folds.

3.4.3 Silhouette Analysis

Silhouette scores [47] measure how similar each sample is to its assigned cluster compared to other clusters. The resulting score ranges from -1 to 1, where higher positive values indicate more distinct and cohesive clusters, while values near zero or negative suggest overlapping or poorly defined structures. Specifically,

$$\text{silhouette}(i) = \frac{b(i) - a(i)}{\max\{a(i),\, b(i)\}},$$

where $a(i)$ is the average distance from sample i to all other points in its cluster, and $b(i)$ is the minimum average distance from i to points in any other cluster. These scores are particularly suitable for data containing potentially overlapping diagnostic categories. They capture both intra-cluster compactness and inter-cluster separation, which is essential for validating whether the latent space meaningfully distinguishes pathological labels.

In this study, individual per-label silhouettes are also computed to assess how effectively each diagnostic or pathological category forms its own cohesive region within the high-dimensional space.

3.4.4 Centroid Distance Analysis

To further quantify how certain groups occupy the learned embedding space, the mean Euclidean distance between samples and their cluster centroid is calculated. This metric captures how tightly points are grouped around the conceptual "center" of each class. This distance-based perspective offers a straightforward way to evaluate how consistently each condition is represented. Comparing centroid distances before and after fine-tuning enables the detection of subtle structural shifts in the latent space, particularly under multi-label or partially overlapping scenarios.

Additionally, comparisons of centroid distances before and after fine-tuning illustrate whether the representations for each condition converge toward or diverge from their cluster centers. In some instances, sub-analyses distinguish single-labeled versus multi-labeled samples, revealing whether primary labels dominate embedding placement more strongly than secondary ones.

3.4.5 Qualitative Embedding Analysis

Final-Layer Embeddings and Dimensionality Reduction

To visually inspect the learned representations and assess their relationship with both classification labels and relevant patient metadata, final-layer embeddings from each model are extracted alongside patient level metadata (age and sex for ECG and BMI for CMR). This is done on

the test partition of the Chapman dataset for ECG and on the validation folds of the ACDC dataset for CMR.

These high-dimensional embeddings are reduced to two dimensions using the Uniform Manifold Approximation and Projection (UMAP) [48], enabling a qualitative appraisal of how samples cluster based on learned features as well as whether clusters correlate with external metadata. UMAP was chosen due to its ability to preserve both local and global structures within the data, making it well-suited for assessing feature separability and underlying patterns in high-dimensional representations [48]. UMAP hyperparameters were picked based on existing research using UMAP on the same datasets[7].

ECG Embeddings and Cross-Domain Visualization
For the multi-label ECG task, embeddings from the Chapman test partition and the PTB-XL subset are *jointly* combined to fit the UMAP transformation. This enables direct visualization of how PTB-XL recordings align within the same latent space as Chapman. Density contours highlight diagnostic clusters in the Chapman subset, while PTB-XL samples are overlaid with matching color coding for shared diagnostic labels. The resulting plots illustrate whether the fine-tuned ECG encoder can position the mapped conditions between the datasets close together, even across datasets with distinct demographic and acquisition characteristics.

3.5 Reproducible DL Framework

This study builds on the foundation established by [6], [30] to implement a reproducible DL framework. The framework addresses two key methodological challenges: first, ensuring exact reproducibility of experiments through comprehensive parameter tracking; second, facilitating rapid iteration on model architectures through modular design. The complete implementation, including technical documentation and configurations, is publicly available[8].

The framework architecture consists of modular components for data handling, model development, and evaluation, allowing for systematic investigation of projection methods. Key design decisions include the adoption of PyTorch (Version 2.4.0) as the core DL framework, with Lightning (Version 2.4.0) providing high-level abstractions for training workflows. Following [30], the framework utilizes their fork of the PyTorch Image Models (timm) library for computer vision models. These technologies were selected based on their established position in the research community, which ensures access to pre-trained weights and standardized implementations of state-of-the-art architectures.

Configuration management through Hydra (Version 1.3.2) ensures that all experimental parameters from data augmentation choices to model hyperparameters are explicitly defined and reproducible. This system stores configurations as hierarchical YAML files, enabling version control and systematic hyperparameter sweeps. Development standards are maintained through Python 3.12.5, ensuring compatibility across different environments.

To facilitate large-scale experimentation, the framework supports execution in both local and distributed computing environments. Bash scripts handle resource allocation and environment configuration for both SLURM clusters and Amazon SageMaker. Data versioning through Data Version Control (DVC) tracks dataset changes using checksums, ensuring that all experiments reference identical input data, further enhancing reproducibility.

[7] For ECG $n\_neighbors = 30$ and $min\_dist = 0.25$ as seen in [21] and CMR $n\_neighbors = 15$ and $min\_dist = 0.1$ as seen in [49].

[8] `https://github.com/IPOLE-BAT-CMR-ECG/dl-framework`

4 Results

4.1 ECG: Improved Diagnostic Specificity after Fine-Tuning

4.1.1 Baseline Model Embeddings Show Limited Diagnostic Separation

Figure 4.1 (left) shows the embedding structure derived from the baseline (pre-trained) model. Overall, these embeddings exhibit limited differentiation between diagnostic categories, as evidenced by Chapman's baseline silhouette score of 0.039 (Table 4.1). While some emerging patterns suggest partial recognition of pathological differences, classes overlap and appear to be all within one homogeneous cluster. For instance, SR (blue dot) is dispersed throughout the space, frequently overlapping with SB (yellow star), and AFIB (green square) intermingles with GSVT (orange rotated-square). PACE (pink x) forms a dense cluster with minimal spread. Notably, a distinct group of records appears to the left of the embedding space, but no clear label-associated differentiation is observed.

Dataset	Baseline	Fine-Tuned	% Improve
Chapman	0.039	0.210	436.75%
PTB-XL	-0.004	0.074	2120.17%

Table 4.1: Overall silhouette scores (baseline vs. fine-tuned) computed across rhythm labels (AFIB, GSVT, SB, SR, PACE). The % Improve column indicates the relative gain in silhouette, where higher values reflect stronger cluster separation.

4.1.2 Fine-Tuning Enhanced Arrhythmia Differentiation

After fine-tuning on Chapman, the embedding space shows visibly improved clustering (Figure 4.1, right). Numerically, Chapman's silhouette rose from 0.039 to 0.21 (+436.75%), signaling stronger within-label cohesion and between-label separation (Table 4.1). Rhythm-related diagnostic categories, such as AFIB, GSVT, PACE, SB, and SR, show more distinct and pronounced grouping, reflecting the encoder's enhanced ability to capture diagnostic-specific ECG features. This is further supported by per rhythm label silhouette scores (Table 4.2) which indicate improvement for all aforementioned labels except for PACE which is less pronounced [11.51%]. While SR and SB remain close (and AFIB and GSVT continue to overlap in some regions), each label achieves a more coherent cluster than in the baseline. PACE keeps its dense formation, isolated from other classes. The isolated group of mixed labels persists after fine-tuning showing no more apparent separation for these records.

Label	Baseline	Fine-Tuned	Change	%Change
AFIB	0.028	0.179	0.152	546.29%
GSVT	0.030	0.142	0.113	386.86%
PACE	0.086	0.096	0.01	11.51%
SB	0.068	0.276	0.208	303.57%
SR	0.010	0.195	0.185	1923.89%

Table 4.2: Mean per-label silhouette scores for Chapman under baseline and fine-tuned embeddings. %Change is relative to the baseline score. Each row indicates one of the rhythm labels.

Rhythm Labels Dominate Cluster Placement

Figures 4.2 (baseline) and 4.3 (fine-tuned) illustrate how single-labeled and multi-labeled recordings occupy the embedding space. In both models, the rhythm label largely determines cluster

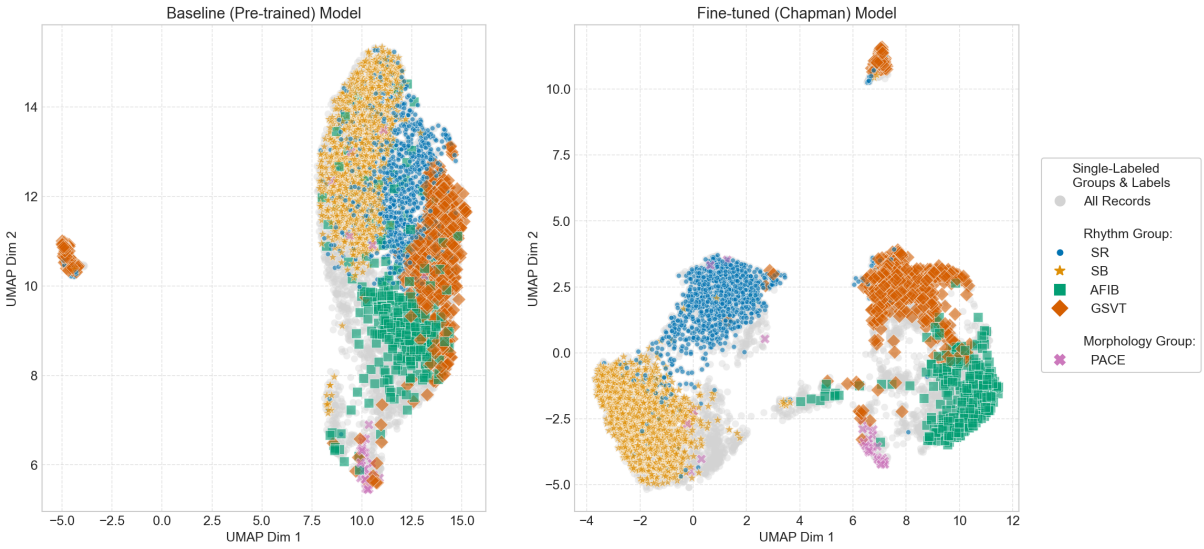

Figure 4.1: Comparative analysis of ECG embeddings before and after fine-tuning on the Chapman dataset. The left plot shows embeddings generated by the baseline (pre-trained) encoder, while the right plot demonstrates the fine-tuned encoder. All points are single-label rhythm.

membership, even when additional non-rhythm labels are present. For instance, multi-labeled AFIB samples still appear in the main AFIB cluster rather than dispersing toward other categories. This pattern remains evident after fine-tuning, indicating that rhythm abnormalities exert the strongest influence on how the model organizes ECG features.

Table 4.3 quantifies this tendency by comparing the mean distance to each label's centroid for single-labeled and multi-labeled recordings in the fine-tuned embedding space. Notably, AFIB cases show a 32.49% increase in average centroid distance for multi-labeled samples (16.868) relative to single-labeled ones (12.732). A similar pattern appears for GSVT, SB, and SR, where multi-labeled samples exceed a 10% gap. An exception is PACE, which exhibits a slight decrease (−2.48%) in centroid distance. Overall, the final row in Table 4.3 indicates that multi-labeled records lie approximately 20% farther from their respective centroids.

Label	Single-Labeled	Multi-Labeled	Ratio	% Diff
AFIB	12.732	16.868	1.325	+32.49%
GSVT	13.027	15.259	1.171	+17.13%
PACE	14.119	13.769	0.975	-2.48%
SB	10.693	12.097	1.131	+13.13%
SR	11.489	16.160	1.407	+40.65%
Mean	12.412	14.831	1.202	+20.19%

Table 4.3: Mean distances to centroids for fine-tuned Chapman embeddings, comparing single-labeled vs. multi-labeled samples for rhythm labels. Single-labeled distance is computed using only records with that label alone, while multi-labeled includes the label plus at least one other. The ratio is (Multi / Single) distance and % Diff is relative to the single-labeled distance. The last row shows the mean across all rhythm labels.

These qualitative observations align with the classification metrics on the Chapman test partition (Table 4.4). The fine-tuned model achieves high F1-scores on rhythm abnormalities (AFIB [0.91], GSVT [0.88], SB [0.94]) and maintains competitive scores even for less represented classes like

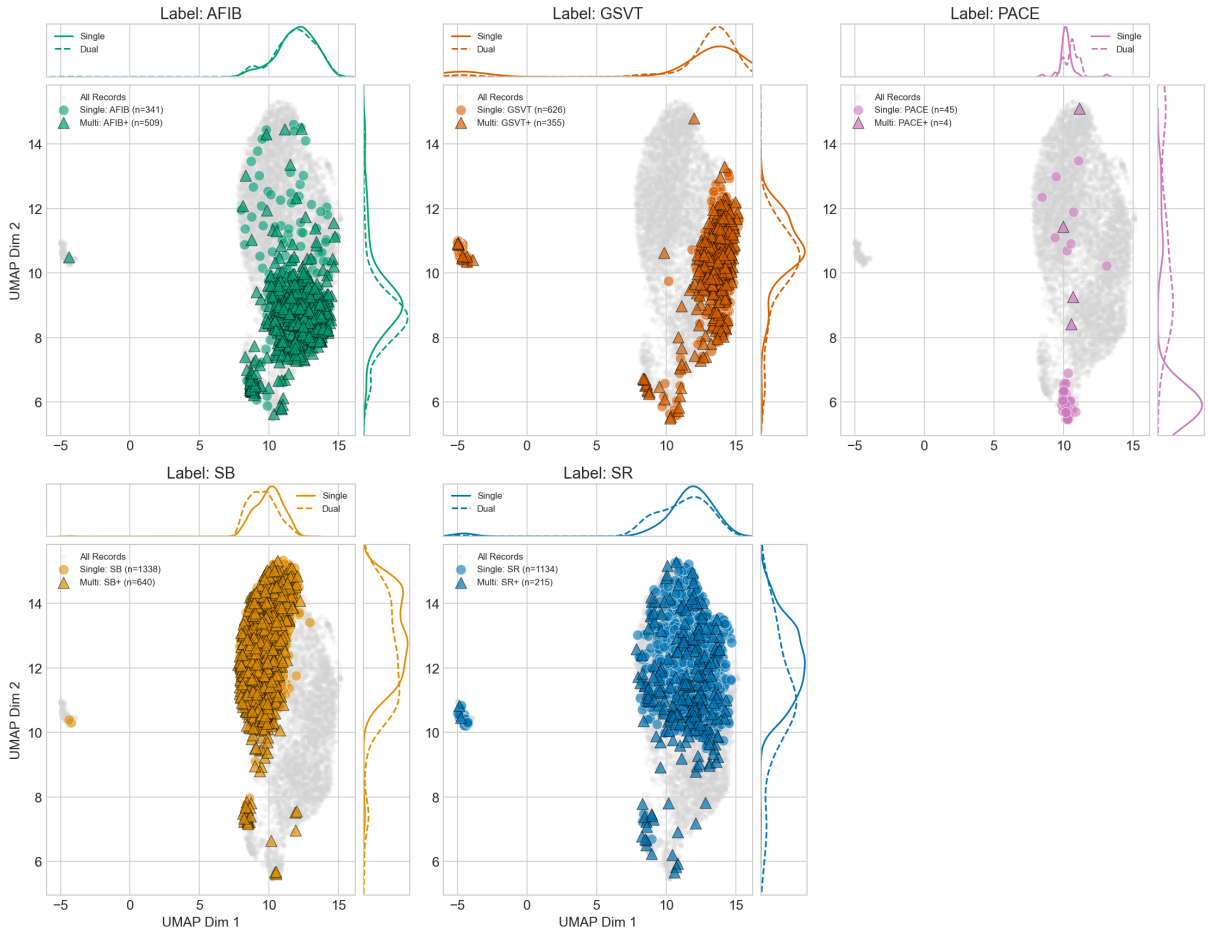

Figure 4.2: Baseline (pre-trained) embeddings showing single-labeled (circular markers) and multi-labeled (triangular markers) recordings. Multi-labeled samples are generally positioned near their corresponding single-label rhythm clusters, indicating that rhythm labels largely drive embedding structure.

PACE [0.72]. However, despite strong results for the aforementioned rhythm abnormalities, the macro-averaged F1-score across all (20) labels remains relatively weak [0.50]. Moreover, this evaluation is further skewed by the presence of low-prevalence labels with near-zero or non-existent F1-score, including WPW [0.00, 0.16%], CAVB [0.00, 0.17%], APB [0.01, 3.09%], and Low QRS Voltage (LQV) [0.09, 2.33%]. Detailed scores for all labels are provided in Appendix D.1.

Figure 4.4 illustrates the relationship between label prevalence and model performance. A positive correlation was observed between label frequency and F1-scores (Pearson's $r = 0.66$, $p = 0.002$, $n = 20$), with label prevalence accounting for 44% of the variance in performance ($R^2 = 0.44$). This indicates improved classification performance for more common classes such AFIB, SB, SR, while the moderate R^2 value suggests additional factors, such as distinctive patterns in rare classes such as PACE, influence model outcomes.

Emergent Sub-Cluster Near PACE Records

In the fine-tuned embedding space (Figure 4.1, right), a sub-cluster arises near the main PACE cluster (around $x = 4, y = -2.4$). Unlike the rhythm-driven groupings, these points do not align with any single rhythm label. Preliminary inspection suggests a higher prevalence of conduction abnormalities, such as Complete Right Bundle Branch Block (CRBBB) or Complete

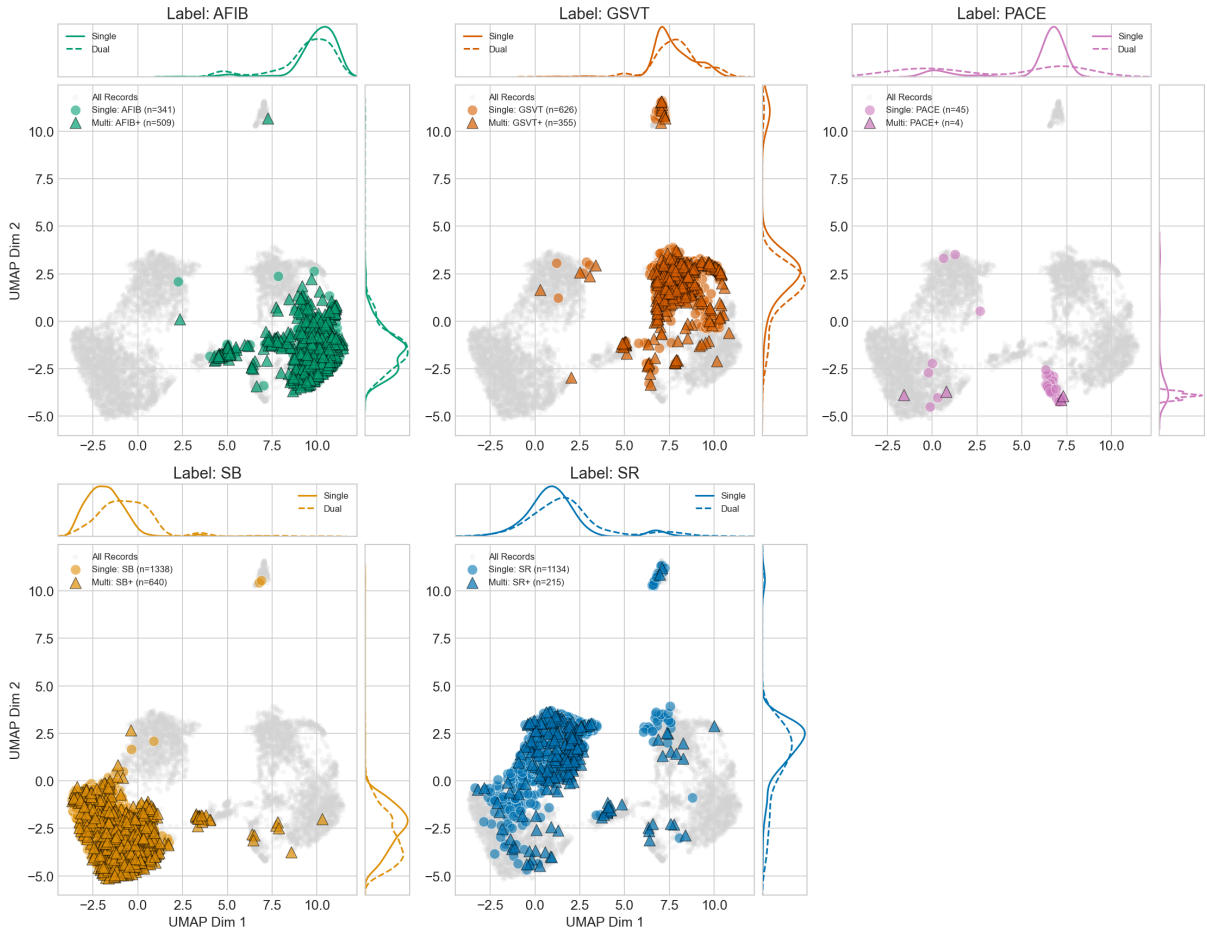

Figure 4.3: Fine-tuned (Chapman) embeddings showing single-labeled (circular markers) and multi-labeled (triangular markers) recordings. Multi-labeled samples are generally positioned near their corresponding single-label rhythm clusters, indicating that rhythm labels largely drive embedding structure.

Left Bundle Branch Block (CLBBB), within this sub-cluster. These conditions are characterized by prolonged QRS durations [50], [51], which may contribute to their separation from other rhythm-driven clusters. Additional faceted plots illustrating these co-occurrences are provided in Appendix D.1.1.

Metadata Analysis: No Clear Clustering by Sex or Age

A brief inspection of embeddings colored by sex or age revealed no globally distinct subgroups. Figures D.8 and D.6 (Appendix D.1.2) show that both male and female participants are interspersed throughout each rhythm cluster, with no clear sex-driven partitioning. Similarly seen in Figure D.9 and D.7, older and younger individuals appear in all parts of the latent space. However, labels such as AFIB do exhibit a slight skew toward older ages, suggesting that some clinical correlates manifest in the embedding to a limited extent. Overall, neither sex nor age forms a strong organizing principle comparable to rhythm labels.

4.1.3 Fine-Tuned Embeddings Generalized Robustly to New Domain

The out-of-distribution generalization is examined by overlapping PTB-XL embeddings onto the Chapman fine-tuned embedding space as shown in Figure 4.5. Quantitatively, PTB-XL's silhouette improved from -0.004 to 0.074, reflecting better label separation in the fine-tuned

Label	F1-Score	Recall	Precision
AFIB	0.91	0.94	0.89
GSVT	0.88	0.85	0.91
PACE	0.72	0.62	0.86
SB	**0.94**	**0.99**	0.90
SR	0.84	0.76	**0.94**

Table 4.4: F1-score, Recall, and Precision for rhythm labels on Chapman test partition after fine-tuning. Best-performing values are shown in bold. Detailed table of results across all labels are in Appendix D.1

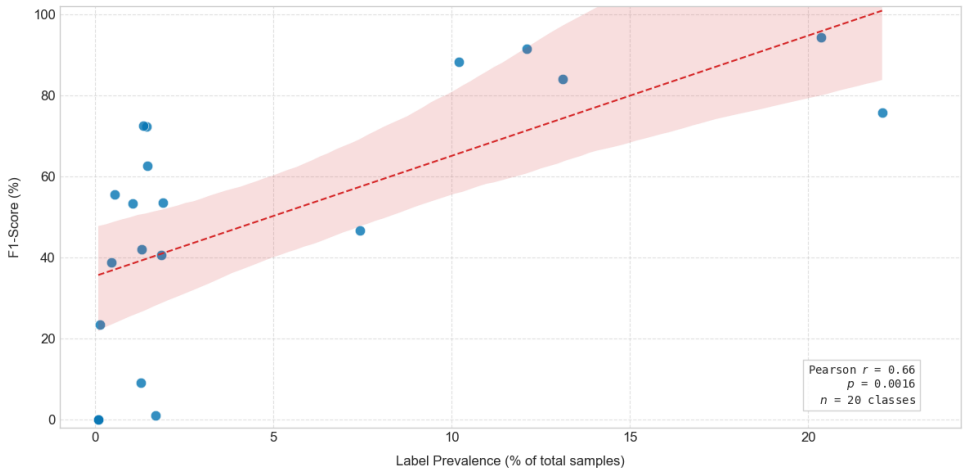

Figure 4.4: Relationship between label prevalence and F1-scores on the Chapman dataset test partition. The dashed regression line (95% confidence interval shaded) reflects the positive association ($r = 0.66$, $p = 0.002$). Analyzed classes: $n = 20$.

space (Table 4.1) across the rhythm labels (AFIB, GSVT, PACE, SB, SR). Looking at individual silhouette scores per rhythm label, just like with Chapman, there is a noticeable improvement except for PACE as seen in Table 4.6.

PTB-XL embeddings map closely with their corresponding Chapman clusters for AFIB, SR, SB, GSVT and PACE demonstrating strong cross-dataset alignment. PACE likewise remains densely grouped, indicating consistent recognition of pacing characteristics. By contrast, a visibly noticeable mix-up between SR and SB points suggests some overlap in how these sinus rhythms are labeled or manifest. This is further supported by looking into at the centroid of SB and SR of the PTB-XL embeddings: 21.62% of **sbr!** (**sbr!**) points are closer to SB centroid than to their SR centroid. However, no SB points (0.00%) are closer to the SR centroid than to their own SB centroid.

Table 4.5 highlights how PTB-XL's average distance to each Chapman label centroid changed after fine-tuning. SB records, for example, moved slightly *closer* (−2.9%) to their Chapman centroid, suggesting tighter alignment post-fine-tuning between datasets. By contrast, PACE exhibited a +19.6%) distance increase, indicating that PACE points shifted in how they map relative to Chapman's baseline centroid.

Additionally, none of the PTB-XL samples appear in the isolated cluster observed in Chapman (bottom right for baseline, top right for fine-tuned), hinting that the features fundamental to

this group may be absent in PTB-XL data and could be due to a dataset-specific factor of Chapman.

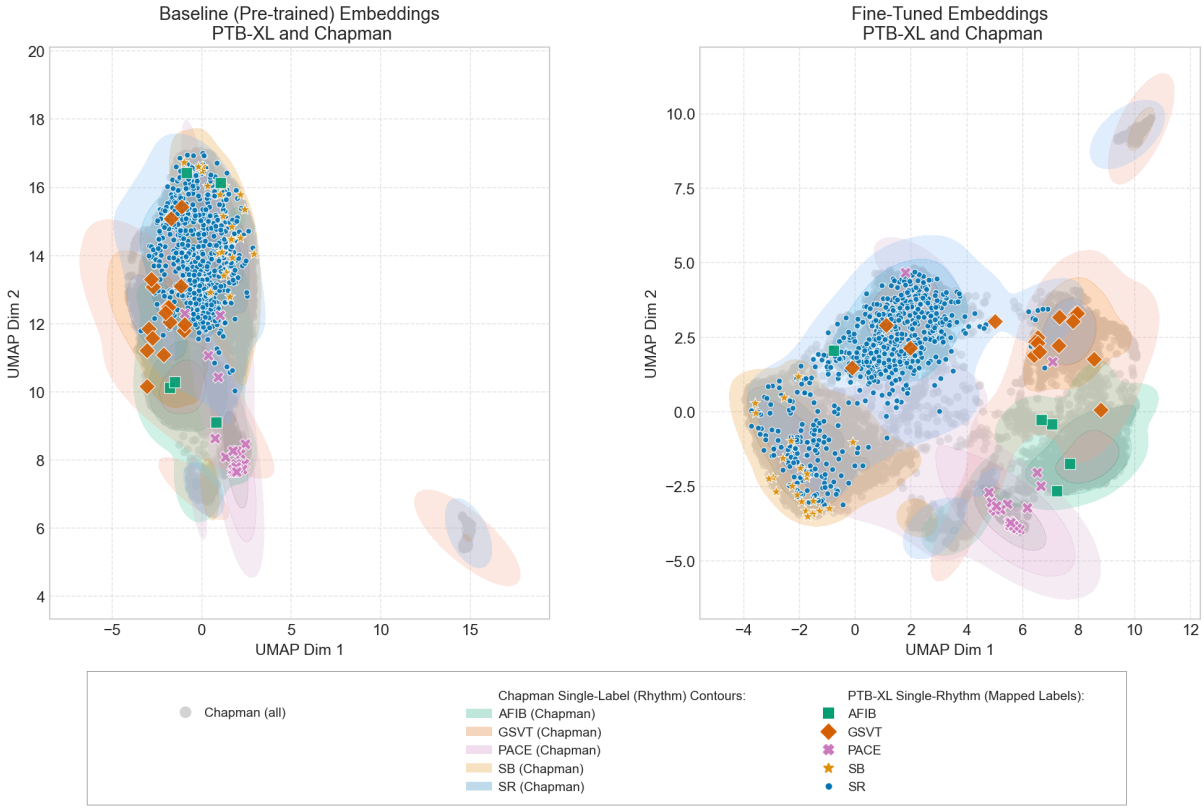

Figure 4.5: Cross-domain visualization of ECG embeddings derived from the Chapman fine-tuned encoder, overlaid with PTB-XL dataset embeddings. Density contours represent rhythm clusters from Chapman, whereas markers indicate PTB-XL records projected onto the same embedding space.

Faceted comparisons for each rhythm label are available in Appendix D.1 and D.2, where separate UMAP plots highlight how PTB-XL points overlay the Chapman embeddings on a label-by-label basis. These plots include density curves along each axis, making it easier to identify where PTB-XL samples lie relative to Chapman clusters.

Label	Baseline	Fine-Tuned	%Change	n
AFIB	12.42	13.57	+9.3%	101
GSVT	12.89	14.31	+11.0%	82
PACE	10.92	13.07	+19.6%	23
SB	11.96	11.61	-2.9%	49
SR	12.92	13.84	+7.1%	1545

Table 4.5: Class-centric distances from PTB-XL samples to each rhythm label's centroid in the Chapman embedding (baseline vs. fine-tuned). Each row shows the average Euclidean distance before and after fine-tuning, along with the percentage change. The final column (n) indicates how many PTB-XL samples per label were measured.

Label	Baseline	Fine-Tuned	Change	%Change
AFIB	-0.001	0.119	0.119	13016.14%
GSVT	0.001	0.061	0.052	538.86%
PACE	0.198	0.223	0.032	16.41%
SB	0.092	0.224	0.133	144.50%
SR	-0.011	0.064	0.075	708.45%

Table 4.6: Mean per-label silhouette scores for PTB-XL under baseline and fine-tuned embeddings, restricted to the rhythm labels. Large changes in silhouette reflect stronger separation post-fine-tuning.

4.2 CMR: Unveils Emerging Pathological Structure

4.2.1 Limited Diagnostic Separation at Baseline

Figure 4.6 (left) illustrates the baseline (pre-trained) model's embeddings for fold 1. The analysis is conducted across all folds; however, for visual simplicity, only the first fold is presented here while the analysis is done for all folds. Notably, no fundamental visual differences are perceivable between folds when examining the embedding space. Full visualizations for all folds can be found in Appendix D.2. Overall, there is a weak separation (silhouette mean of -0.117 ± 0.110) among all categories, evidenced by negative or near-zero silhouettes such as HCM (-0.218 ± 0.121) and NOR (-0.054 ± 0.136) (Table 4.8). Likewise, the intra/inter-label distance ratios remain relatively high (e.g., HCM at 0.871 ± 0.121; Table 4.7), resulting in strong overlap between healthy (NOR) and pathological cases. This suggests that, in its initial state, the model does not yet capture meaningful relationships between conditions, leading to highly overlapping representations.

4.2.2 Emerging Pathological Clustering After Fine-Tuning

After fine-tuning, the model's latent space becomes notably more organized as seen in Figure 4.6 (right). There is a noticeable shift toward more defined clusters, indicating that fine-tuning helped encode more meaningful diagnostic features than present in the baseline. From a clustering standpoint, silhouette scores transition from negative (mean -0.117 ± 0.110) to positive or near-positive (mean 0.055 ± 0.106), indicating clearer separations. Meanwhile, the intra/inter-label ratio in Table 4.7 decreases across all conditions (e.g., NOR from 0.808 to 0.551), highlighting greater separation between healthy and pathological clusters.

Certain conditions, such as DCM (blue) and MINF (green), show visual overlap, yet they remain relatively separated from other groups. HCM seem to separate reasonably well showing minor overlap with RV and NOR cases. The overall increase in structural organization suggests that fine-tuning enhances the representation of pathological variation, resulting in a more informative latent space both qualitatively and quantitatively.

The model achieves a final accuracy of 0.63 ± 0.046, reflecting moderate performance in distinguishing between diagnostic categories. The reported value represents the mean accuracy across the five validation folds, with the standard error indicating variability.

Looking at the clustering based on BMI, no strong pattern is perceivable. To further analyze potential trends, BMI values were categorized into four groups: underweight (BMI < 18.5), normal weight ($18.5 \leq$ BMI < 25), overweight ($25 \leq$ BMI < 30), and obese (BMI ≥ 30). This classification was applied to the dataset and used to color-code the embeddings. Notably, NOR

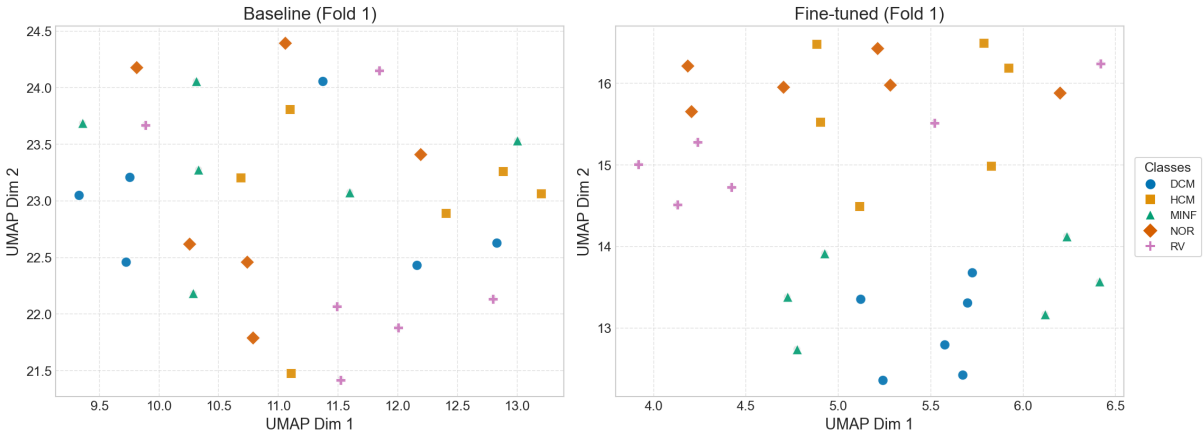

Figure 4.6: Comparative analysis of CMR embeddings before and after fine-tuning on the ACDC dataset (Fold 1). The left plot shows embeddings generated by the baseline (pre-trained) encoder, while the right plot demonstrates the same embeddings of the fine-tuned encoder.

Label	Baseline Ratio	Fine-Tuned Ratio	Ratio Change	Ratio Change %
DCM	0.772 ± 0.097	0.510 ± 0.022	0.262 ± 0.088	33.10%
HCM	0.871 ± 0.121	0.513 ± 0.080	0.358 ± 0.116	40.52%
MINF	0.820 ± 0.060	0.544 ± 0.075	0.276 ± 0.088	33.42%
NOR	0.808 ± 0.051	0.551 ± 0.031	0.258 ± 0.036	31.78%
RV	0.802 ± 0.056	0.606 ± 0.076	0.196 ± 0.054	24.61%

Table 4.7: Intra/Inter-Label Distance Ratio Comparison (Mean ± SD)

Label	Baseline (Mean ± Std)	Fine-Tuned (Mean ± Std)	Change	Change (%)
DCM	-0.050 ± 0.125	0.071 ± 0.076	0.121	420.51%
HCM	-0.218 ± 0.121	0.122 ± 0.117	0.339	325.83%
MINF	-0.143 ± 0.110	0.036 ± 0.102	0.179	113.95%
NOR	-0.054 ± 0.136	0.049 ± 0.119	0.103	250.76%
RV	-0.120 ± 0.059	-0.002 ± 0.115	0.118	159.43%
Mean	-0.117 ± 0.110	0.055 ± 0.106	0.172	147.18%

Table 4.8: Per-label silhouette scores (mean ± std) for the baseline and fine-tuned CMR embeddings, along with the absolute change and relative improvement (%). Higher silhouette values indicate clearer label separations.

cases seem to be mostly non-obese, but given the limited sample size, this observation remains inconclusive.

5 Discussion & Limitations

This section examines how fine-tuning impacts the learned representations of pre-trained ECG and CMR encoders and addresses the research questions outlined in Section 1. Three primary concerns are discussed: (1) the effect of fine-tuning on emergent clusters of pathological cases,

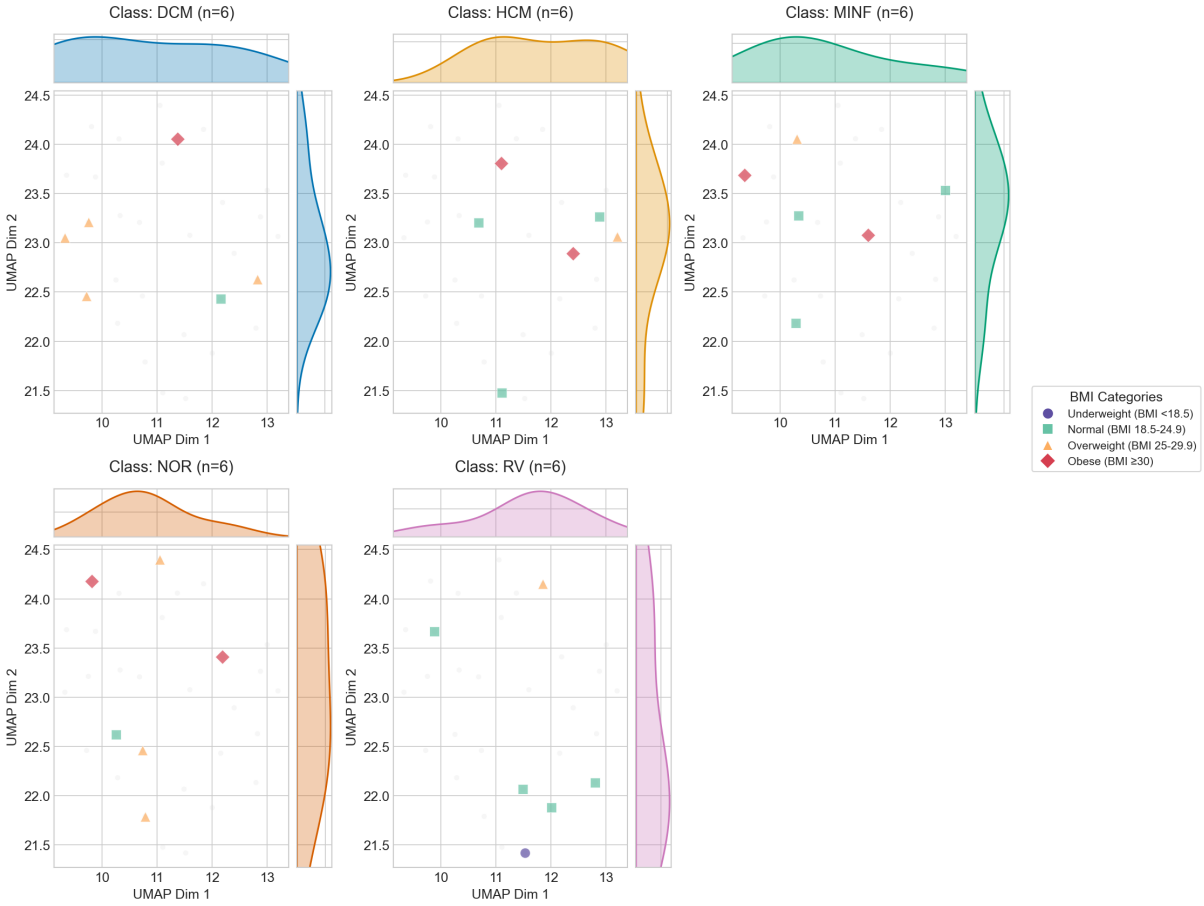

Figure 4.7: Baseline (pre-trained) CMR embeddings for fold 1, color-symbol-coded by BMI category.

(2) the extent to which demographic factors manifest in the embedding space, and (3) the robustness of fine-tuned representations under domain shift.

5.1 Fine-Tuning and Emergent Pathological Clusters (RQ1)

Improved Arrhythmia Separation in ECG

The results (Section 4.1) demonstrated that fine-tuning on a pathology-rich dataset (Chapman) noticeably improved the cluster separability for rhythm labels (AFIB, GSVT, PACE, SB and SR). This outcome indicates that representations learned on a primarily healthy UKBB cohort can be successfully adapted to highlight disease-specific features, in line with prior work on transfer learning for small domain shifts [6], [21], [29].

Moreover, the achieved F1-scores for rhythm labels are comparable to those reported in [8], which used the same dataset and label definitions. While the overall results across all labels are weaker than those reported in the single-model approach of [8], it is important to note that their study employed a different model architecture.

A further limitation of this study is the lack of a baseline comparison using linear probing, which is commonly used in self-supervised settings to assess the quality of learned representations [52]–[55]. This would have allowed for a direct assessment of the baseline embeddings in terms of classification performance.

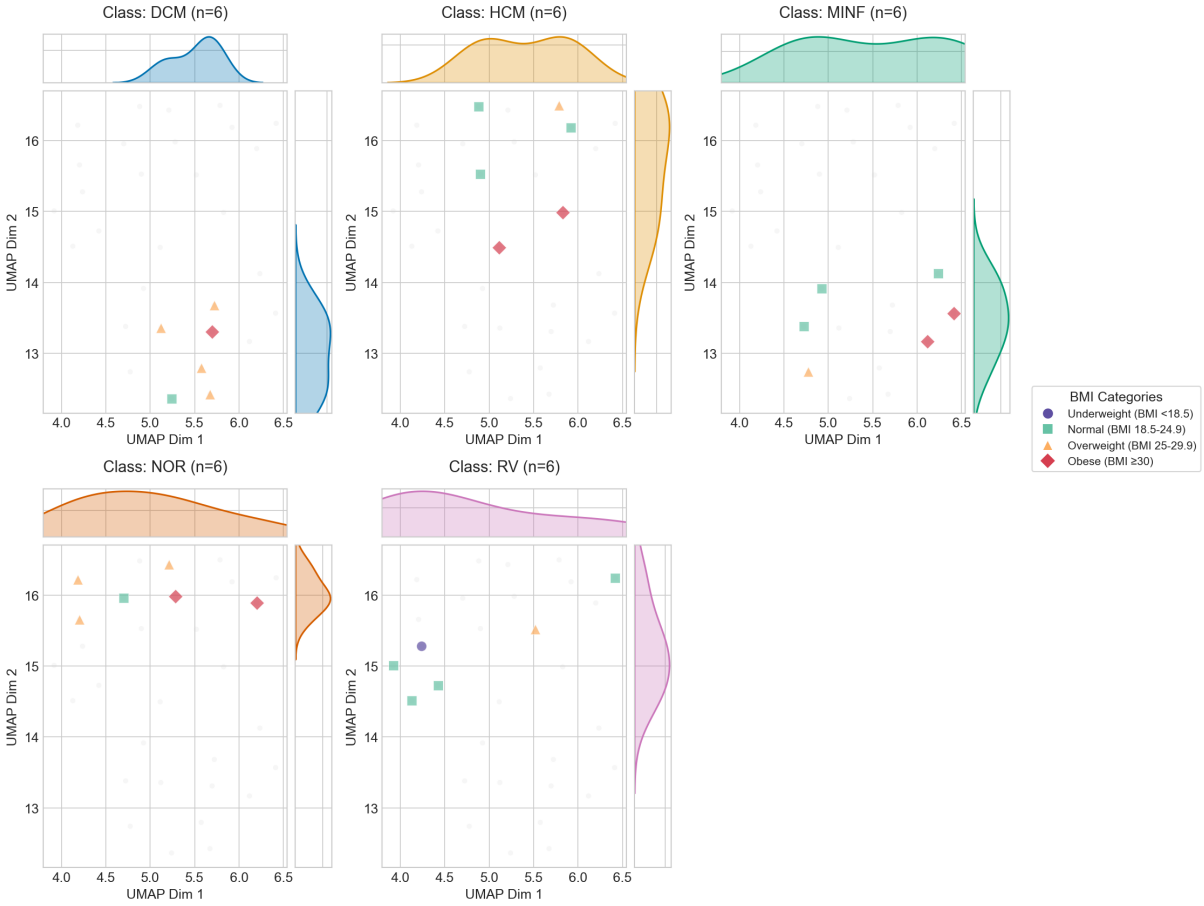

Figure 4.8: Fine-tuned CMR embeddings for Fold 1, color-coded by BMI category.

Sub-Clusters of Conduction Abnormalities

In addition to the primary rhythm-driven clusters, a smaller sub-cluster emerged near the main PACE region (Section 4.1 Figure 4.1). Preliminary inspection suggested that conduction abnormalities (CRBBB, CLBBB) were prevalent within this sub-cluster. It is plausible that these conditions as characterized by prolonged QRS durations [50], [51] get captured by the encoder and are therefore distinctively represented in the latent space.

However, in the same Figure 4.1 a distinct cluster was also examined but did not yield any conclusive clinical or demographic correlates, suggesting that other latent factors may be influencing its formation. Since neither the conduction-based groupings nor the unexplained cluster were the main focus of this study, further investigation is warranted to determine the diagnostic relevance of these emergent structures, whether they stem from specific pathologies, borderline arrhythmic states, or entirely novel waveform features.

Pathological Differentiation in CMR

In parallel to the ECG findings, fine-tuning on the ACDC dataset (Section 4.2) produced a more distinct separation of pathological categories in the CMR encoder. While the baseline embeddings showed minimal differentiation, post-fine-tuning clusters emerged for MINF, RV, HCM, and, to a lesser degree, DCM. According to clinical feedback with medical experts, MINF and DCM can share certain morphological and functional similarities (e.g., reduced ejection fraction, ventricular remodeling), which aligns with the partial overlap observed in the embedding space. Meanwhile, HCM often presents distinctly thickened myocardium, explaining why it appears

more separate from MINF and DCM, yet still exhibits minor overlap in specific cases as clarified by medical experts. Although the ACDC dataset is limited to 150 scans, the transition from a primarily healthy feature space to one emphasizing structural abnormalities demonstrates the potential of fine-tuning for better capturing pathological variation. These observations are consistent with findings in other multi-disease imaging studies [6], although additional research with larger datasets is needed to confirm broader clinical applicability.

It is important to note that UMAP may not be ideal for datasets with fewer than 500 samples due to its reliance on approximate nearest neighbor algorithms and negative sampling, which can lead to suboptimal embeddings on the ACDC dataset with a sample size of 150 [48], posing a further limitation.

5.2 Demographic Influences on the Embeddings (RQ2)

Lack of Strong Age- or Sex-Driven Subclusters

Embeddings colored by age and sex (Appendix D.1.2) indicated no globally distinct clusters based on demographic metadata. Male and female participants were evenly distributed across rhythm clusters, and older versus younger individuals did not form separate subregions. However, some clinical conditions, such as AFIB, tended to skew older, reflecting real-world epidemiological trends [56], [57]. Hence, demographic factors did not strongly reorient the embedding; instead, pathological features remained the primary organizing principle. This outcome suggests that while pre-trained models can learn subtle demographic correlates, fine-tuning on pathology-rich data may prioritize disease-related waveform or imaging characteristics over basic demographic distinctions. However, given prior findings that ECGs can be used to predict age- and sex [58], it is not unexpected that these demographic signals become overshadowed when a model is fine-tuned primarily to distinguish pathological conditions.

Implications for Potential Biases

The limited imprint of demographic factors in the latent space can be beneficial if the primary diagnostic goal is to differentiate disease states rather than demographic differences. However, the possibility remains that subtle biases could emerge under other training regimes or for less frequent demographics, consistent with broader concerns in medical AI [11], [12]. Ongoing vigilance is necessary, especially as models are deployed in clinical settings with more diverse populations.

5.3 Robustness Under Domain Shifts (RQ3)

Cross-Domain Alignment in ECG

The PTB-XL projection (Section 4.1) showed that fine-tuned embeddings aligned PTB-XL samples with the corresponding rhythm clusters in Chapman. Notably, a distinct cluster in Chapman that lacked its PTB-XL counterpart suggests that certain dataset-specific factors—possibly instrumentation or labeling nuances—can shape localized embedding regions. Nevertheless, the model successfully transferred the learned arrhythmia-specific representation to a new clinical cohort, achieving consistent detection of key rhythm abnormalities such as AFIB and PACE. These results are consistent with research indicating that self-supervised or transfer-learned features can exhibit robust out-of-distribution performance when tasks and modalities overlap [21].

Moderate Domain Shifts in CMR

Although the CMR domain shift was less extensively tested on external datasets, the improvement on ACDC from a baseline pre-trained on largely healthy UKBB data suggests domain adaptability might be achievable through fine-tuning as shown by the improved organization in

the latent space Figure 4.6. The robustness of this adaptation is still questionable lacking the external dataset, analogous to PTB-XL for ECG. The ACDC dataset with a sample size of only 150 limits strong conclusions about generalizability to broader CMR populations. Nevertheless, the observed gains in pathological clustering suggest a pathway for more effective CMR fine-tuning if larger multi-disease cohorts become available.

Additional factors further intensify the domain shift challenge. Differences in scanner magnet strengths within ACDC introduce heterogeneity that may confound adaptation, as variations in field strength (T1.5 and T3.0) and imaging protocols can significantly influence image contrast and resolution [38], [39]. This influence is known to have an impact on performance as seen in [13]. Moreover, the approximated mid-phase frame, may not align with true physiological midpoints due to differences in acquisition protocols. Similarly, the approximated middle baso-apical slice with variations in slice counts across acquisitions introduces potential anatomical misalignment. While necessary for consistency, these approximations, along with scanner-induced variability, contribute to domain adaptation challenges and may impact model robustness and influence the results.

5.4 Extensibility of DL Pipeline for Multiple Modalities

While primarily validated on unimodal ECG and CMR tasks as part of this study, the established DL framework can easily accommodate additional data modalities. Researchers can introduce new preprocessing scripts, model architectures, or data-augmentation strategies as separate modules. Because each module is configured independently, new modalities or tasks do not require fundamental changes to the overall pipeline structure.

6 Future Work

Building on the insights gained from examining ECG and CMR encoders separately, several directions emerge for continued research and methodological refinement. First, extending the framework to multimodal fusion remains a particularly promising avenue: combining arrhythmic findings from ECG with anatomical insights from CMR could enhance the diagnostic specificity of deep learning models, as similarly suggested in [6]. The modular design of the proposed DL pipeline lends itself well to this task, facilitating a seamless integration of multiple data types into a unified representation.

Additionally, a closer examination of newly formed ECG subclusters could further reveal potential mechanistic or clinical correlates. In particular, several conduction block categories, including CLBBB and CRBBB, formed distinct groups separate from the primary rhythm labels (AFIB, GSVT, PACE, SB, SR), whereas others remained less clearly characterized. An in-depth analysis of waveform-level features could provide greater clarity on the factors driving these subclusters. Meanwhile, for CMR, it would be worthwhile to reexamine the slice-extraction strategy. Automated slice segmentation could be a promising method to extract slices without the need for human annotation [59]. By moving beyond a single middle slice and incorporating multi-slice or even volumetric data, future work could capture a more detailed depiction of the left and right ventricular structures. At the same time, a systematic assessment of device-related domain shifts—such as comparing scans from 1.5 T versus 3.0 T machines—might prove critical for robust model generalization across heterogeneous clinical environments.

Although the present CMR analysis demonstrated improvements on the ACDC set, the limited sample size (150 scans) constrains definitive conclusions. Repeating these experiments with larger cohorts and broader disease categories would allow for a more comprehensive characterization of the encoder's capabilities and a more precise understanding of how fine-tuning reshapes

representation spaces. In a similar vein, greater emphasis on domain shift analysis—similar to how ECG embeddings were tested on PTB-XL—would benefit the CMR domain. Evaluating the model on out-of-distribution CMR data from additional institutions or publicly available repositories would ascertain whether it maintains discriminative power in new settings.

To better quantify how much of the performance boost derives from the underlying representations, evaluating baseline models via linear probing is a potentially illuminating next step. With this approach, the encoder remains frozen while a classification head is trained, allowing direct comparisons between baseline and fine-tuned feature spaces. This is already widely used in the SSL context to assess the quality of learned representations [52]–[55]. Furthermore, training the entire model from scratch on these smaller, pathology-rich datasets, as performed in [8], [30], would help clarify whether transfer learning confers advantages not easily matched by purely supervised methods.

Lastly, the present study's framework is primed for exploring alternative ECG architectures. For instance, a combined ResNet-DenseNet structure with ResU blocks has shown promise for multi-label ECG classification [8], [60] and could be incorporated into this pipeline with minimal overhead. Considering rare labels or subtle waveform abnormalities might especially benefit from architectures that excel at capturing fine-grained nuances. In parallel, extending these efforts to the Brazilian ECG/CMR dataset, prepared in collaboration with the INC and UFRJ, would further evaluate the framework's adaptability. As soon as this data becomes available, researchers could examine region-specific cardiovascular traits and additional domain shifts introduced by a distinct clinical population.

7 Conclusion

This thesis investigated how fine-tuning pre-trained ECG and CMR encoders influences their ability to detect and represent pathological cardiac conditions. The study was motivated by the challenge of adapting models originally trained on predominantly healthy cohorts, such as the UKBB, to more clinically diverse data. Specific goals included determining whether fine-tuning yields clearer pathological clusters, assessing how demographic factors manifest within the embedding space, and evaluating the robustness of the refined embeddings to domain shifts.

Quantitative and qualitative showed that training on pathology-rich datasets improved cluster separation for conditions such as atrial fibrillation, sinus bradycardia, and various cardiomyopathies. More cohesive disease-related embeddings emerged, indicating that models pre-trained on healthy data can learn to highlight pathological features when re-exposed to clinical cases. Meanwhile, demographic attributes (age, sex) did not strongly reorganize the latent space, implying that disease-specific patterns took priority. Testing with a second ECG dataset (PTB-XL) demonstrated coherent cluster alignment under a moderate domain shift.

The main research questions focused on whether new pathological clusters appeared after fine-tuning, whether demographic signals were amplified or reduced and whether the adapted embeddings remained stable when transferred to a separate dataset. The findings showed that new, more distinct pathological groupings did indeed form, demographic signals did not become more dominant, and strong alignment of out-of-distribution ECG data pointed to reasonable cross-domain adaptability. A direct comparison for CMR was not performed due to the absence of a second external dataset, although improvements on ACDC suggest that further generalization may be possible if larger CMR cohorts become available.

The small size of the ACDC dataset constrained CMR analyses and the lack of linear probing baseline evaluations limited the understanding of how much the deeper representation updates alone contributed to performance gains. Employing multiple slices or volumetric CMR

data, incorporating domain-shift analyses akin to those used for ECG, and including more diverse populations would strengthen confidence in the findings. Examining additional real-world datasets—such as the planned Brazilian cohort—would also offer valuable insights into the generalizability of the models to broader clinical settings.

Overall, the experiments demonstrated that fine-tuning can guide pre-trained ECG and CMR models to focus on clinically meaningful disease patterns without unduly amplifying demographic factors. By clarifying how representation spaces reorganize under pathological labels and moderate domain shifts, this work provides a structured basis for adapting large-scale pre-trained cardiac encoders to more specialized medical contexts. The developed framework, featuring modular design and reproducible tooling, stands ready for future extensions to multimodal pipelines, more extensive datasets, and real-world clinical applications.

References

[1] S. Dattani, F. Spooner, H. Ritchie, and M. Roser, „Causes of Death“, en, *Our World in Data*, Sep. 2023. [Online]. Available: https://ourworldindata.org/causes-of-death (visited on 02/16/2025).

[2] K. C. Siontis, P. A. Noseworthy, Z. I. Attia, and P. A. Friedman, „Artificial intelligence-enhanced electrocardiography in cardiovascular disease management“, en, *Nature Reviews Cardiology*, vol. 18, no. 7, pp. 465–478, Jul. 2021, Publisher: Nature Publishing Group, ISSN: 1759-5010. DOI: 10.1038/s41569-020-00503-2. [Online]. Available: https://www.nature.com/articles/s41569-020-00503-2 (visited on 03/18/2025).

[3] J. Hampton and J. Hampton, *The ECG Made Easy*, en. Elsevier, 2019, Google-Books-ID: 2r5_tAEACAAJ, ISBN: 978-0-7020-7457-8.

[4] R. Gulrajani, „The forward and inverse problems of electrocardiography“, *IEEE Engineering in Medicine and Biology Magazine*, vol. 17, no. 5, pp. 84–101, Sep. 1998, Conference Name: IEEE Engineering in Medicine and Biology Magazine, ISSN: 1937-4186. DOI: 10.1109/51.715491. [Online]. Available: https://ieeexplore.ieee.org/document/715491 (visited on 03/21/2025).

[5] E. Schenone, A. Collin, and J.-F. Gerbeau, „Numerical simulation of electrocardiograms for full cardiac cycles in healthy and pathological conditions“, en, *International Journal for Numerical Methods in Biomedical Engineering*, vol. 32, no. 5, e02744, 2016, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cnm.2744, ISSN: 2040-7947. DOI: 10.1002/cnm.2744. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/cnm.2744 (visited on 03/21/2025).

[6] Ö. Turgut, P. Müller, P. Hager, *et al.*, „Unlocking the diagnostic potential of electrocardiograms through information transfer from cardiac magnetic resonance imaging“, en, *Medical Image Analysis*, vol. 101, p. 103 451, Jan. 2025, ISSN: 13618415. DOI: 10.1016/j.media.2024.103451. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1361841524003785 (visited on 01/11/2025).

[7] F. Girlanda, O. Demler, B. Menze, and N. Davoudi, *Enhancing Cardiovascular Disease Prediction through Multi-Modal Self-Supervised Learning*, arXiv:2411.05900 [cs], Nov. 2024. DOI: 10.48550/arXiv.2411.05900. [Online]. Available: http://arxiv.org/abs/2411.05900 (visited on 12/11/2024).

[8] J.-B. Cha, S.-R. Hwang, and Y.-C. Park, „Diagnosis-specific Multi-model Design for 12-lead ECG Multi-label Classification“, ko, *Journal of the Institute of Electronics and Information Engineers*, vol. 61, no. 8, pp. 39–46, Aug. 2024, ISSN: 2288-159X, 2287-5026. DOI: 10.5573/ieie.2024.61.8.39. [Online]. Available: http://www.dbpia.co.kr/Journal/ArticleDetail/NODE11947341 (visited on 02/17/2025).

[9] C. Sudlow, J. Gallacher, N. Allen, *et al.*, „UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age“, en, *PLOS Medicine*, vol. 12, no. 3, e1001779, Mar. 2015, Publisher: Public Library of Science, ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1001779. [Online]. Available: https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001779 (visited on 02/16/2025).

[10] A. Fry, T. J. Littlejohns, C. Sudlow, *et al.*, „Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population“, eng, *American Journal of Epidemiology*, vol. 186, no. 9, pp. 1026–1034, Nov. 2017, ISSN: 1476-6256. DOI: 10.1093/aje/kwx246.

[11] K. M. Keyes and D. Westreich, „UK Biobank, big data, and the consequences of non-representativeness“, *The Lancet*, vol. 393, no. 10178, p. 1297, Mar. 2019, ISSN: 0140-6736. DOI: 10.1016/S0140-6736(18)33067-8. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0140673618330678 (visited on 03/02/2025).

[12] R. J. Chen, J. J. Wang, D. F. Williamson, *et al.*, „Algorithm fairness in artificial intelligence for medicine and healthcare“, *Nature biomedical engineering*, vol. 7, no. 6, pp. 719–742, Jun. 2023, ISSN: 2157-846X. DOI: 10.1038/s41551-023-01056-8. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10632090/ (visited on 03/09/2025).

[13] B. Guo, D. Lu, G. Szumel, *et al.*, *The Impact of Scanner Domain Shift on Deep Learning Performance in Medical Imaging: An Experimental Study*, en, arXiv:2409.04368 [eess], Oct. 2024. DOI: 10.48550/arXiv.2409.04368. [Online]. Available: http://arxiv.org/abs/2409.04368 (visited on 03/09/2025).

[14] D. Kaur, J. W. Hughes, A. J. Rogers, *et al.*, „Race, Sex, and Age Disparities in the Performance of ECG Deep Learning Models Predicting Heart Failure“, eng, *Circulation. Heart Failure*, vol. 17, no. 1, e010879, Jan. 2024, ISSN: 1941-3297. DOI: 10.1161/CIRCHEARTFAILURE.123.010879.

[15] Y. Li and Y. Fan, *Medical Image Segmentation with Domain Adaptation: A Survey*, arXiv:2311.01702 [eess], Nov. 2023. DOI: 10.48550/arXiv.2311.01702. [Online]. Available: http://arxiv.org/abs/2311.01702 (visited on 03/02/2025).

[16] K. Weimann and T. O. F. Conrad, „Transfer learning for ECG classification“, *Scientific Reports*, vol. 11, p. 5251, Mar. 2021, ISSN: 2045-2322. DOI: 10.1038/s41598-021-84374-8. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7933237/ (visited on 02/16/2025).

[17] H. Liu, Z. Zhao, and Q. She, „Self-supervised ECG pre-training“, *Biomedical Signal Processing and Control*, vol. 70, p. 103 010, Sep. 2021, ISSN: 1746-8094. DOI: 10.1016/j.bspc.2021.103010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809421006078 (visited on 03/02/2025).

[18] L. Ericsson, H. Gouk, C. C. Loy, and T. M. Hospedales, „Self-Supervised Representation Learning: Introduction, advances, and challenges“, *IEEE Signal Processing Magazine*, vol. 39, no. 3, pp. 42–62, May 2022, Conference Name: IEEE Signal Processing Magazine, ISSN: 1558-0792. DOI: 10.1109/MSP.2021.3134634. [Online]. Available: https://ieeexplore.ieee.org/document/9770283?denied= (visited on 03/02/2025).

[19] K. Weiss, T. M. Khoshgoftaar, and D. Wang, „A survey of transfer learning“, *Journal of Big Data*, vol. 3, no. 1, p. 9, May 2016, ISSN: 2196-1115. DOI: 10.1186/s40537-016-0043-6. [Online]. Available: https://doi.org/10.1186/s40537-016-0043-6 (visited on 03/02/2025).

[20] S. Al-Stouhi and C. K. Reddy, „Transfer learning for class imbalance problems with inadequate data“, en, *Knowledge and Information Systems*, vol. 48, no. 1, pp. 201–228, Jul. 2016, ISSN: 0219-3116. DOI: 10.1007/s10115-015-0870-3. [Online]. Available: https://doi.org/10.1007/s10115-015-0870-3 (visited on 03/02/2025).

[21] S. Soltanieh, J. Hashemi, and A. Etemad, „In-Distribution and Out-of-Distribution Self-supervised ECG Representation Learning for Arrhythmia Detection“, *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 2, pp. 789–800, Feb. 2024, arXiv:2304.06427 [cs], ISSN: 2168-2194, 2168-2208. DOI: 10.1109/JBHI.2023.3331626. [Online]. Available: http://arxiv.org/abs/2304.06427 (visited on 03/02/2025).

[22] S.-C. Huang, A. Pareek, M. Jensen, M. P. Lungren, S. Yeung, and A. S. Chaudhari, „Self-supervised learning for medical image classification: A systematic review and implementation guidelines“, en, *npj Digital Medicine*, vol. 6, no. 1, pp. 1–16, Apr. 2023, Publisher: Nature Publishing Group, ISSN: 2398-6352. DOI: 10.1038/s41746-023-00811-0. [On-

line]. Available: https://www.nature.com/articles/s41746-023-00811-0 (visited on 03/02/2025).

[23] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, *A Simple Framework for Contrastive Learning of Visual Representations*, arXiv:2002.05709, Jul. 2020. [Online]. Available: http://arxiv.org/abs/2002.05709 (visited on 10/16/2024).

[24] J.-B. Grill, F. Strub, F. Altché, *et al.*, *Bootstrap your own latent: A new approach to self-supervised Learning*, arXiv:2006.07733 [cs], Sep. 2020. DOI: 10.48550/arXiv.2006.07733. [Online]. Available: http://arxiv.org/abs/2006.07733 (visited on 03/21/2025).

[25] L. Ericsson, H. Gouk, and T. M. Hospedales, „How Well Do Self-Supervised Models Transfer?“, in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, ISSN: 2575-7075, Jun. 2021, pp. 5410–5419. DOI: 10.1109/CVPR46437.2021.00537. [Online]. Available: https://ieeexplore.ieee.org/document/9577835 (visited on 03/02/2025).

[26] S. Konstantakos, J. Cani, I. Mademlis, *et al.*, „Self-supervised visual learning in the low-data regime: A comparative evaluation“, *Neurocomputing*, vol. 620, p. 129 199, Mar. 2025, arXiv:2404.17202 [cs], ISSN: 09252312. DOI: 10.1016/j.neucom.2024.129199. [Online]. Available: http://arxiv.org/abs/2404.17202 (visited on 03/02/2025).

[27] S. Shurrab and R. Duwairi, „Self-supervised learning methods and applications in medical imaging analysis: A survey“, *PeerJ Computer Science*, vol. 8, e1045, Jul. 2022, ISSN: 2376-5992. DOI: 10.7717/peerj-cs.1045. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9455147/ (visited on 03/21/2025).

[28] Z. Ding, Y. Hu, Z. Li, *et al.*, „Cross-Modality Cardiac Insight Transfer: A Contrastive Learning Approach to Enrich ECG with CMR Features“, en, in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, M. G. Linguraru, Q. Dou, A. Feragen, *et al.*, Eds., vol. 15003, Series Title: Lecture Notes in Computer Science, Cham: Springer Nature Switzerland, 2024, pp. 109–119, ISBN: 978-3-031-72383-4 978-3-031-72384-1. DOI: 10.1007/978-3-031-72384-1_11. [Online]. Available: https://link.springer.com/10.1007/978-3-031-72384-1_11 (visited on 10/21/2024).

[29] H. A. A. K. Hammoud, T. Das, F. Pizzati, P. Torr, A. Bibi, and B. Ghanem, *On Pretraining Data Diversity for Self-Supervised Learning*, arXiv:2403.13808 [cs], Jul. 2024. DOI: 10.48550/arXiv.2403.13808. [Online]. Available: http://arxiv.org/abs/2403.13808 (visited on 03/02/2025).

[30] O. Turgut, *Oetu/MMCL-ECG-CMR*, original-date: 2023-08-08T08:12:15Z, Jan. 2025. [Online]. Available: https://github.com/oetu/MMCL-ECG-CMR (visited on 02/14/2025).

[31] J. Wang, C. Lan, C. Liu, *et al.*, *Generalizing to Unseen Domains: A Survey on Domain Generalization*, en, arXiv:2103.03097 [cs], May 2022. DOI: 10.48550/arXiv.2103.03097. [Online]. Available: http://arxiv.org/abs/2103.03097 (visited on 03/09/2025).

[32] J. Zheng, H. Chu, D. Struppa, *et al.*, „Optimal Multi-Stage Arrhythmia Classification Approach“, en, *Scientific Reports*, vol. 10, no. 1, p. 2898, Feb. 2020, Publisher: Nature Publishing Group, ISSN: 2045-2322. DOI: 10.1038/s41598-020-59821-7. [Online]. Available: https://www.nature.com/articles/s41598-020-59821-7 (visited on 02/13/2025).

[33] J. Zheng, H. Guo, and H. Chu, *A large scale 12-lead electrocardiogram database for arrhythmia study*, Aug. 2022. DOI: 10.13026/WGEX-ER52. [Online]. Available: https://physionet.org/content/ecg-arrhythmia/1.0.0/ (visited on 10/27/2024).

[34] P. Wagner, N. Strodthoff, R.-D. Bousseljot, W. Samek, and T. Schaeffter, *PTB-XL, a large publicly available electrocardiography dataset*, Nov. 2022. DOI: 10.13026/KFZX-AW45. [Online]. Available: https://physionet.org/content/ptb-xl/1.0.3/ (visited on 02/13/2025).

[35] P. Wagner, N. Strodthoff, R.-D. Bousseljot, *et al.*, „PTB-XL, a large publicly available electrocardiography dataset“, en, *Scientific Data*, vol. 7, no. 1, p. 154, May 2020, Publisher: Nature Publishing Group, ISSN: 2052-4463. DOI: 10.1038/s41597-020-0495-6. [Online]. Available: https://www.nature.com/articles/s41597-020-0495-6 (visited on 02/13/2025).

[36] O. Bernard, A. Lalande, C. Zotti, *et al.*, „Deep Learning Techniques for Automatic MRI Cardiac Multi-Structures Segmentation and Diagnosis: Is the Problem Solved?“, *IEEE Transactions on Medical Imaging*, vol. 37, no. 11, pp. 2514–2525, Nov. 2018, ISSN: 0278-0062, 1558-254X. DOI: 10.1109/TMI.2018.2837502. [Online]. Available: https://ieeexplore.ieee.org/document/8360453/ (visited on 10/28/2024).

[37] R. J. Prineas, R. S. Crow, and Z.-M. Zhang, *The Minnesota Code Manual of Electrocardiographic Findings*, en. London: Springer, 2010, ISBN: 978-1-84882-777-6 978-1-84882-778-3. DOI: 10.1007/978-1-84882-778-3. [Online]. Available: http://link.springer.com/10.1007/978-1-84882-778-3 (visited on 02/17/2025).

[38] B. J. Soher, B. M. Dale, and E. M. Merkle, „A review of MR physics: 3T versus 1.5T“, eng, *Magnetic Resonance Imaging Clinics of North America*, vol. 15, no. 3, pp. 277–290, v, Aug. 2007, ISSN: 1064-9689. DOI: 10.1016/j.mric.2007.06.002.

[39] D. P. Hinton, L. L. Wald, J. Pitts, and F. Schmitt, „Comparison of Cardiac MRI on 1.5 and 3.0 Tesla Clinical Whole Body Systems“, en-US, *Investigative Radiology*, vol. 38, no. 7, p. 436, Jul. 2003. DOI: 10.1097/01.RLI.0000067489.31556.70. [Online]. Available: https://journals.lww.com/investigativeradiology/abstract/2003/07000/comparison_of_cardiac_mri_on_1_5_and_3_0_tesla.9.aspx (visited on 02/20/2025).

[40] Z.-M. Zhang, S. Chen, and Y.-Z. Liang, „Baseline correction using adaptive iteratively reweighted penalized least squares“, en, *Analyst*, vol. 135, no. 5, pp. 1138–1146, Apr. 2010, Publisher: The Royal Society of Chemistry Read_Status: New Read_Status_Date: 2025-02-13T09:59:36.184Z, ISSN: 1364-5528. DOI: 10.1039/B922045C. [Online]. Available: https://pubs.rsc.org/en/content/articlelanding/2010/an/b922045c (visited on 02/13/2025).

[41] M. G. Khan, Ed., *Rapid ECG Interpretation*, en. Totowa, NJ: Humana Press, 2008, Read_Status: New Read_Status_Date: 2025-02-13T10:04:15.435Z, ISBN: 978-1-58829-979-6 978-1-59745-408-7. DOI: 10.1007/978-1-59745-408-7. [Online]. Available: http://link.springer.com/10.1007/978-1-59745-408-7 (visited on 02/13/2025).

[42] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, arXiv:2010.11929, Jun. 2021. DOI: 10.48550/arXiv.2010.11929. [Online]. Available: http://arxiv.org/abs/2010.11929 (visited on 10/23/2024).

[43] I. Loshchilov and F. Hutter, *Decoupled Weight Decay Regularization*, arXiv:1711.05101 [cs] version: 3, Jan. 2019. DOI: 10.48550/arXiv.1711.05101. [Online]. Available: http://arxiv.org/abs/1711.05101 (visited on 02/21/2025).

[44] J. T. C. Schwabedal, J. C. Snyder, A. Cakmak, S. Nemati, and G. D. Clifford, *Addressing Class Imbalance in Classification Problems of Noisy Signals by using Fourier Transform Surrogates*, arXiv:1806.08675 [eess], Jan. 2019. DOI: 10.48550/arXiv.1806.08675. [Online]. Available: http://arxiv.org/abs/1806.08675 (visited on 03/18/2025).

[45] K. He, X. Zhang, S. Ren, and J. Sun, *Deep Residual Learning for Image Recognition*, arXiv:1512.03385 [cs], Dec. 2015. DOI: 10.48550/arXiv.1512.03385. [Online]. Available: http://arxiv.org/abs/1512.03385 (visited on 02/13/2025).

[46] S. Watanabe, *Tree-Structured Parzen Estimator: Understanding Its Algorithm Components and Their Roles for Better Empirical Performance*, arXiv:2304.11127 [cs], May 2023. DOI: 10.48550/arXiv.2304.11127. [Online]. Available: http://arxiv.org/abs/2304.11127 (visited on 02/25/2025).

[47] P. J. Rousseeuw, „Silhouettes: A graphical aid to the interpretation and validation of cluster analysis“, *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, Nov. 1987, ISSN: 0377-0427. DOI: 10.1016/0377-0427(87)90125-7. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0377042787901257 (visited on 03/20/2025).

[48] L. McInnes, J. Healy, and J. Melville, *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, arXiv:1802.03426 [stat], Sep. 2020. DOI: 10.48550/arXiv.1802.03426. [Online]. Available: http://arxiv.org/abs/1802.03426 (visited on 02/11/2025).

[49] R. Shad, C. Zakka, D. Kaur, *et al.*, *A Generalizable Deep Learning System for Cardiac MRI*, arXiv:2312.00357 [eess], Dec. 2023. DOI: 10.48550/arXiv.2312.00357. [Online]. Available: http://arxiv.org/abs/2312.00357 (visited on 03/06/2025).

[50] M. M. Søndergaard, J. Riis, K. W. Bodker, *et al.*, „Associations between left bundle branch block with different PR intervals, QRS durations, heart rates and the risk of heart failure: A register-based cohort study using ECG data from the primary care setting“, *Open Heart*, vol. 8, no. 1, e001425, Feb. 2021, ISSN: 2053-3624. DOI: 10.1136/openhrt-2020-001425. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7880100/ (visited on 03/20/2025).

[51] E. R. Uyguanco, A. Mirandi, G. Qureshi, J. Lazar, A. Chhabra, and J. Kassotis, „Prolongation of QRS duration and axis deviation in the right bundle branch block are not markers for left ventricular systolic dysfunction“, *The International Journal of Angiology : Official Publication of the International College of Angiology, Inc*, vol. 19, no. 2, e83–e85, 2010, ISSN: 1061-1711. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3005414/ (visited on 03/20/2025).

[52] P. Bachman, R. D. Hjelm, and W. Buchwalter, *Learning Representations by Maximizing Mutual Information Across Views*, arXiv:1906.00910 [cs], Jul. 2019. DOI: 10.48550/arXiv.1906.00910. [Online]. Available: http://arxiv.org/abs/1906.00910 (visited on 03/18/2025).

[53] A. v. d. Oord, Y. Li, and O. Vinyals, *Representation Learning with Contrastive Predictive Coding*, arXiv:1807.03748 [cs], Jan. 2019. DOI: 10.48550/arXiv.1807.03748. [Online]. Available: http://arxiv.org/abs/1807.03748 (visited on 03/18/2025).

[54] A. Kolesnikov, X. Zhai, and L. Beyer, *Revisiting Self-Supervised Visual Representation Learning*, arXiv:1901.09005 [cs], Jan. 2019. DOI: 10.48550/arXiv.1901.09005. [Online]. Available: http://arxiv.org/abs/1901.09005 (visited on 03/18/2025).

[55] D. Gündüzalp and C. Zou, „Combined Self-Supervised Learning for ECG Classification Based on the Classification Token“, en, in *2024 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, Houston, TX, USA: IEEE, Nov. 2024, pp. 1–8, ISBN: 9798350351552. DOI: 10.1109/BHI62660.2024.10913542. [Online]. Available: https://ieeexplore.ieee.org/document/10913542/ (visited on 03/18/2025).

[56] D. M. Lloyd-Jones, T. J. Wang, E. P. Leip, *et al.*, „Lifetime Risk for Development of Atrial Fibrillation“, *Circulation*, vol. 110, no. 9, pp. 1042–1046, Aug. 2004, Publisher: American Heart Association. DOI: 10.1161/01.CIR.0000140263.20897.42. [Online]. Available: https://www.ahajournals.org/doi/full/10.1161/01.CIR.0000140263.20897.42 (visited on 03/20/2025).

[57] Q. Chen, Z. Yi, and J. Cheng, „Atrial fibrillation in aging population“, en, *AGING MEDICINE*, vol. 1, no. 1, pp. 67–74, 2018, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/agm2 ISSN: 2475-0360. DOI: 10.1002/agm2.12015. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/agm2.12015 (visited on 03/20/2025).

[58] Z. I. Attia, P. A. Friedman, P. A. Noseworthy, *et al.*, „Age and Sex Estimation Using Artificial Intelligence From Standard 12-Lead ECGs“, *Circulation. Arrhythmia and Electrophysiology*, vol. 12, no. 9, e007284, Aug. 2019, ISSN: 1941-3149. DOI: 10.1161/CIRCEP.119.007284. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7661045/ (visited on 02/19/2025).

[59] M. Paknezhad, S. Marchesseau, and M. S. Brown, „Automatic basal slice detection for cardiac analysis“, *Journal of Medical Imaging*, vol. 3, no. 3, p. 034 004, Jul. 2016, ISSN: 2329-4302. DOI: 10.1117/1.JMI.3.3.034004. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5028419/ (visited on 03/20/2025).

[60] S. Hwang, J. Cha, J. Heo, S. Cho, and Y. Park, „Multi-label ECG Abnormality Classification Using A Combined ResNet-DenseNet Architecture with ResU Blocks“, in *2023 IEEE EMBS Special Topic Conference on Data Science and Engineering in Healthcare, Medicine and Biology*, Dec. 2023, pp. 111–112. DOI: 10.1109/IEEECONF58974.2023.10404234. [Online]. Available: https://ieeexplore.ieee.org/document/10404234 (visited on 03/20/2025).

Declaration of Authenticity

I hereby declare that any individual work / pair work / team work submitted for assessment is entirely the product of my own / my own and my partner's / my own and my team's effort,

- that I/we have correctly cited all text passages that do not originate from me/us, in accordance with standard academic citation rules[9], and that I/we have clearly mentioned all sources used;
- that I/we have declared in footnotes or in an index of auxiliary tools all aids used (AI assistance systems such as chatbots[10], translation[11], paraphrasing[12], or programming applications[13], and indicated their use at the corresponding text passages;
- that I/we have acquired all intangible rights to any materials I/we may have used, such as images or graphics, or that these materials were created by me/us;
- that the topic, the thesis or parts of it have not been used in an assessment of another module, unless this has been expressly agreed with the lecturer in advance and is stated as such;
- that I/we am/are aware that my/our work may be checked for plagiarism and for third-party authorship of human or technical origin (artificial intelligence);
- that I/we am/are aware that the FHNW School of Engineering will pursue a violation of this declaration of authenticity and that disciplinary consequences (reprimand or expulsion from the study program) may result from this.

Windisch, 21. March 2025

Name: Dominik Filliger

Signature: 

Name: Noah Leuenberger

Signature: 

[9] e.g. APA oder IEEE
[10] e.g., ChatGPT
[11] e.g., DeepL
[12] e.g., Quillbot
[13] e.g., Github Copilot

Appendices

A Datasets

A.1 Chapman

A.1.1 Demographic and Label Distributions

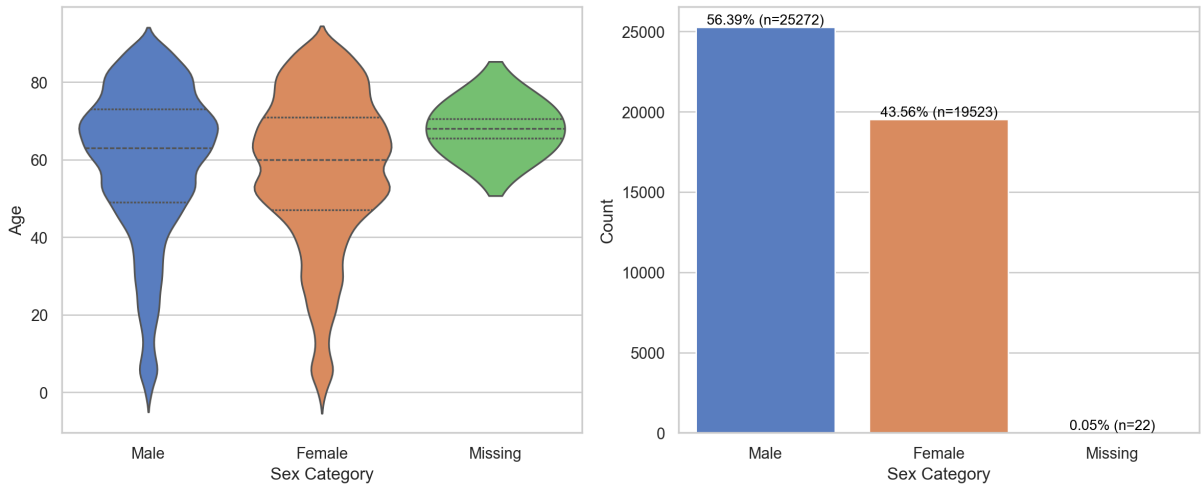

Figure A.1: Age distribution by sex category (left) and sex category distribution (right) for the dataset ($n = 44,817$). The violin plot illustrates the age distribution within each sex category, including missing values, with quartiles marked. The bar plot displays the proportion of each sex category.

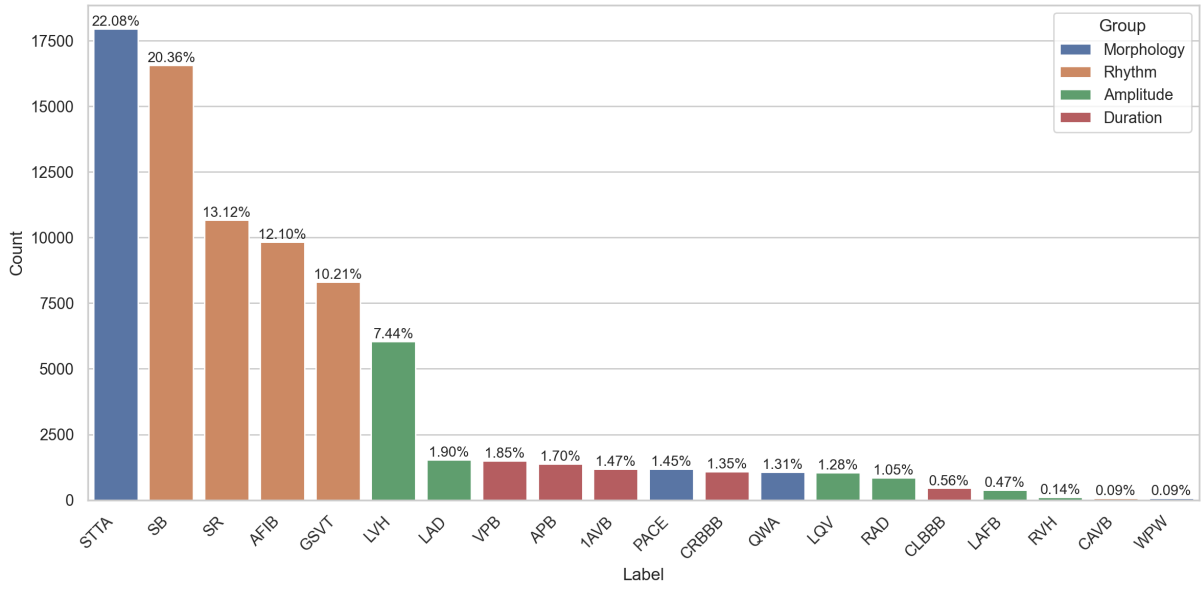

Figure A.2: Distribution of labels by group. The bar plot represents the frequency of each label, with colors indicating the associated group. Percentages are annotated to show the relative occurrence of each label within the dataset. As this is a multi-label dataset, instances may have multiple labels, so percentages exceed 100%.

A.2 PTB-XL

A.2.1 Demographic and Label Distributions

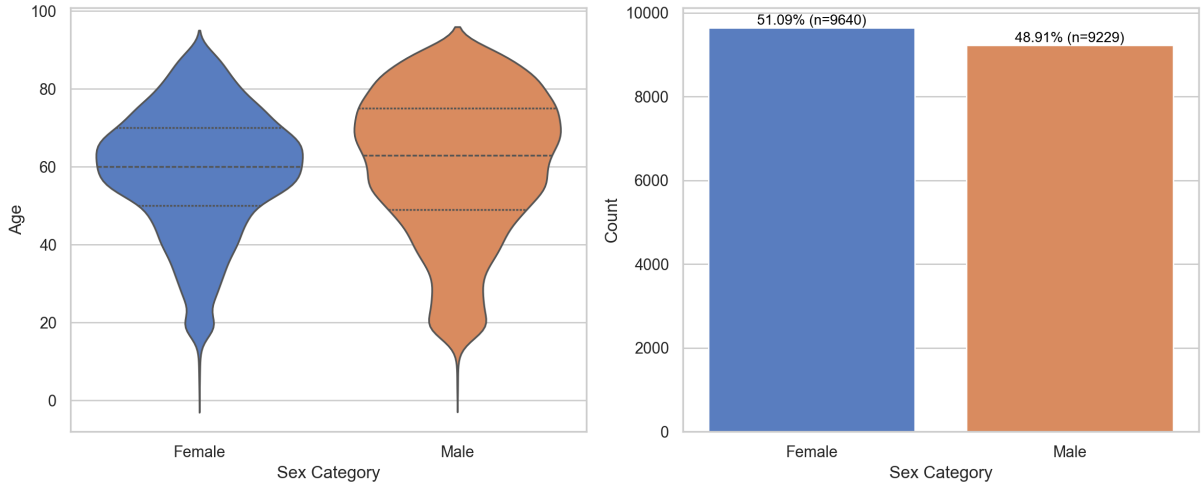

Figure A.3: Age distribution by sex category (left) and sex category distribution (right) for the dataset ($n = 18,869$). The violin plot illustrates the age distribution within each sex category, with quartiles marked. Age values greater than 89 are clipped to 90 per HIPAA guidelines.

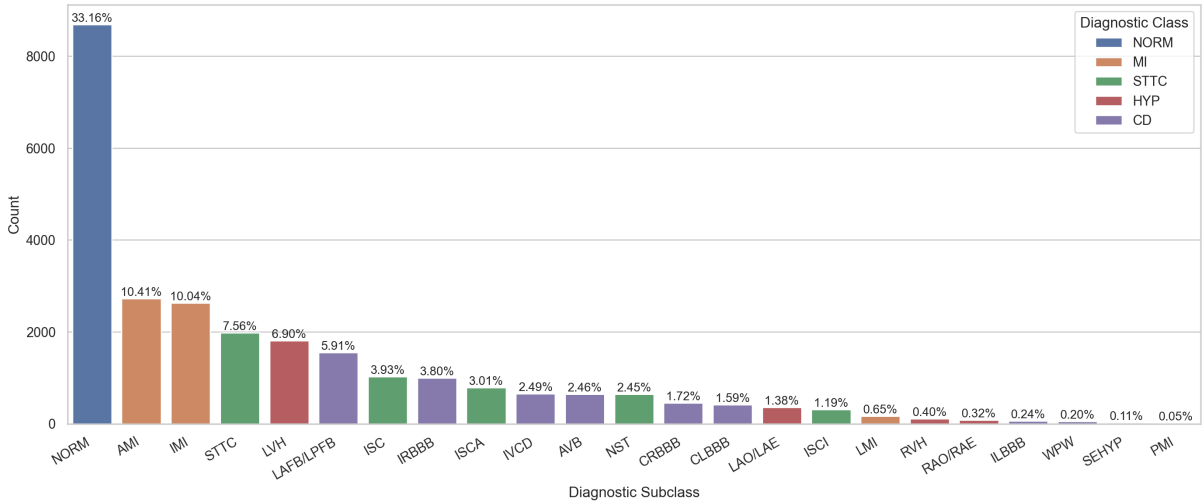

Figure A.4: Distribution of SCP statement labels by diagnostic subclass. The bar plot represents the frequency of each diagnostic subclass, with colors indicating the associated diagnostic class. Percentages are annotated to show the relative occurrence of each subclass within the dataset. As this is a multi-label dataset, instances may have multiple labels, so percentages exceed 100%.

A.2.2 Label Structure

Group	Abbr.	Full Name	Count (%)
Conduction Disturbances (8)	CLBBB	Complete Left Bundle Branch Block	416 (2.20%)
	CRBBB	Complete Right Bundle Branch Block	451 (2.39%)
	ILBBB	Incomplete Left Bundle Branch Block	62 (0.33%)
	IRBBB	Incomplete Right Bundle Branch Block	996 (5.28%)
	IVCD	Non-specific Intraventricular Conduction Disturbance	653 (3.46%)
	LAFB/LPFB	Left Posterior Fascicular Block	1549 (8.21%)
	WPW	Wolff-Parkinson-White Syndrome	53 (0.28%)
	AVB	Second Degree AV Block	645 (3.42%)
Hypertrophy (5)	LAO/LAE	Left Atrial Overload/Enlargement	362 (1.92%)
	LVH	Left Ventricular Hypertrophy	1807 (9.58%)
	RAO/RAE	Right Atrial Overload/Enlargement	84 (0.45%)
	RVH	Right Ventricular Hypertrophy	106 (0.56%)
	SEHYP	Septal Hypertrophy	28 (0.15%)
Myocardial Infarction (4)	AMI	Subendocardial Injury in Lateral Leads	2728 (14.46%)
	IMI	Subendocardial Injury in Inferolateral Leads	2629 (13.93%)
	LMI	Lateral Myocardial Infarction	171 (0.91%)
	PMI	Posterior Myocardial Infarction	14 (0.07%)
Normal (1)	NORM	Normal ECG	8686 (46.03%)
ST-T Changes (5)	ISCA	Ischemic in Anterior Leads	789 (4.18%)
	ISCI	Ischemic in Inferolateral Leads	312 (1.65%)
	ISC	Non-specific Ischemic	1029 (5.45%)
	NST	Non-specific ST Changes	643 (3.41%)
	STTC	Electrolytic Disturbance or Drug Effect	1980 (10.49%)

Table A.1: Hierarchical overview of PTB-XL dataset labels, with counts and percentages relative to the total dataset sample size ($n = 18,869$).

Abbr.	Full Name	Count	(%)
AFIB	Atrial Fibrillation	1514	6.95%
AFLT	Atrial Flutter	73	0.33%
BIGU	Bigeminal Pattern (Unknown Origin)	82	0.38%
PACE	Pacemaker Rhythm	296	1.36%
PSVT	Paroxysmal Supraventricular Tachycardia	24	0.11%
SARRH	Sinus Arrhythmia	772	3.54%
SBRAD	Sinus Bradycardia	637	2.92%
SVARR	Supraventricular Arrhythmia	157	0.72%
SR	Sinus Rhythm	16782	76.99%
STACH	Sinus Tachycardia	826	3.79%
SVTAC	Supraventricular Tachycardia	27	0.12%
TRIGU	Trigeminal Pattern (Unknown Origin)	20	0.09%

Table A.2: Overview of rhythm-related diagnoses in the PTB-XL dataset ($n = 21,799$), with counts and relative frequencies.

PTB-XL Abbr.	Chapman Abbr.	PTB-XL Full Name
AFIB	AFIB	Atrial Fibrillation
AFLT	AFIB	Atrial Flutter
SR	SR	Sinus Rhythm
SARRH	SR	Sinus Arrhythmia
SBRAD	SB	Sinus Bradycardia
PACE	PACE	Pacemaker Rhythm
STACH	GSVT	Sinus Tachycardia
SVARR	GSVT	Sinus Tachycardia
SVTAC	GSVT	Supraventricular Tachycardia
PSVT	GSVT	Paroxysmal Supraventricular Tachycardia

Table A.3: Mapping of PTB-XL rhythm labels to Chapman labels.

A.3 ACDC

A.3.1 Demographic and Physical Characteristics

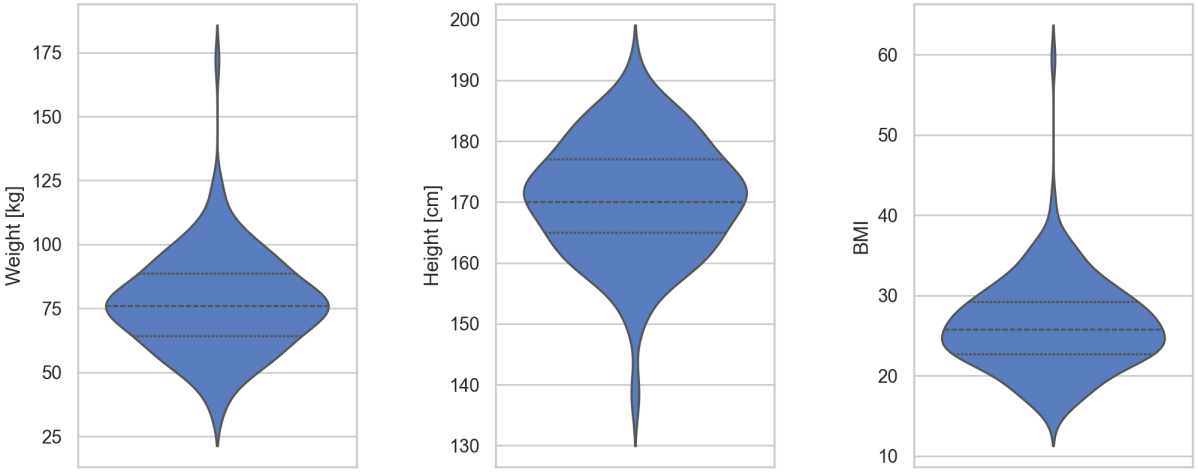

Figure A.5: Weight distribution (left), height distribution (center), and BMI distribution (right) for the ACDC dataset. Quartiles are marked.

B Classifier Model Architectures

Layer	Type	Details
Patch Embedding		
patch_embed	Conv2d	$1 \to 384$, 1×100, Stride 100
norm	Identity	-
Transformer Encoder		
pos_drop	Dropout	$p = 0.1$
Transformer Block 1		
norm1	LayerNorm	384
Self-Attention Mechanism		
qkv	Linear	$384 \to 1152$
attn_drop	Dropout	$p = 0.0$
proj	Linear	$384 \to 384$
proj_drop	Dropout	$p = 0.1$
ls1	Identity	Skip Connection
drop_path1	Identity	-
norm2	LayerNorm	384
Feedforward MLP		
fc1	Linear	$384 \to 1536$
act	GELU	Activation
drop1	Dropout	$p = 0.1$
fc2	Linear	$1536 \to 384$
drop2	Dropout	$p = 0.1$
ls2	Identity	Skip Connection
drop_path2	Identity	-
block2-3	Transformer Block $\times 2$	Same as Block 1
Classification Head		
fc_norm	LayerNorm	384
attention_pool	MultiheadAttention	$384 \to 384$
head	Linear	$384 \to 20$

Table B.1: ECG Classifier Model Architecture (Vision Transformer [42] Backbone)

Layer	Type	Details
Stem		
conv1	Conv2d	$3 \to 64$, 7×7, Stride 2
bn1	BatchNorm2d	64
relu	ReLU	Inplace
maxpool	MaxPool2d	3×3, Stride 2
Residual Block 1		
res1_1	Conv2d	$64 \to 64$, 1×1
	BatchNorm2d	64
	ReLU	Inplace
	Conv2d	$64 \to 64$, 3×3
	BatchNorm2d	64
	ReLU	Inplace
	Conv2d	$64 \to 256$, 1×1
	BatchNorm2d	256
	Downsample	Conv2d $64 \to 256$, 1×1
res1_2-1_3	Bottleneck $\times 2$	$256 \to 64 \to 64 \to 256$
Residual Block 2		
res2_1	Conv2d	$256 \to 128$, 1×1
	BatchNorm2d	128
	ReLU	Inplace
	Conv2d	$128 \to 128$, 3×3, Stride 2
	BatchNorm2d	128
	ReLU	Inplace
	Conv2d	$128 \to 512$, 1×1
	BatchNorm2d	512
	Downsample	Conv2d $256 \to 512$, 1×1, Stride 2
res2_2-2_4	Bottleneck $\times 3$	$512 \to 128 \to 128 \to 512$
Residual Block 3		
res3_1	Conv2d	$512 \to 256$, 1×1
	BatchNorm2d	256
	ReLU	Inplace
	Conv2d	$256 \to 256$, 3×3, Stride 2
	BatchNorm2d	256
	ReLU	Inplace
	Conv2d	$256 \to 1024$, 1×1
	BatchNorm2d	1024
	Downsample	Conv2d $512 \to 1024$, 1×1, Stride 2
res3_2-3_6	Bottleneck $\times 5$	$1024 \to 256 \to 256 \to 1024$
Residual Block 4		
res4_1	Conv2d	$1024 \to 512$, 1×1
	BatchNorm2d	512
	ReLU	Inplace
	Conv2d	$512 \to 512$, 3×3, Stride 2
	BatchNorm2d	512
	ReLU	Inplace
	Conv2d	$512 \to 2048$, 1×1
	BatchNorm2d	2048
	Downsample	Conv2d $1024 \to 2048$, 1×1, Stride 2
res4_2-4_3	Bottleneck $\times 2$	$2048 \to 512 \to 512 \to 2048$
Classification Head		
avgpool	AdaptiveAvgPool2d	Output 1×1
fc	Linear	$2048 \to 5$

Table B.2: CMR Classifier Model Architecture (ResNet-50 [45] Backbone)

C Experiment Setups

The experimental configurations follow [6], with augmentation strategies directly adopted from their implementation [30].

Optimization Parameters	
Optimizer	AdamW [43]
Base learning rate	3×10^{-6}
Weight decay	0.05
Layer decay	0.75
Drop path rate	0.1
Loss function	BCE
Training Schedule	
Total epochs	200
Checkpoint interval	Every 10 epochs
Warmup period	20 epochs (10% of total)
Final epochs selected	50 (best validation performance)
Data Augmentations	
Training:	Fourier surrogates (phase noise=0.075) Gaussian jitter (σ=0.2) Amplitude rescaling (σ=0.5)

Table C.1: ECG fine-tuning configuration. Augmentations follow [6], [30].

Optimization Parameters	
Optimizer	AdamW
Base learning rate	3×10^{-3}
Weight decay	10^{-4}
Loss function	Cross-entropy
Training Schedule	
Total epochs	150
Checkpoint interval	Every 10 epochs
Warmup period	15 epochs
Final epochs selected	80 (best validation accuracy across folds)
Data Augmentations	
Training:	Horizontal flips (p=0.95) Rotation ($\pm 45°$) Color jitter (bright/cont=0.5) Random resized crops (0.6-1.0 scale)

Table C.2: CMR fine-tuning configuration. Augmentations follow [6], [30].

D Additional Results

This appendix contains additional results from the study that were not included in the main body of the thesis.

D.1 ECG

Figure D.1: Baseline (pre-trained) ECG embeddings from the Chapman dataset (faceted by diagnostic label). Each subplot compares the distribution of Chapman embeddings (grey dots and contours) with PTB-XL embeddings (colored markers) for shared rhythm labels.

D.1.1 Sub-Cluster Formation for CRBBB and CLBBB

Figures D.4 and D.5 present faceted plots of records labeled with CRBBB and CLBBB in conjunction with different rhythm categories, focusing on the smaller sub-cluster observed near the PACE region. These UMAP projections suggest that conduction-block patterns form a

Label	F1-Score	Recall	Precision
PACE	0.72	0.62	0.86
AFIB	0.91	0.94	0.89
CLBBB	0.55	0.75	0.44
QWA	0.42	0.35	0.52
LQV	0.09	0.05	0.35
1AVB	0.63	0.57	0.69
CAVB	0.00	0.00	0.00
APB	0.01	0.01	0.14
LAD	0.54	0.57	0.51
SB	**0.94**	**0.99**	0.90
SR	0.84	0.76	0.94
GSVT	0.88	0.85	0.91
VPB	0.41	0.28	0.76
LAFB	0.39	0.51	0.31
RAD	0.53	0.51	0.55
LVH	0.47	0.84	0.32
STTA	0.76	0.75	0.77
CRBBB	0.72	0.89	0.61
WPW	0.00	0.00	0.00
RVH	0.24	0.13	**1.00**
Overall	0.50	0.52	0.57

Table D.1: F1-score, Recall, and Precision for each label and overall on Chapman test partition after fine-tuning. Best-performing values per metric across all labels are highlighted in bold.

compact group adjacent to PACE, indicating that certain conduction-driven features reside in a separate neighborhood of the latent space.

Examination of these dual-labeled embeddings confirms the co-location of conduction abnormalities adjacent to PACE within the latent space, distinguishing them from the primary clusters dominated by rhythm labels. This phenomenon may reflect shared waveform characteristics, such as prolonged QRS complexes or pacing spikes, that set these subgroups apart from the main arrhythmia clusters.

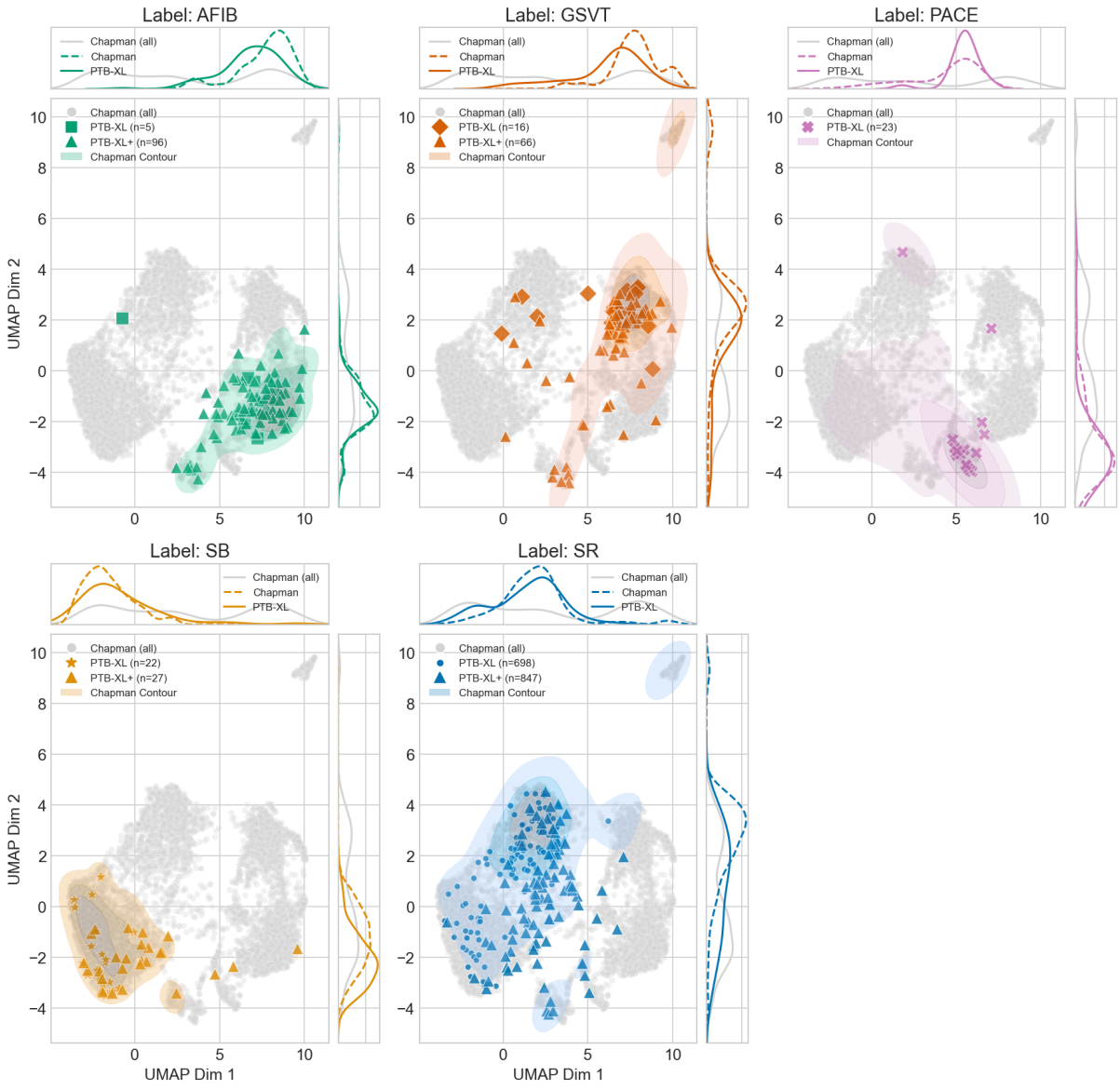

Figure D.2: Fine-tuned ECG embeddings from the Chapman dataset (faceted by diagnostic label). Each subplot highlights the enhanced clustering and overlap between Chapman and PTB-XL embeddings within corresponding rhythm categories post-fine-tuning. Marginal KDE plots emphasize the improved inter-dataset coherence and diagnostic specificity of the embeddings.

Figure D.3: Evolution of the UMAP-embedded ECG features as the model undergoes increasing fine-tuning epochs on the Chapman dataset. **Top-left:** baseline (pre-trained) model, **Top-right:** fine-tuned for 50 epochs, **Bottom-left:** fine-tuned for 100 epochs, **Bottom-right:** fine-tuned for 200 epochs. Single-labeled rhythm conditions are color-coded (SR in blue, AFIB in green, SB in yellow, GSVT in orange, and PACE in pink), whereas gray points represent multi-labeled or other records.

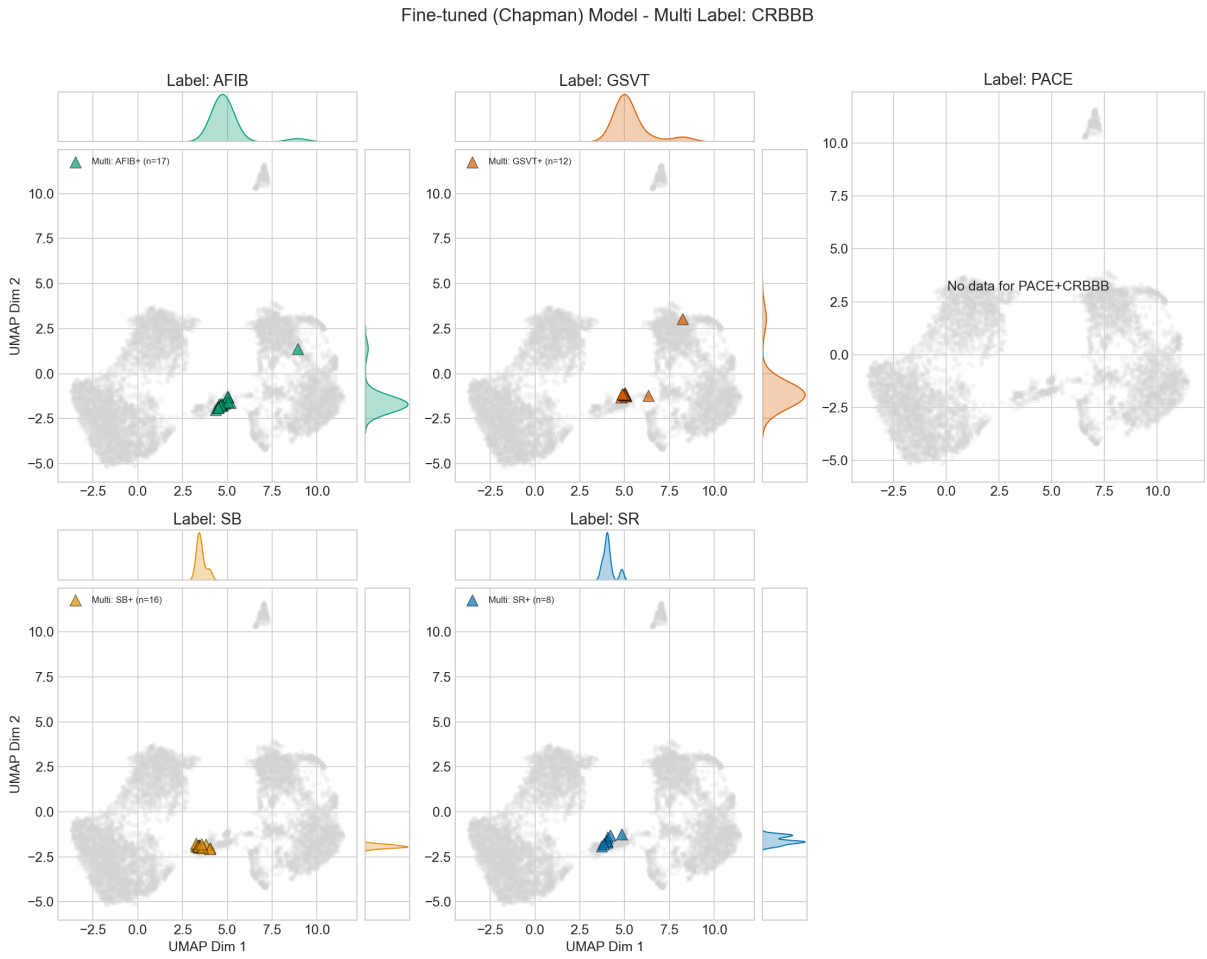

Figure D.4: Faceted UMAP plots showing how records with CRBBB (in combination with each rhythm label) cluster within the fine-tuned (Chapman) embeddings. Each subplot highlights the samples (triangular markers) overlaid on all embeddings (gray points). Kernel density curves indicate the distribution along each UMAP axis.

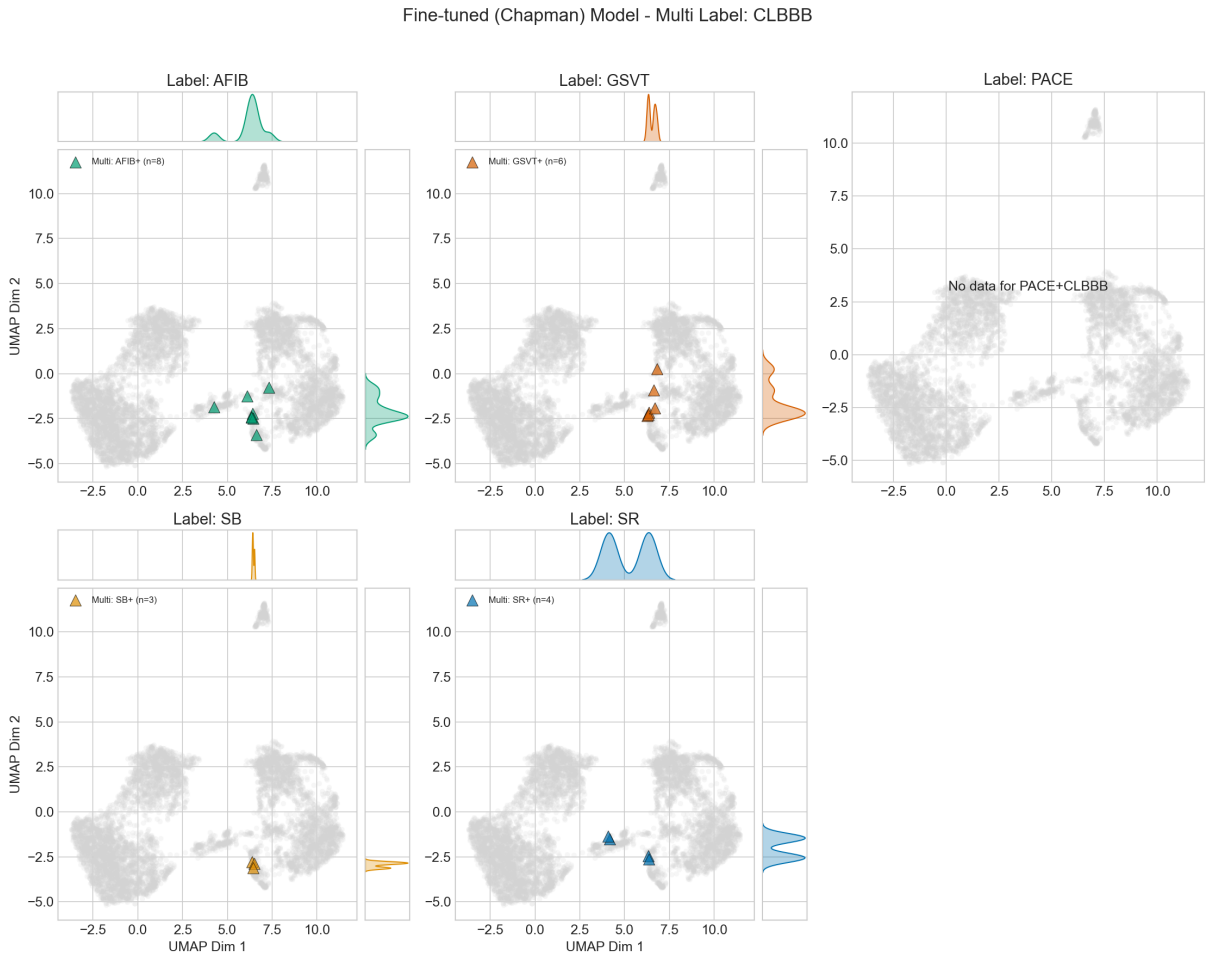

Figure D.5: Faceted UMAP plots illustrating the positioning of CLBBB samples combined with each rhythm label in the fine-tuned (Chapman) latent space. Similar to Figure D.4, each subplot displays examples (triangular markers) alongside all other points (gray). Density curves along the axes reflect each cohort's spread.

D.1.2 Metadata Clustering

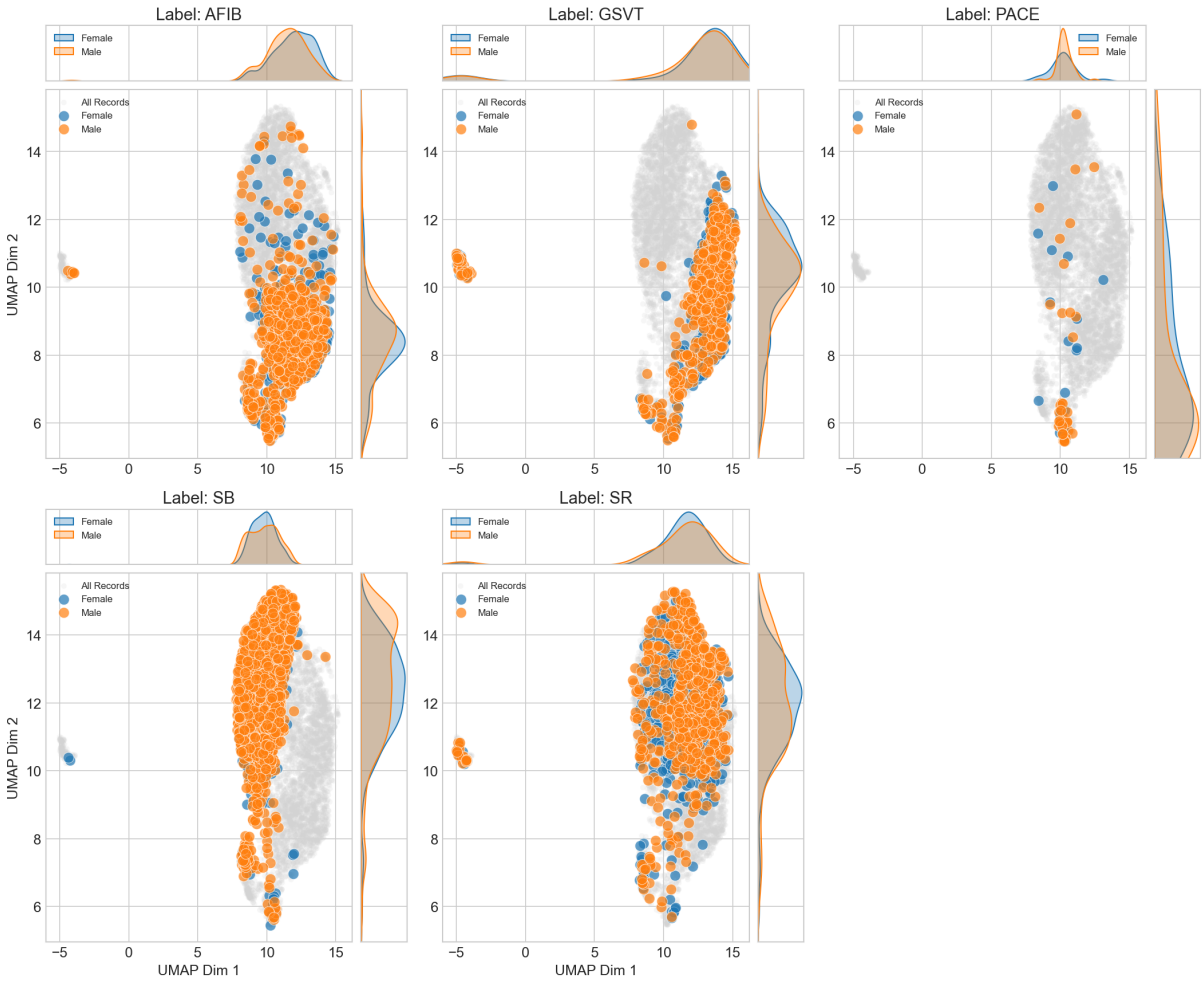

Figure D.6: Baseline (pre-trained) embeddings colored by subjects sex. Orange and blue markers indicate male and female subjects, respectively.

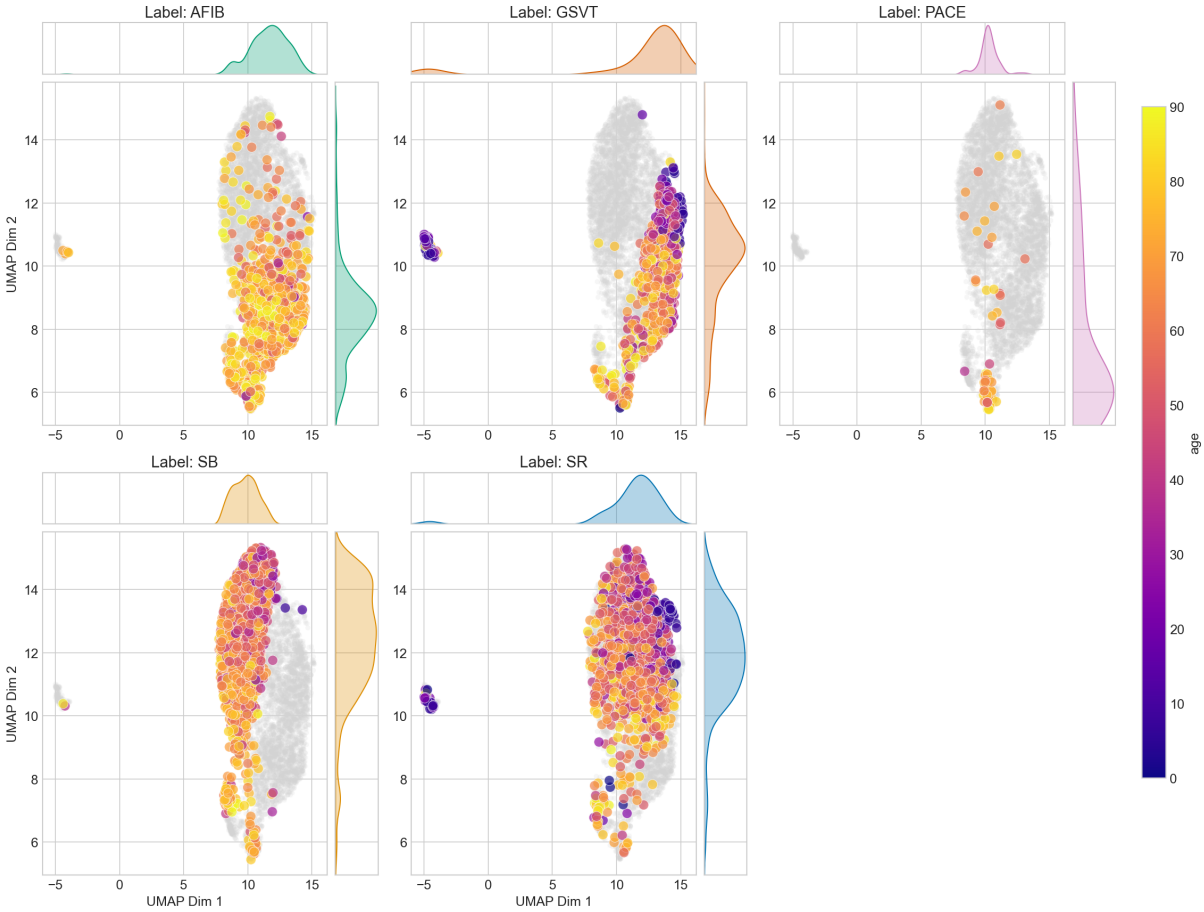

Figure D.7: Baseline (pre-trained) embeddings colored by age (continuous scale). Older subjects do not form isolated clusters and are spread throughout the embedding.

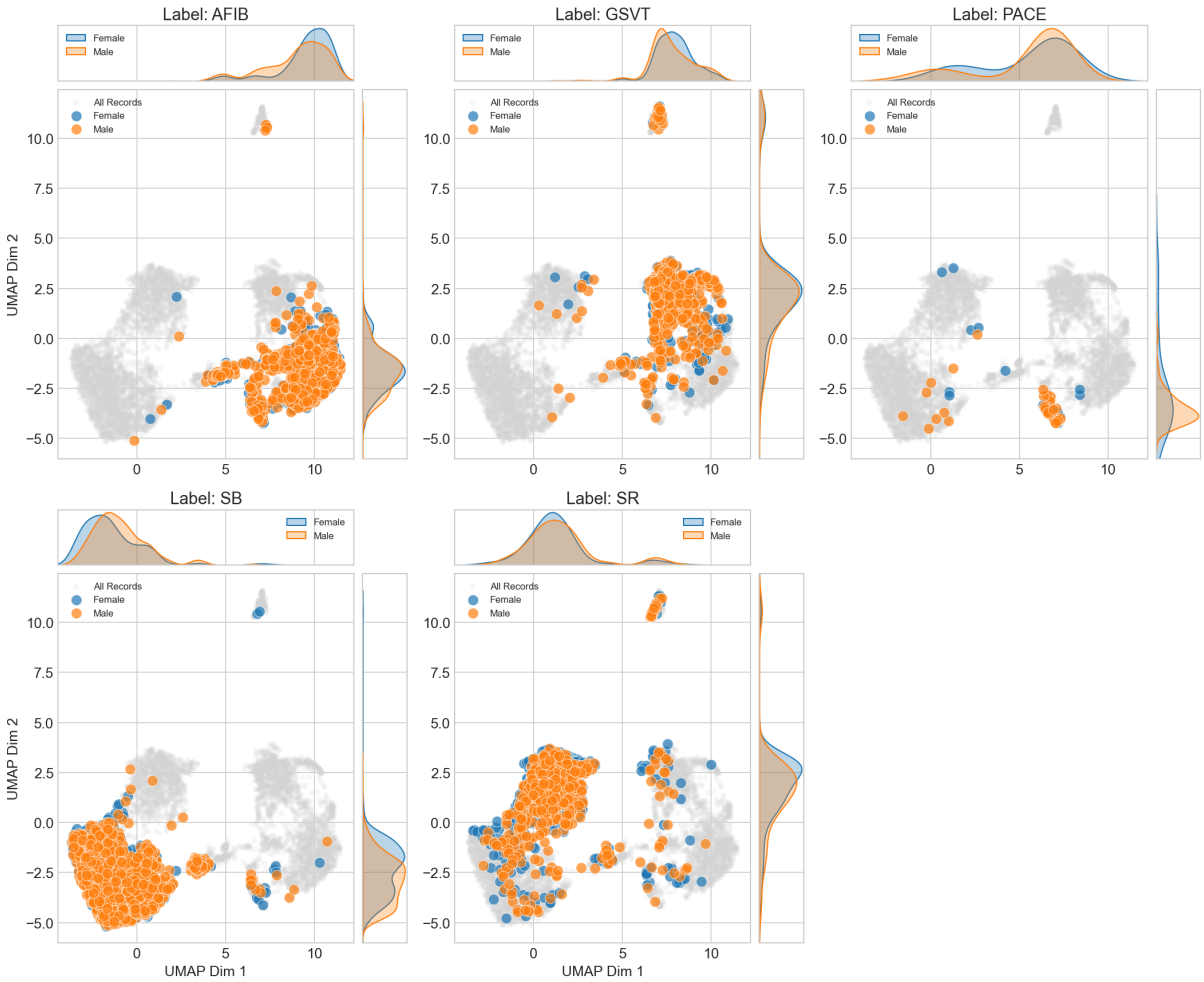

Figure D.8: Fine-tuned (Chapman) embeddings color-coded by subjects sex (orange for male, blue for female).

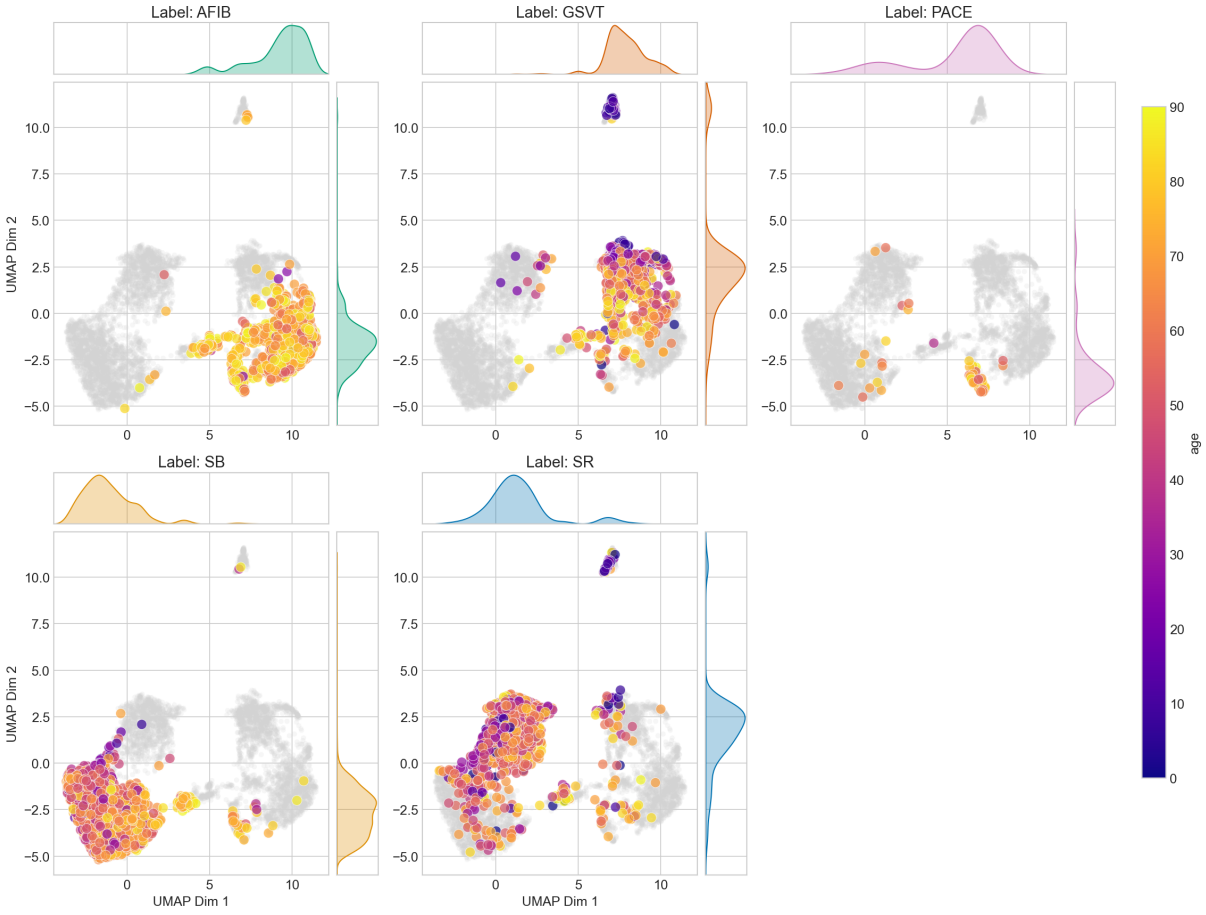

Figure D.9: Fine-tuned (Chapman) embeddings colored by age, illustrating that older and younger individuals appear in most parts of the latent space. Labels such as AFIB do show a mild skew toward older ages, but no separate age-based clusters are observed.

D.2 CMR

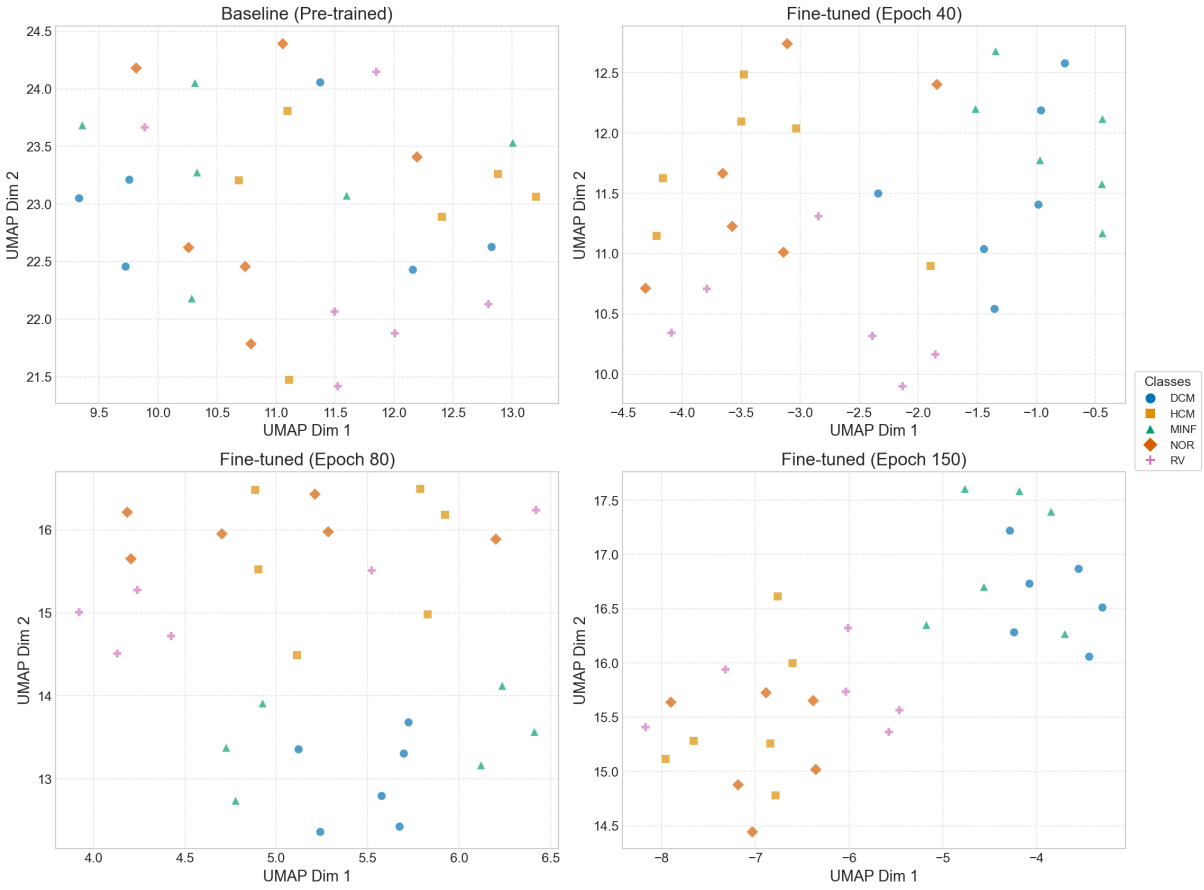

Figure D.10: Progressive evolution of the CMR embeddings for Fold 1 across training epochs. Each point is color-coded by the diagnosed class.

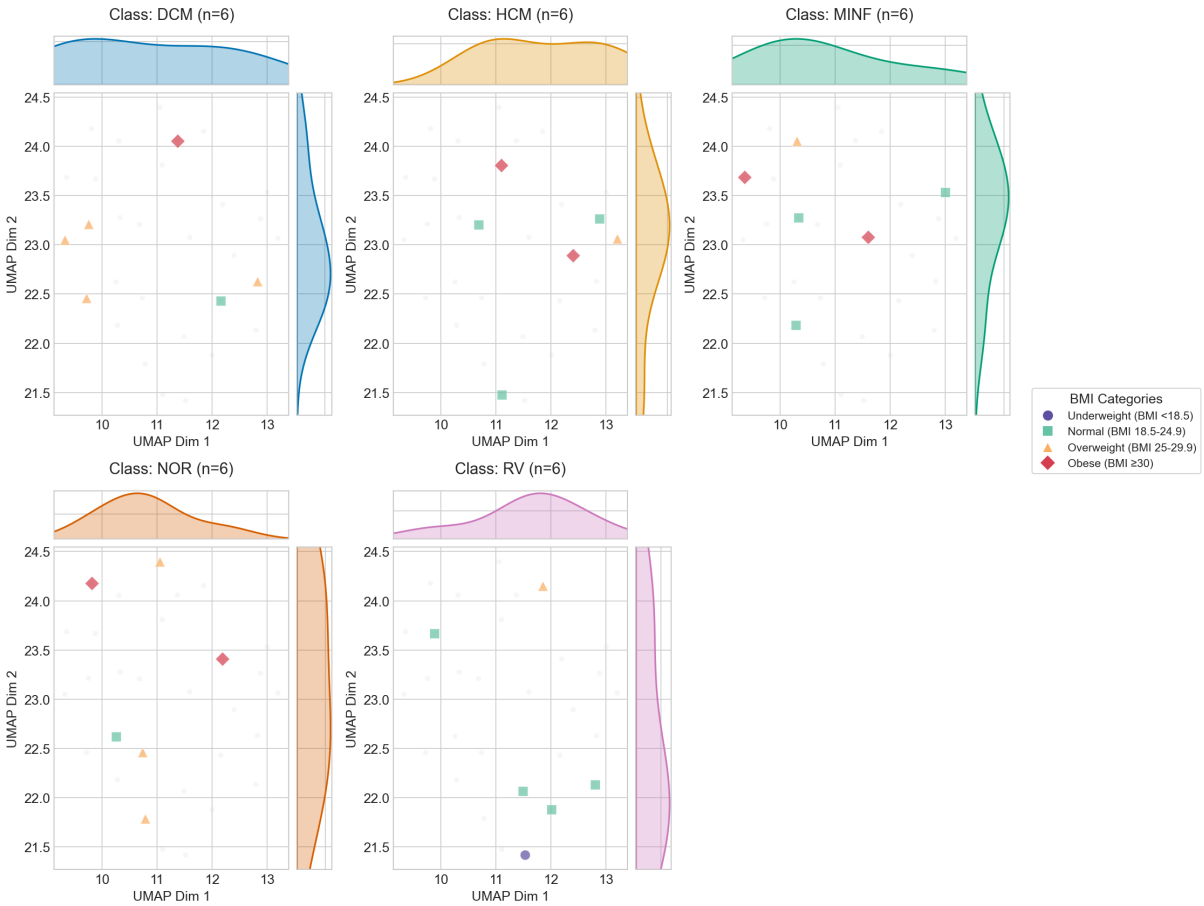

Figure D.11: Baseline (pre-trained) embeddings for Fold 1, faceted by BMI category. Different markers/colors represent normal, overweight, or obese subjects.

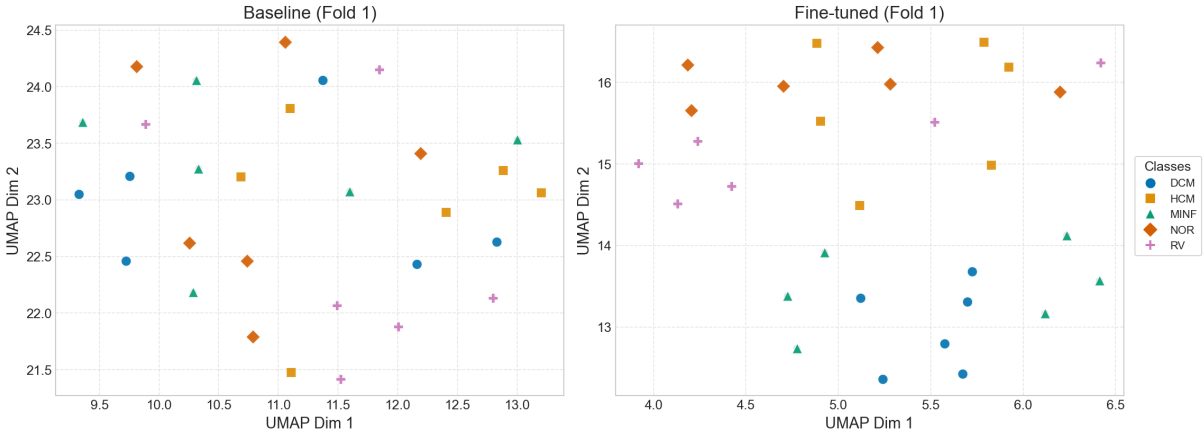

Figure D.12: Side-by-side visualization of baseline and fine-tuned embeddings for Fold 1. Classes are color-coded (DCM, HCM, MINF, NOR, RV) to highlight any clustering.

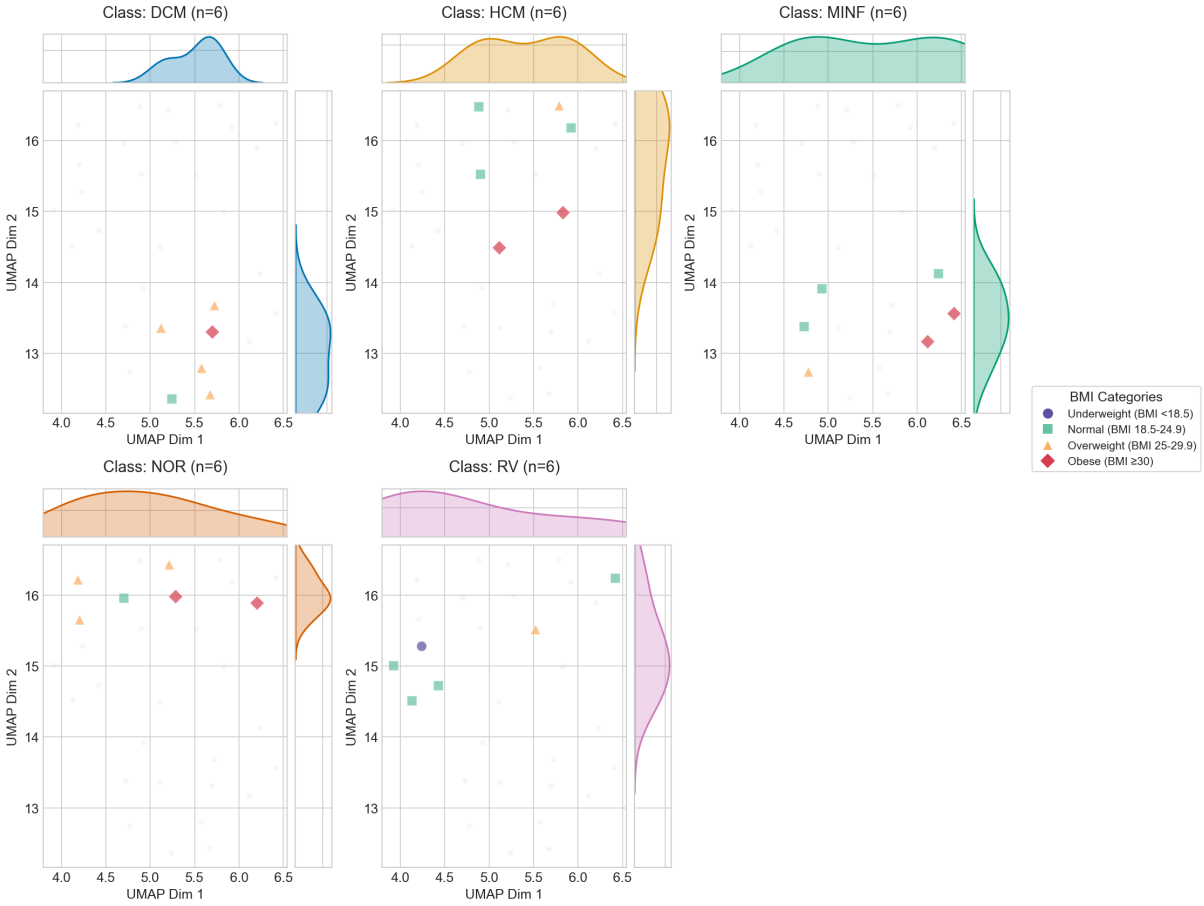

Figure D.13: Fine-tuned (Fold 1) embeddings are faceted by BMI category to illustrate the distribution shift in the latent space after training.

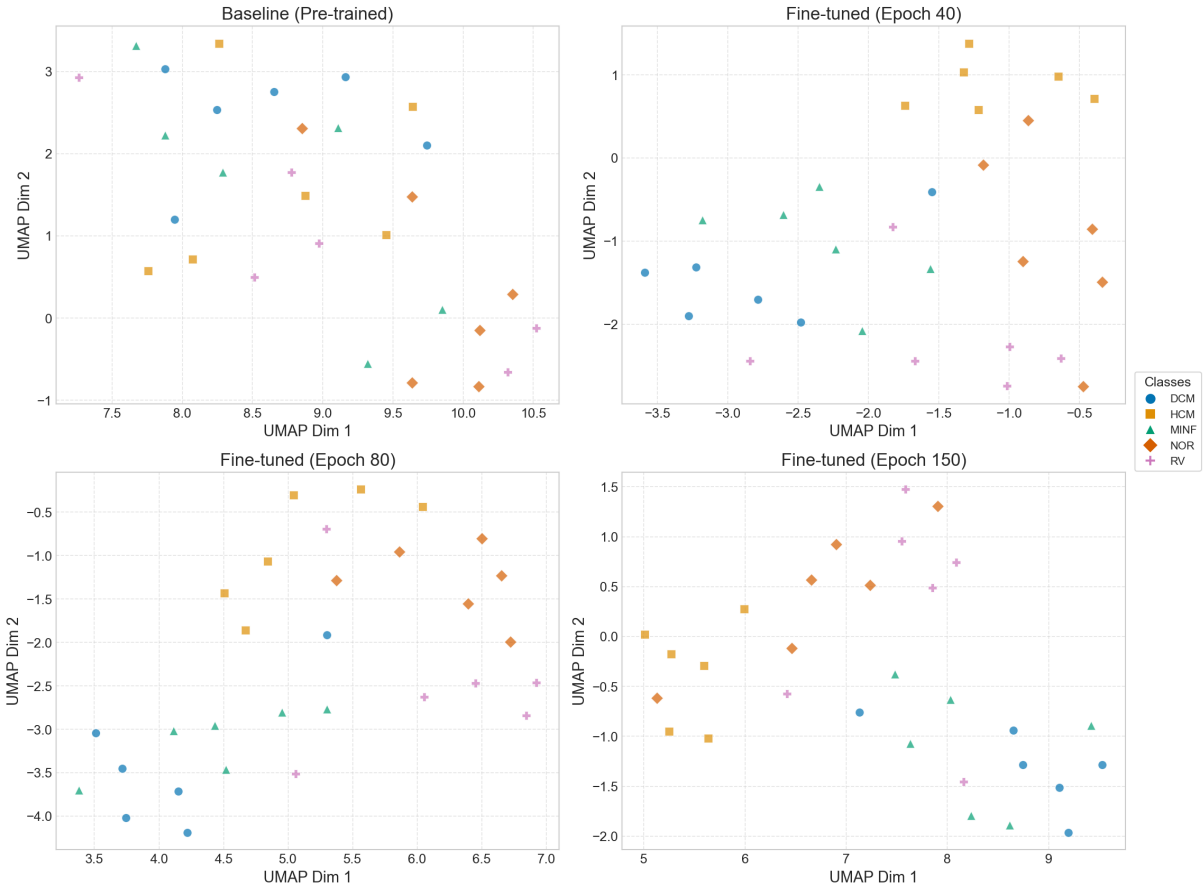

Figure D.14: Progressive evolution of the CMR embeddings for Fold 2 across training epochs.

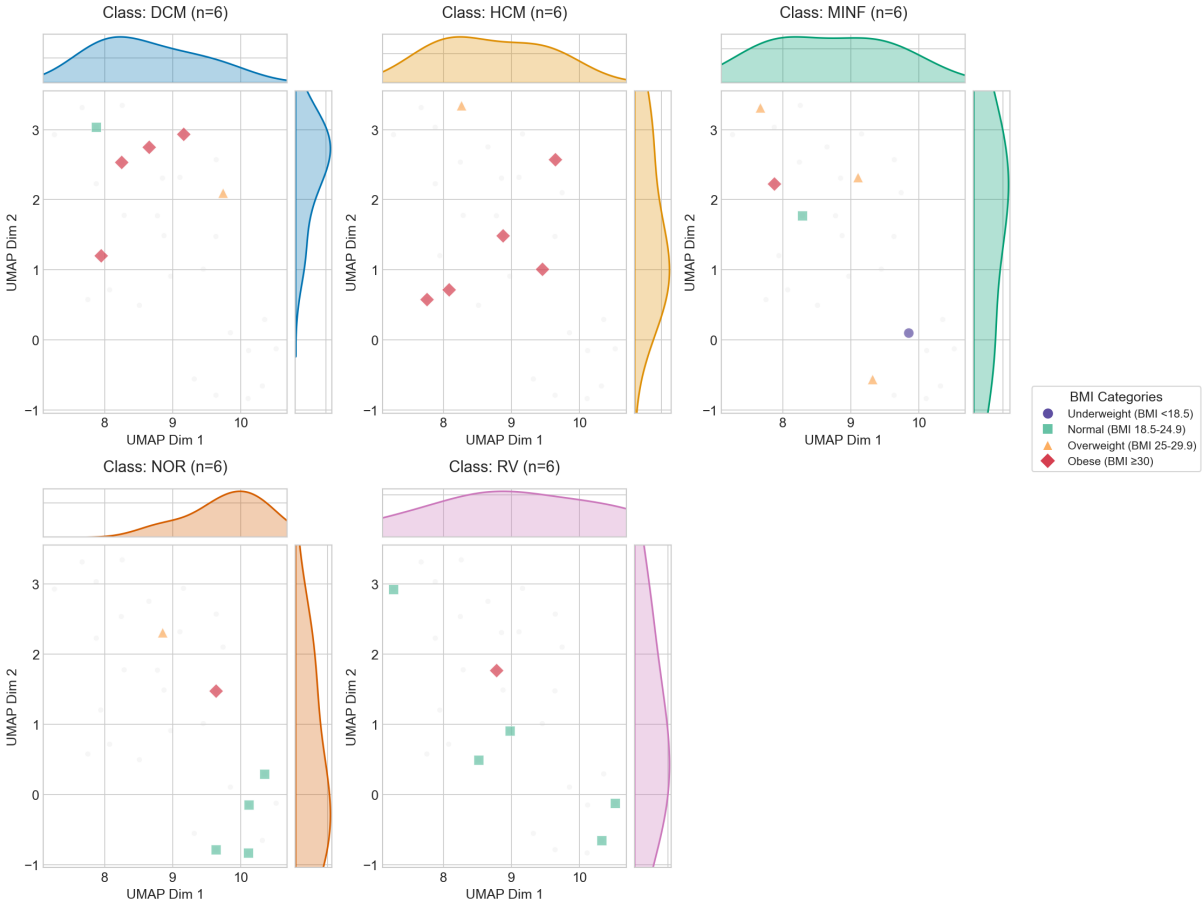

Figure D.15: Baseline (pre-trained) embeddings for Fold 2, faceted by BMI category.

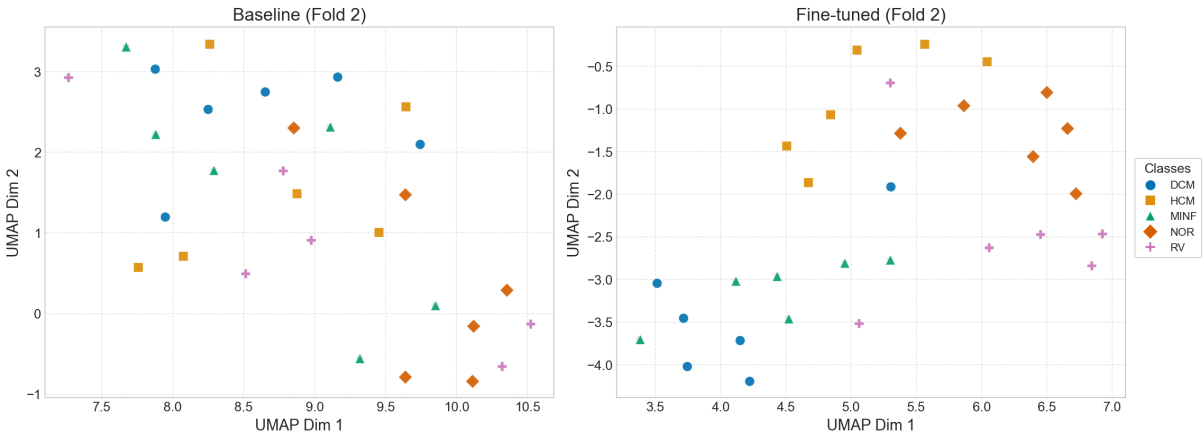

Figure D.16: Comparison of baseline and fine-tuned embeddings for Fold 2.

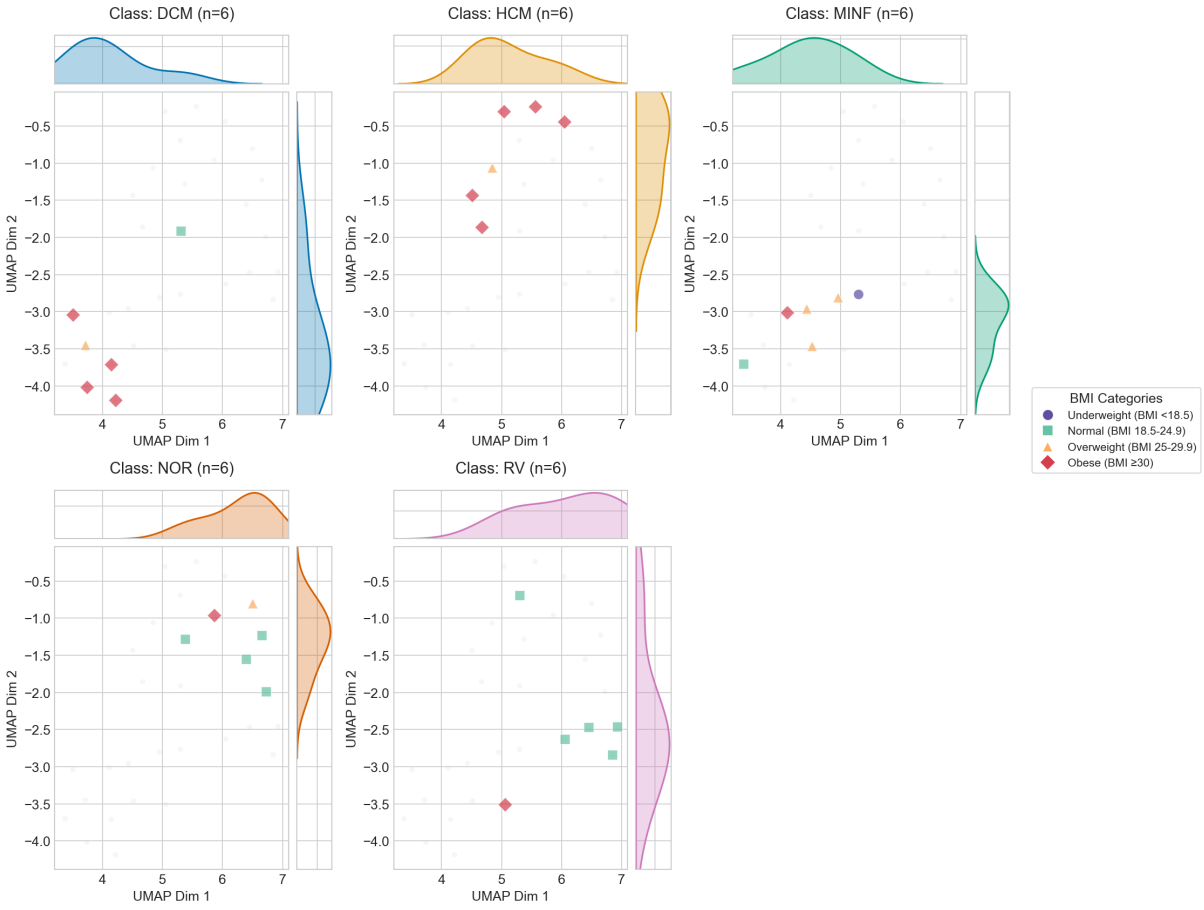

Figure D.17: Fine-tuned (Fold 2) embeddings, colored by BMI category.

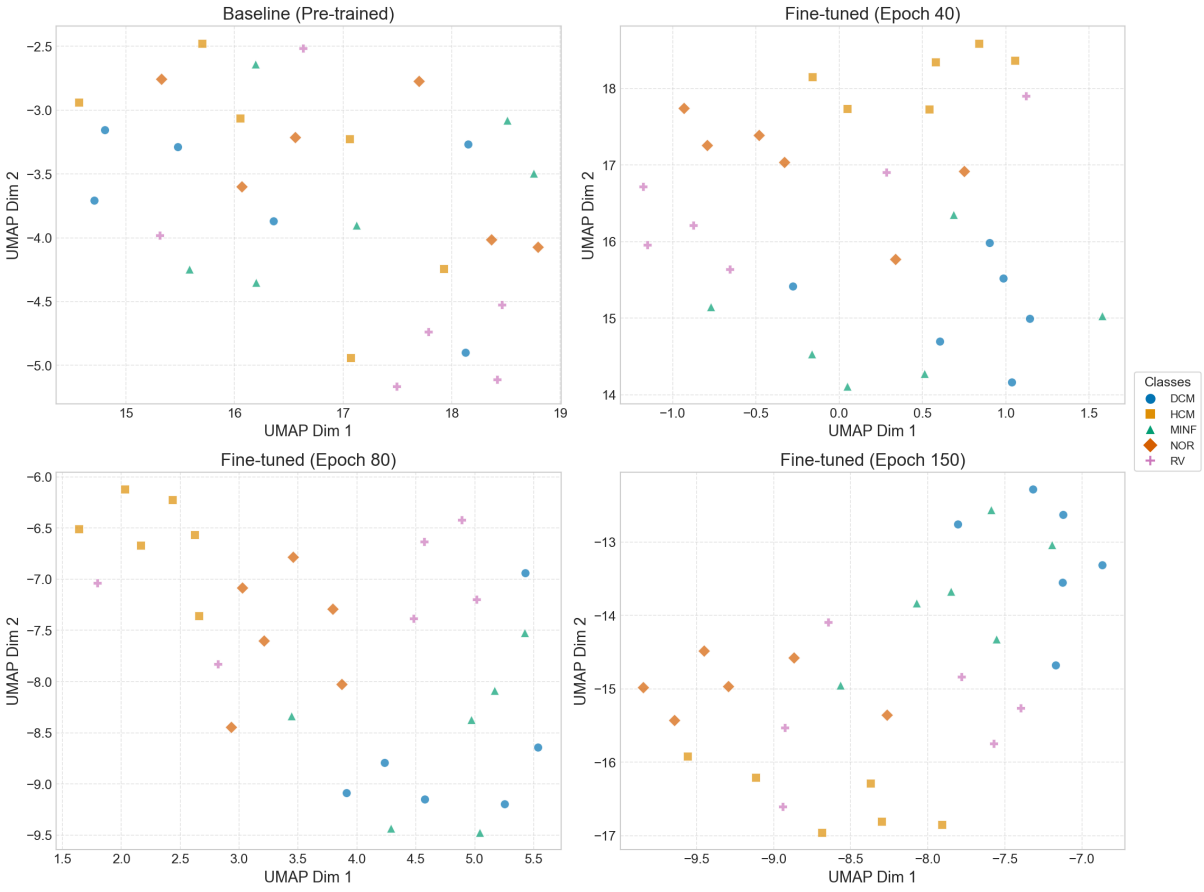

Figure D.18: Progressive evolution of the CMR embeddings for Fold 3 across training epochs.

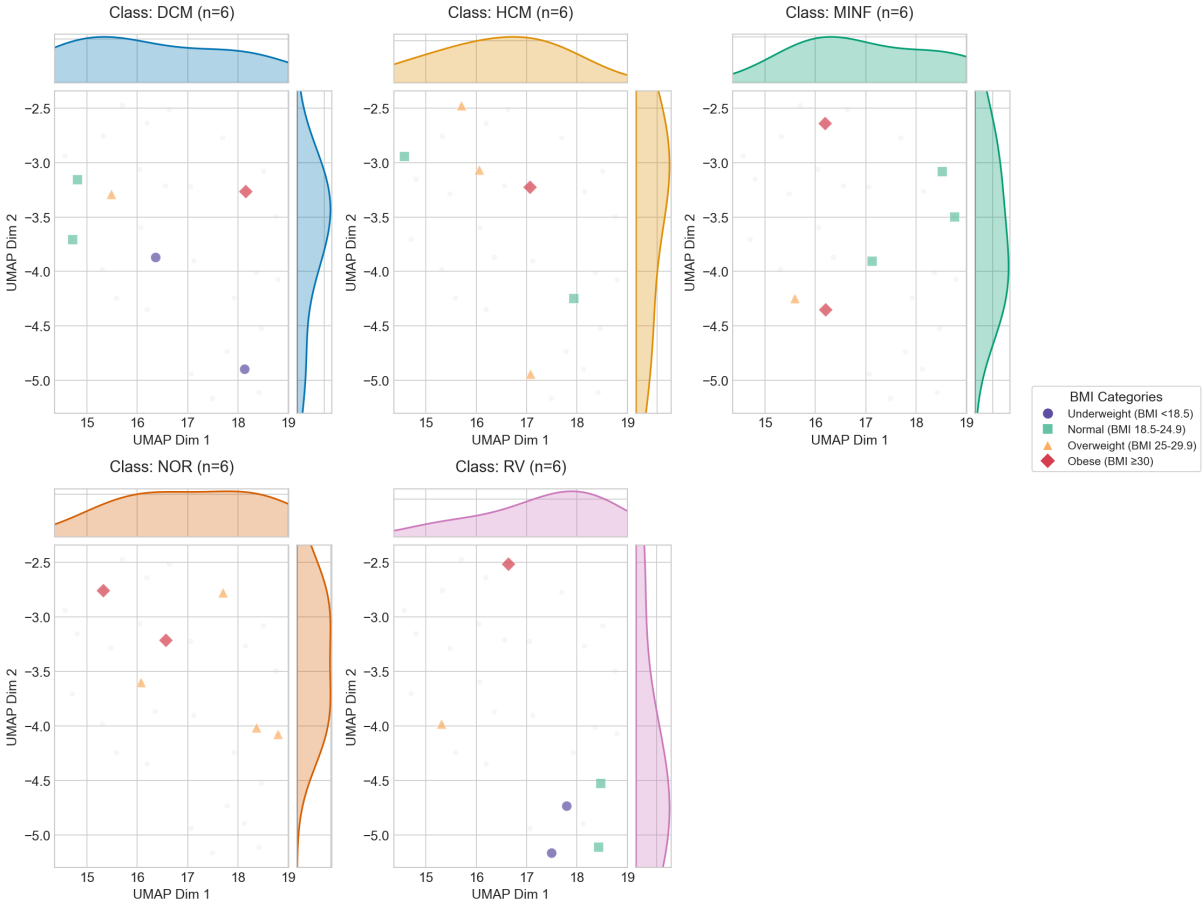

Figure D.19: Baseline (pre-trained) embeddings for Fold 3, faceted by BMI category.

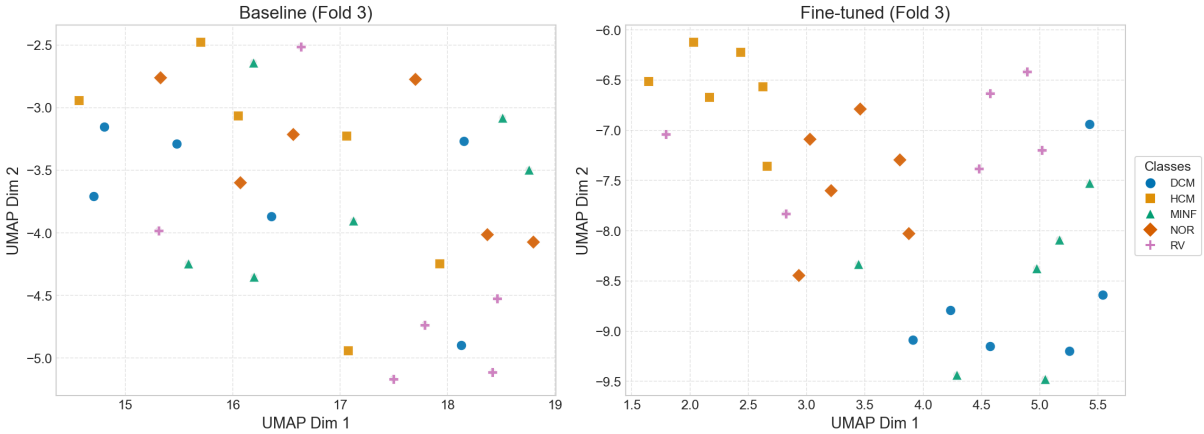

Figure D.20: Comparison of baseline and fine-tuned embeddings for Fold 3.

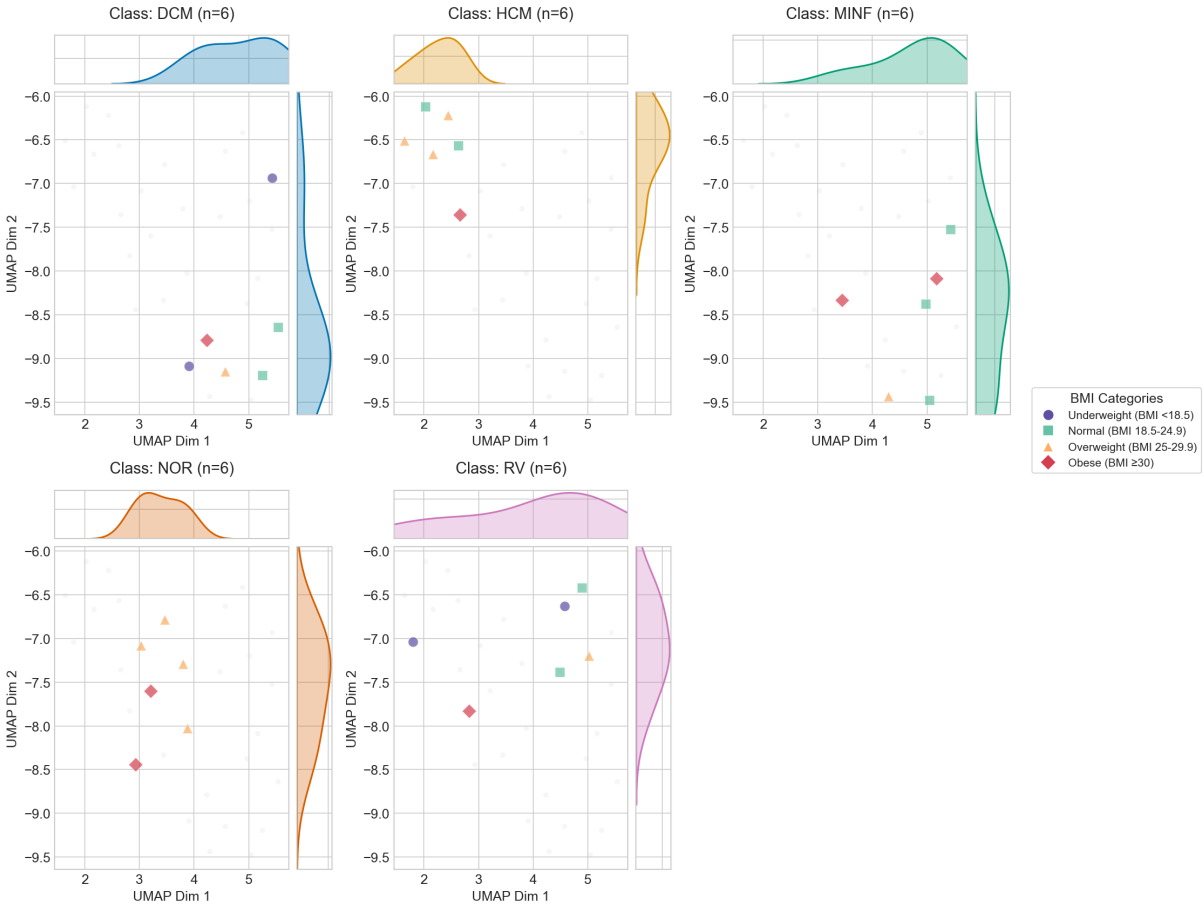

Figure D.21: Fine-tuned (Fold 3) embeddings, color-coded by BMI category.

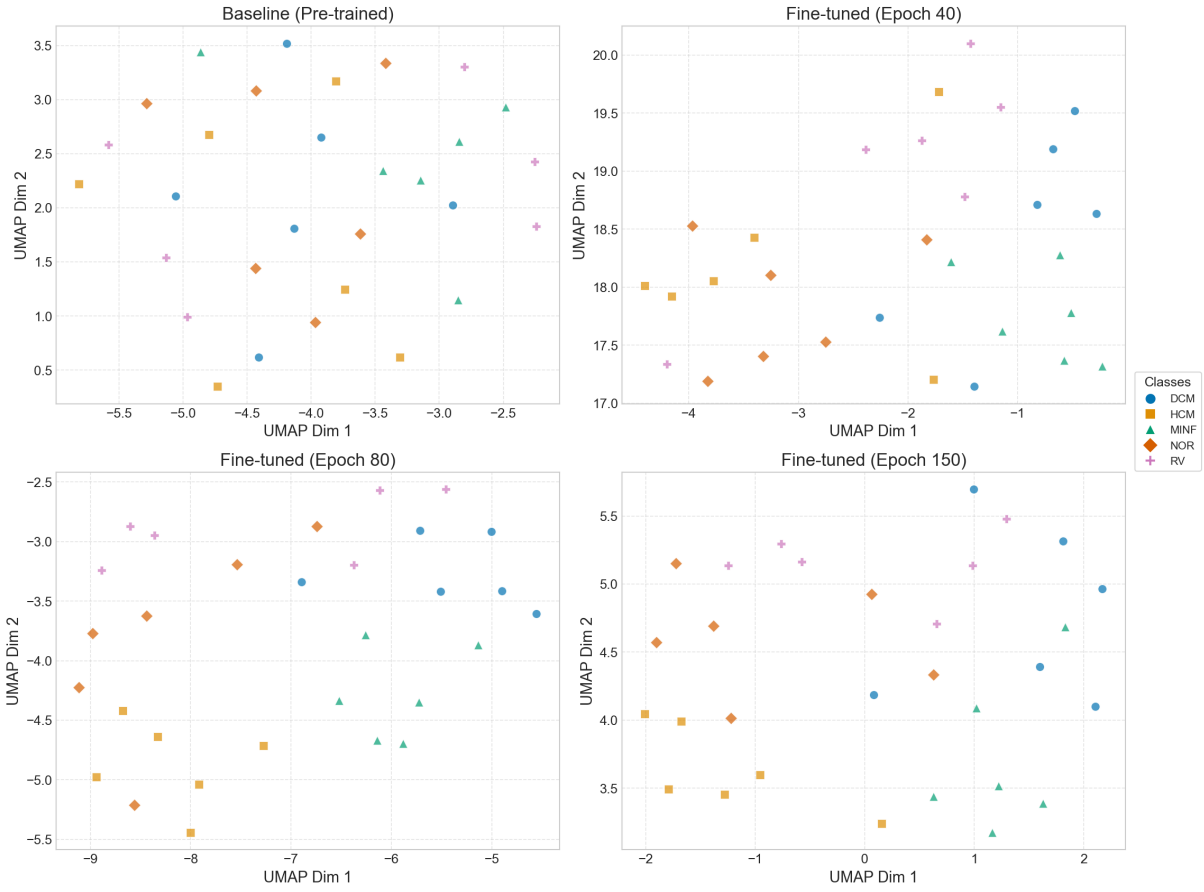

Figure D.22: Progressive evolution of the CMR embeddings for Fold 4 across training epochs.

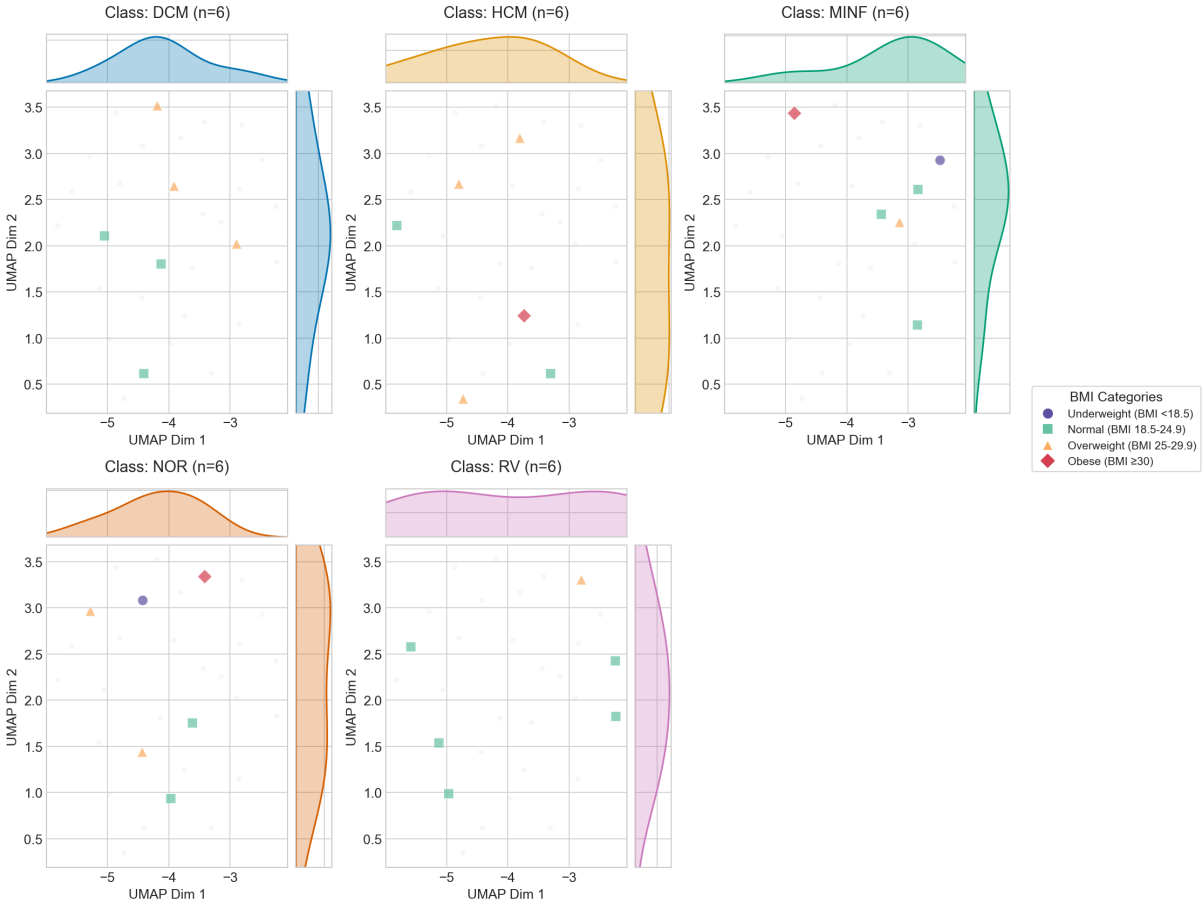

Figure D.23: Baseline (pre-trained) embeddings for Fold 4, faceted by BMI category.

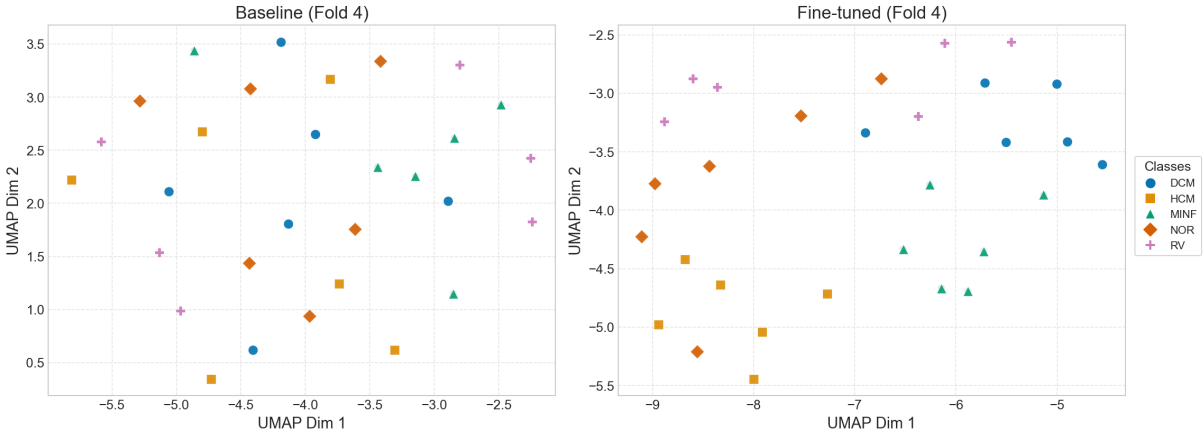

Figure D.24: Comparison of baseline and fine-tuned embeddings for Fold 4.

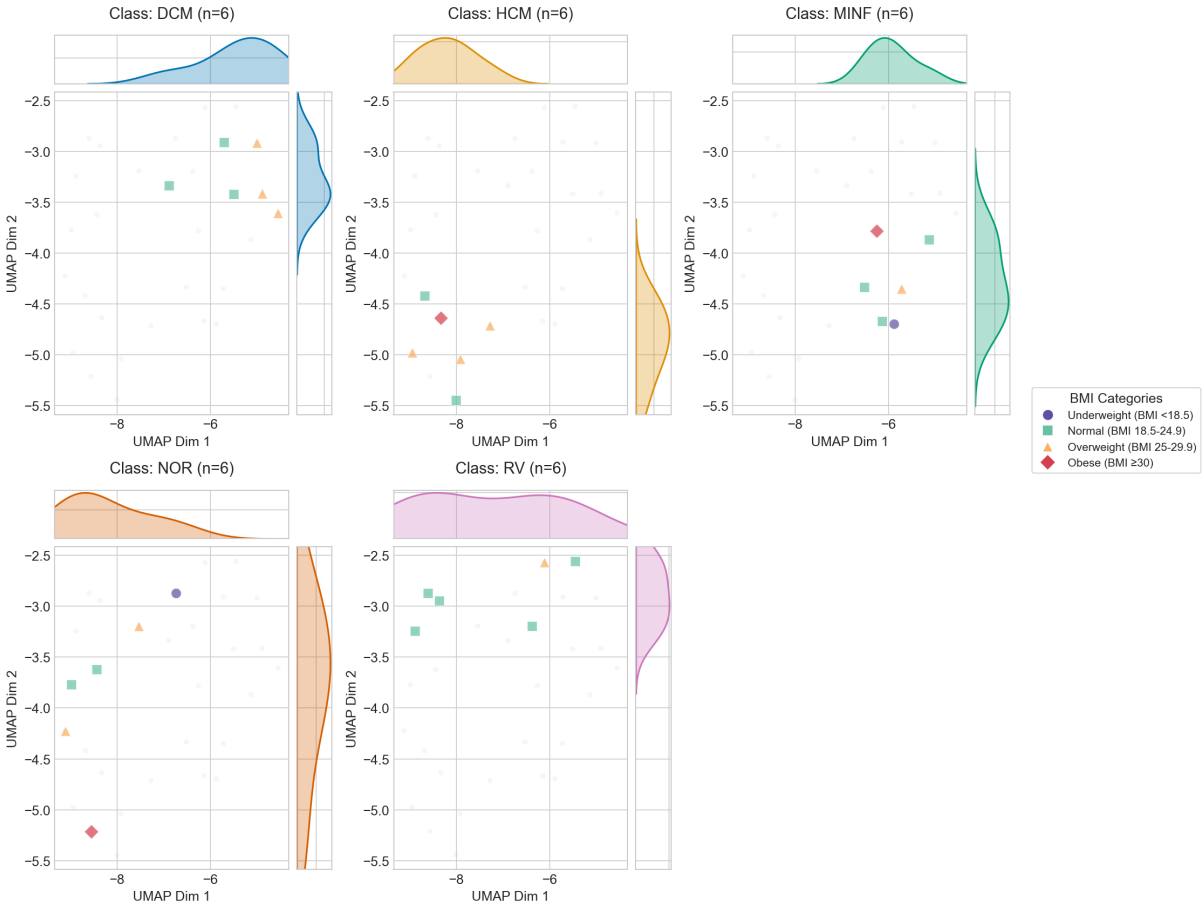

Figure D.25: Fine-tuned (Fold 4) embeddings, color-coded by BMI category.

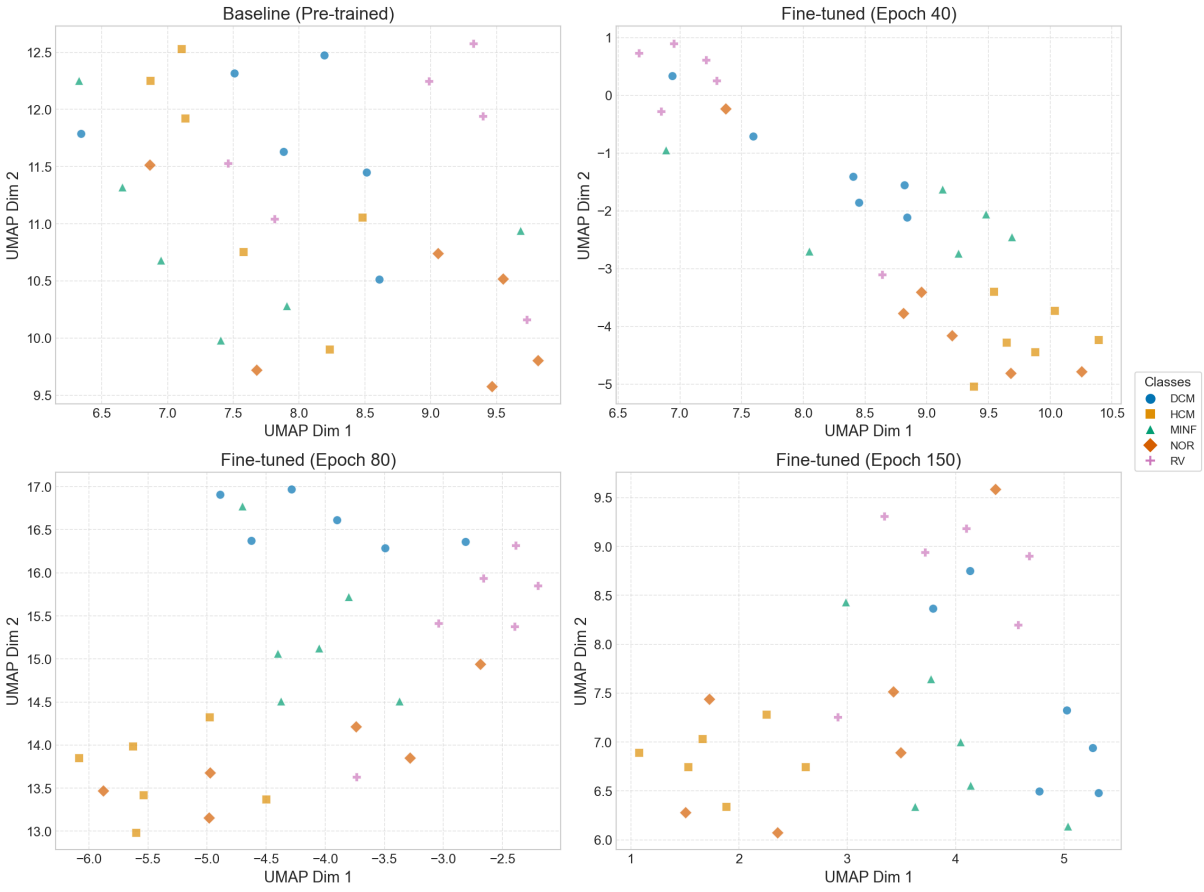

Figure D.26: Progressive evolution of the CMR embeddings for Fold 5 across training epochs.

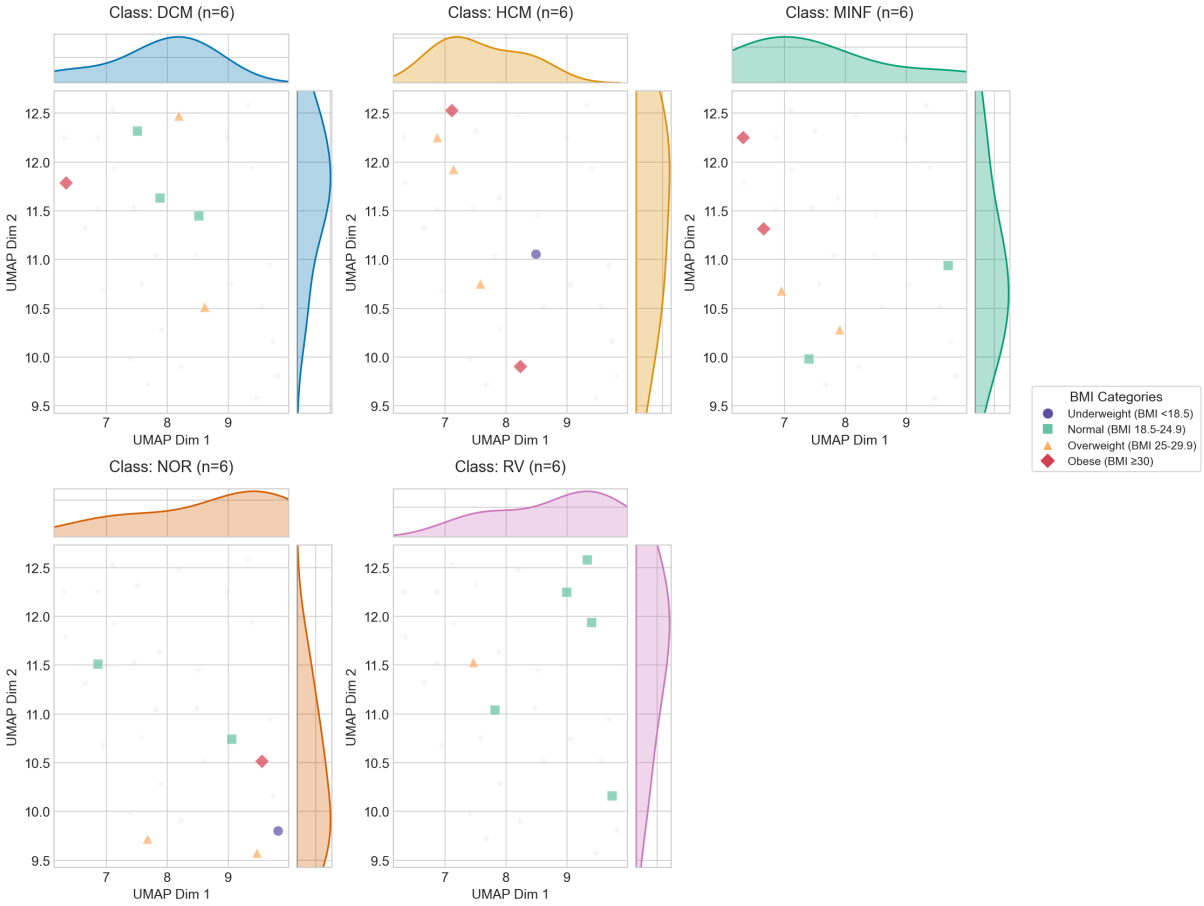

Figure D.27: Baseline (pre-trained) embeddings for Fold 5, faceted by BMI category.

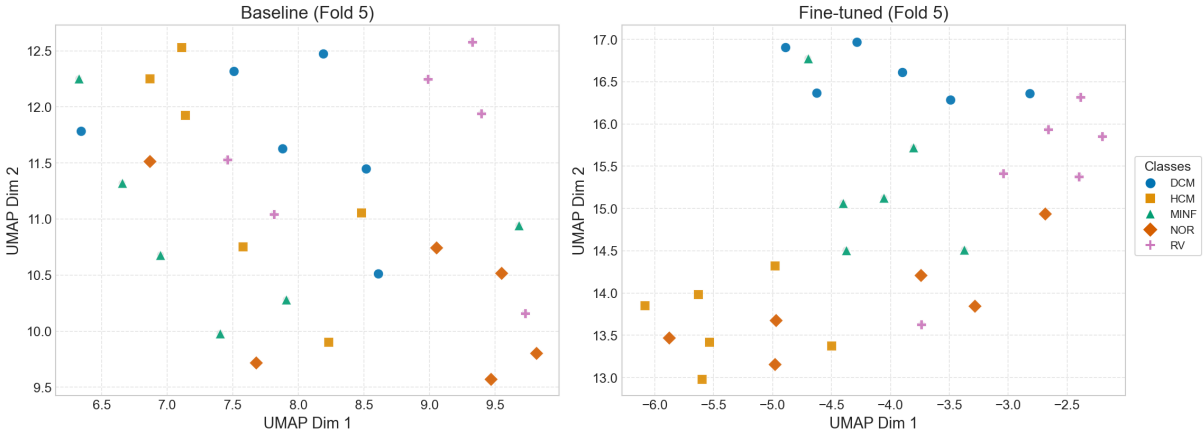

Figure D.28: Comparison of baseline and fine-tuned embeddings for Fold 5.

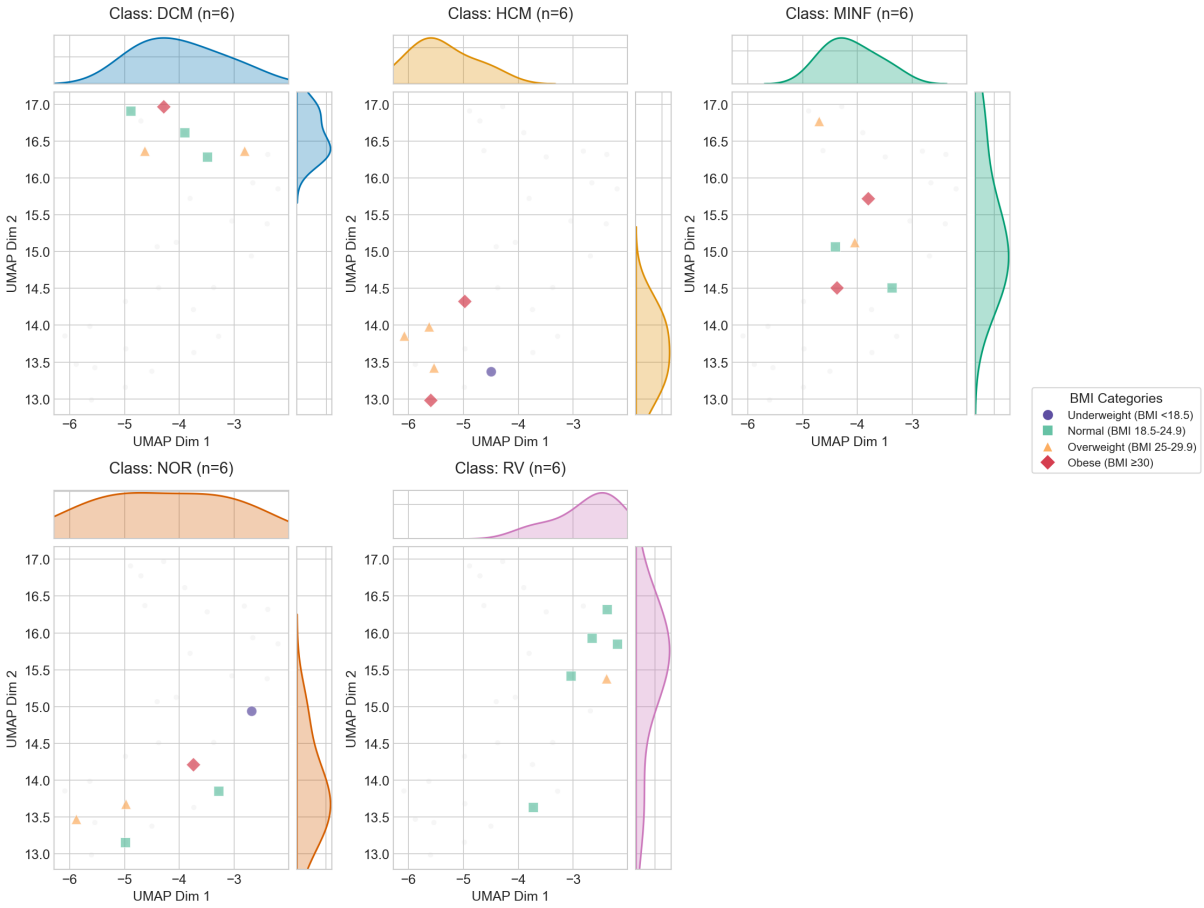

Figure D.29: Fine-tuned (Fold 5) embeddings, color-coded by BMI category.